Round-off error analysis of the conjugate gradient algorithm

TECHNISCHE HOGESCHOOL EINDHOVEN
NEDERLAND
ONDERAFDELING DER WISKUNDE

TECHNOLOGICAL UNIVERSITY EINDHOVEN
THE NETHERLANDS
DEPARTMENT OF MATHEMATICS

Round-off error analysis of the conjugate gradient algorithm

by

J.A.M. Bollen

T.H. Report 79-WSK-06

August 1979

Contents

		Page
Abstract		1
Chapter 1.	Introduction	2
1.1.	Introduction	2
1.2.	Summary	3
1.3.	Preliminary on rounding errors and floating point arithmetic	4
1.4.	Notations and conventions	8
Chapter 2.	The cg and iscg algorithm	10
2.1.	The cg algorithm	10
2.2.	The iscg algorithm	11
2.3.	Algebraic properties of the iscg algorithm	13
Chapter 3.	Round-off error analysis of one step of the iscg algorithm	18
3.1.	Introduction	18
3.2.	Round-off error analysis	18
Chapter 4.	The convergence of r_i	29
4.1.	Introduction	29
4.2.	Two steps of iscg	29
4.3.	The linear convergence of r_i	32
Chapter 5.	The convergence of x_i	35
5.1.	Introduction	35
5.2.	An estimate for the natural relative error	35
Chapter 6.	Final comments	43
6.1.	Comparison with Wozniakowski's results	43
6.2.	A class of iscg algorithms	48
References		50

Abstract

We perform the rounding error analysis of a conjugate gradient algorithm, using recursive residuals, for the computation of the solution of a system of linear equation $Ax = b$, where A is a $n \times n$ positive definite matrix.
We prove that (when the occurence of underflow is ignored) these recursively computed residual vectors r_i tend to zero if $106\ \varepsilon\ (C_1 + 2C_2 + 8)\ \kappa^2 < 1$.
Here κ is the condition number of A in the spectral norm, ε is the relative machine precision of the floating point arithmetic and C_1 and C_2 are constants depending on n and connected with the calculation of Ax and with the calculation of inner products.
This result not only holds if the initial conjugate direction vector p_0 is taken equal to the initial residual vector $r_0 := b - Ax_0$ but also if p_0 is chosen arbitrarily.
Furthermore we show that the computed sequences $\{r_i\}$ and $\{p_i\}$ converge at worst at a linear rate and that this rate is bounded by the convergence rate of the steepest descent method.
For the computed sequence $\{x_i\}$ we are only able to prove that ultimately $\|A^{\frac{1}{2}}(\hat{x} - x_i)\| / \|A^{\frac{1}{2}}\hat{x}\|$ is of order $\varepsilon(\kappa^{3/2} \log 1/\varepsilon + \kappa^2)$, where $\hat{x}$ is the solution of $Ax = b$.
Similar results are proved for the gradient algorithm, using recursive residuals.

AMS classification scheme 1979 65G05 65F10

1. Introduction

1.1. Introduction

We study a classical conjugate gradient method (cg) for the solution of a linear system $Ax = b$, where A is a $n \times n$ positive definite matrix. It is one of the variants of the cg-method developed by E. Stiefel and M.R. Hestenes [3]. In the classification of Reid [7] it is the cg-algorithm given by the formulas (2.3a), (2.4), (2.5b), (2.6a) and (2.7) of that paper. Especially we mention here our computation of the residual vector r_i. Instead of actually computing the residual vector $r_i = b - Ax_i$ at each step for computation of the conjugate direction vector p_i, we use for $i \geq 0$ the recursion relation (2.5b) :

$$r_{i+1} = r_i - a_i A p_i$$

The vectors r_i which are obtained by using this updating formula will be referred to as *recursive residual vectors*. In exact arithmetic these vectors are equal to the residual vectors $b - Ax_i$ at each step.
Algebraically cg produces the solution $\hat{x} = A^{-1}b$ after at most n steps. In the presence of round-off however the n-th computed vector x_n is not even a reasonable approximation of $\hat{x}$ if we have an ill-conditioned system. This is caused by the fact that the theoretical orthogonality relations are disturbed in the presence of round-off. However, regarded as an iterative method for the solution of large and sparse systems, continuing after more than n iterations, the method has several very pleasant features, that already have been mentioned by Reid [7].
Until now only a few theoretical analyses have been carried out to explain the numerical behaviour of cg. Wozniakowski [8] is the only one who gives a full error-analysis of a conjugate gradient algorithm. It is a version of the cg-method that is not contained in the paper of Stiefel and Hestenes [3] or in the paper of Reid [7]. One important difference with our cg-version is that Wozniakowski's version uses true residuals $r_i := b - Ax_i$.
We consider an implementation of cg in floating point arithmetic with relative machine precision ε. We will show that the computed recursive residual vectors r_i and the computed conjugate gradient vectors p_i tend to zero if $106\varepsilon(C_1 + 2C_2 + 8)\kappa^2 < 1$. Here C_1 and C_2 are constants depending on the implementation of the calculation of Ax and of inner products respectively. κ is the condition number of the matrix A. We even prove that

$$(1) \qquad \|A^{-\frac{1}{2}} r_{i+1}\| \leq (1 + \varepsilon(13C_1 + 3C_2 + 38)\kappa)\, L^i \, \|A^{-\frac{1}{2}} r_0\|$$

where L is a number close to $(\kappa-1)/(\kappa+1)$, which is the convergence rate of the steepest descent method (≡ gradient method). Hence the numerical convergence of cg is at worst linear (as far as the convergence of r_i is concerned).

We will prove that the approximants x_i ultimately satisfy

$$(2) \qquad \frac{\| A^{\frac{1}{2}}(x_i - \hat{x}) \|}{\| A^{\frac{1}{2}}\hat{x} \|} \leq 6\varepsilon\{(119 \log 1/\varepsilon + 17\beta)\kappa^{3/2} + 25(C_1 + 3)\kappa^2\}.$$

We realize that this last result is rather poor in that it involves a factor κ^2. We ascribe the appearance of this factor to the fact that we use recursive residuals. An analysis of the gradient algorithm with recursive residuals reveals the same factor.

The numerical experiments that we have carried out, confirmed the limit-properties $r_i \to 0$, $p_i \to 0$ $(i \to \infty)$ and the convergence rate expressed by (1).
Since we have executed only a rather limited number of experiments, we dare not say whether the factor κ^2 in the estimate (2) is realistic or not. We will report on these numerical experiments in another paper.

1.2. Summary

We summarize the contents of the paper.
In chapter 2 we formulate the cg-algorithm and we briefly state some basic algebraic properties of the algorithm that are important for the error analysis. We also consider the so-callled independent start conjugate gradient method (iscg). This method differs from cg only by the fact that p_0 is not coupled with r_0 but can be chosen freely. Hence cg is a special case of iscg and we will concentrate on the last method. We will derive some basic results for iscg. Most of these results were known already by Crowder and Wolfe [1], but they did not write them down explicitely. We also report on results of Powell [6] in connection with iscg.
Chapter 3 deals with the rounding error analysis of one step of iscg. We only consider the computation of r_{i+1} and p_{i+1}. We here mention the fact that in this report we have not ignored terms of any order in ε.

In chapter 4 we prove the convergence to zero of the computed vectors r_i and p_i. Furthermore we will show that the speed of convergence can be expressed by (1).
The computation of x_{i+1} is studied in chapter 5. Since $r_i = b - Ax_i$ does not hold anymore in the presence of round off, we need to analyse the difference between r_i and $b - Ax_i$. This analysis is carried out in chapter 5 where we finally prove (2).
In the final chapter we consider the gradient algorithm for the computation of the solution of $Ax = b$. We sum up the results of Wozniakowski [8] for the case when true residuals are used and we give new results for the case when recursive residuals are used. We also compare our results for iscg with Wozniakowski's results for his cg-method. Besides we introduce a class of conjugate gradient methods for which we can prove similar results on numerical behaviour as in the iscg-case.

1.3. Preliminary on rounding errors and floating point arithmetic

Throughout this report we assume that the algorithms are performed in floating point arithmetic. The floating point numbers will be assumed to have base β and a mantissa with t digits ($\beta \geq 2$, $t \geq 1$). 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 $\varepsilon = \frac{1}{2}\beta^{1-t}$.
Furthermore we assume that we have a machine with *proper rounding arithmetic* in the sense of T.J. Dekker [2].
This means that the execution of any arithmetical operation $\oplus$ (this can be $+, -, \times, /$) 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 two relations hold

$$fl(a \oplus b) = (a \oplus b)(1 + \xi), \tag{3}$$

$$(1 + \eta)fl(a \oplus b) = a \oplus b \tag{4}$$

where both

$$|\xi| \leq \varepsilon , \quad |\eta| \leq \varepsilon . \tag{5}$$

Hence, adding or subtracting two machine vectors x and y and multiplying a machine number a and a machine vector x gives computed vectors $fl(x \pm y)$ and $fl(ax)$ satisfying

(6) $\quad fl(x \pm y) = (I + F_1)(x \pm y)$,

(7) $\quad fl(ax) = (I + F_2)\, ax$,

(8) $\quad (I + G_1)fl(x \pm y) = x \pm y$,

(9) $\quad (I + G_2)fl(ax) = ax$,

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

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

and consequently

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

We suppose that the computation of Ax is implemented in such a way that the computed vector fl(Ax) 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 depends only on n.
We assume that the algorithm for inner product calculation of two machine vectors x and y satisfies

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

where D is a diagonal matrix such that

(15) $\quad \|D\| \leq \varepsilon C_2$.

The constant C_2 also depends only on n.
For many straightforward implementations $C_1 = n^{3/2}$ and $C_2 = n$.

Remark 1.
Note that 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. □

If two vectors are added then the rounding errors occuring in this operation can be expressed by (6) and (8). Another, rather unusual way to express this rounding errors is given in the following lemma. It will be of special interest if one vector is much smaller than the other vector. We will meet this situation in chapter 5.

Note that it follows from the assumption that we have proper rounding arithmetic that if a and b are machine numbers and if $|b| < (\varepsilon/\beta)|a|$ then

$$(16) \qquad fl(a + b) = a .$$

From this we can prove the following lemma.

Lemma 2. If x and y are machine vectors then

$$(17) \qquad fl(x + y) = x + (I + H)y ,$$

where H is a diagonal matrix satisfying

$$(18) \qquad |H| \leq (\beta + \varepsilon)I ,$$

and hence

$$(19) \qquad \|H\| \leq \beta + \varepsilon$$

Proof. Let $fl(x + y) = x + y + \delta s$ and let x^j denote the j^{th} component of x.
If $|y^j| < (\varepsilon/\beta)|x^j|$ then if follows from (16) that

$$(20) \qquad \delta s^j = -y^j .$$

Consequently, in that case

$$(21) \qquad |\delta s^j| = |y^j| .$$

If $|y^j| \geq (\varepsilon/\beta)|x^j|$ then it follows with (3)

$$(22) \qquad |\delta s^j| \leq \varepsilon|x^j + y^j| \leq (\beta + \varepsilon)|y^j| .$$

Hence in both cases $|\delta s^j| \leq (\beta + \varepsilon)|y^j|$.
Defining $H_{jj} := \delta s^j$, $H_{ij} = 0$ $(i \neq j)$ completes the proof. □

As an illustration of the use of lemma 2 we prove the following theorem. A similar result will be derived in chapter 5.

Theorem 3. Let y_i $(i \geq 0)$ be machine vectors satisfying

$$(23) \qquad \|y_i\| \leq L^i \|y_0\| ,$$

where $0 < L < 1$ and let s_k $(k \geq 0)$ be computed from

(24) $s_0 := y_0$;
$\underline{\text{for}}\ k \geq 1\ \underline{\text{do}}\ s_k := fl(s_{k-1} + y_k)$.

Then we have

(25) $$\overline{\lim_{k\to\infty}} \| s_k - \sum_{i=1}^{k} y_i \| \leq \varepsilon(\beta + 2)\{\frac{\log 1/\varepsilon}{\log 1/L} + 1\}\frac{\| y_0 \|}{1 - L}$$

Proof. We have for $k \geq 1$:

(26) $$s_k = s_{k-1} + y_k + t_k ,$$

where, from (6):

(27) $$\| t_k \| \leq \varepsilon(\| s_{k-1} \| + \| y_k \|)$$

and, from lemma 2:

(28) $$\| t_k \| \leq (\beta + \varepsilon)\| y_k \| .$$

From (26) we conclude

(29) $$s_k - \sum_{i=0}^{k} y_i = \sum_{j=1}^{k} t_j$$

From (23) and (28) it follows that Σt_j converges. Therefore we devide the sum in (29) into two parts:

(30) $$\| s_k - \sum_{i=0}^{k} y_i \| \leq \sum_{j=1}^{\ell} \| t_j \| + \sum_{j=\ell+1}^{k} \| t_j \|, \quad (k > \ell \geq 2).$$

For indices $i \leq \ell$ we use estimate (27) and as soon as y_i is small (of order ε) we use estimate (28) (this last restriction gives the condition for ℓ). From (27), (28) and (29) we obtain

(31) $$\| t_j \| \leq \varepsilon(\| y_j \| + \sum_{i=0}^{j-1} \| y_i \| + \sum_{i=1}^{j-1} \| t_i \|) \leq \varepsilon(\beta + 2) \sum_{i=0}^{\infty} \| y_i \|$$

$$\leq \varepsilon(\beta + 2)\| y_0 \|/(1 - L)$$

and consequently

(32) $$\sum_{j=1}^{\ell} \| t_j \| \leq \varepsilon(\beta + 2)\ell \| y_o \| /(1 - L)$$

For the second sum in (30) we find, using (28),

$$(33) \qquad \sum_{j=\ell+1}^{k} \|t_j\| \le (\beta + \varepsilon) \sum_{j=\ell+1}^{\infty} \|y_j\| \le (\beta + \varepsilon) L^{\ell+1} \|y_0\| / (1 - L)$$

Substitution of (32) and (33) in (30) yields

$$(34) \qquad \|s_k - \sum_{i=0}^{k} y_i\| \le (\beta + 2)\|y_0\|(\varepsilon\ell + L^{\ell+1})/(1 - L) \ .$$

Now let $\ell \ge 2$ be the smallest integer such that $L^{\ell+1} \le \varepsilon$. Then certainly

$$(35) \qquad \ell < (\log 1/\varepsilon)/(\log 1/L) \ .$$

Hence form (34) we finally get for k sufficiently large

$$\|s_k - \sum_{i=0}^{k} y_i\| \le \varepsilon(\beta + 2)\{\frac{\log 1/\varepsilon}{\log 1/L} + 1\}\frac{\|y_0\|}{1 - L}$$

□

1.4. Notations and conventions

Matrices are denoted by capital letters, vectors and numbers by small letters.

The linear equations to be solved are written as $Ax = b$, where A is supposed to be an $n \times n$ real (symmetric) positive definite matrix and b is supposed to be a real (column) vector with n components.

We further mean by

A^T	the transposed matrix of A,
A^{-1}	the inverse of A,
$A^{\frac{1}{2}}$	the unique positive definite matrix satisfying $A^{\frac{1}{2}}.A^{\frac{1}{2}} = A$,
$A^{-\frac{1}{2}}$	the inverse of $A^{\frac{1}{2}}$,
$\|A\|$	the matrix which elements are defined by $(\|A\|)_{ij} := \|A_{ij}\|$,
$A < A'$	that for all elements $A_{ij} < A'_{ij}$,
I	the unit matrix,
$\hat{x}$	the solution vector $A^{-1}b$ of the linear system $Ax = b$,
x^j	the j^{th} component of vector x,
(x,y)	the ordinary Euclidean inner product $x^T y$ of the vectors x and y,
$\|x\|$	the Euclidean norm $(x,x)^{\frac{1}{2}}$ of vector x,
$\|A\|$	the spectral norm $\max_{x \neq 0} (\|Ax\|/\|x\|)$ of A,
κ	the condition number $\|A\|.\|A^{-1}\|$ of A,

β	the base of the floating point numbers in use,
t	the length of the mantissa of the floating point number,
ε	the relative machine precision; $\varepsilon = \frac{1}{2}\beta^{1-t}$
$fl(\cdot)$	the computed value, using floating point arithmetic, of the expression between brackets,
C_1	a constant depending on n and appearing in the upperbound for the relative error for the computation of Ax (see section 1.3),
C_2	a constant depending on n and appearing in the upperbound for the relative error for inner product calculation (see section 1.3),
$\lim x_i$	the limit superior of the sequence $\{x_i\}$,
cg	the conjugate gradient algorithm defined in section 2.1 ,
iscg	the independent start conjugate gradient algorithm defined in section 2.2 ,
trg	the gradient algorithm defined in section 6.1, using true residuals (formula 6.4)),
rrg	the gradient algorithm defined in section 6.1, using recursive residuals (formula 6.5)),
wcg	Wozniakowski's version of the conjugate gradient method, described in section 6.1.

In any chapter theorems, lemma's, definitions, algoritms and remarks are numbered 1, 2, ... and formulas are numbered (1), (2),...
If we refer to theorem 2 (say) in some chapter, this means theorem 2 of the same chapter. If we refer to theorem 1.2, this means theorem 2 of chapter 1.

2. The cg and iscg algorithm

2.1. The cg algorithm

In this section we formulate the conjugate gradient algorithm (cg) and sum up some of its most important algebraic properties.
We will follow the notation of Hestens and Stiefel [3].

Given a system

(1) $Ax = b$

of n linear equations whose matrix is symmetric and positive definite, then the cg-algorithm can be formulated by the following statements.

Algorithm 1. *The conjugate gradient algorithm:*

take x_0; $p_0 := r_0 := b - Ax_0$;
$i := 0$;
while $r_i \neq 0 \vee p_i \neq 0$ **do**
begin

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

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

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

(5) $b_i := -(r_{i+1}, Ap_i)/(p_i, Ap_i)$;

(6) $p_{i+1} := r_{i+1} + b_i p_i$;

(7) $i := i + 1$

end.

By the inner product we mean the ordinary scalar product $(x,y) = x^T y$.

Remark 2.
The formulas (2) and (5) are not the formulas that were used as basis relations in the cg-algorithm by Hestenes and Stiefel (see [3], section 5). Actually, they used the following two relations:

(8) $\quad a_i = (r_i, r_i)/(p_i, Ap_i)$,

(9) $\quad b_i = (r_{i+1}, r_{i+1})/(r_i, r_i)$.

(These are the formulas (2.3b) and (2.6b) of Reid [7]).
Taking either (2) or (8) for a_i and taking either (5) or (9) for b_i in algorithm 1, we obtain 4 different algorithms which algebraically give the same results. From a numerical point of view however they are different and in the presence of round off we will only consider the choices (2) and (5) in this report. □

Before mentioning some properties of cg we first give a definition.

Definition 3. Let A be a symmetric n × n matrix, then the vectors $x, y \in \mathbb{R}^n$ are said to be *conjugate* if $(x, Ay) = 0$ whereas $x \neq 0$ and $y \neq 0$.

Note that mutually conjugate vectors are linearly independent, if A is positive definite.
The most important property of cg is *the finite termination property*: As long as $x_i \neq \hat{x}$ the successive directions $p_0, p_1, \ldots, p_i$ are mutually conjugate and consequently $x_i = \hat{x}$ for some $i < n$.

A further property of cg is that x_{i+1} minimizes $\| A^{\frac{1}{2}}(\hat{x} - x) \|$ on the affine set passing trough x_0 and spanned by $p_0, p_1, \ldots, p_i$. Hence $\| A^{\frac{1}{2}}(\hat{x} - x_i \|$ decreases monotonically.

Another property of cg is that $\| \hat{x} - x_i \|$ decreases monotonically as i increases. The following relation holds

(10) $\quad \| \hat{x} - x_{i+1} \|^2 = \| \hat{x} - x_i \|^2 - (\| p_i \| / \| A^{\frac{1}{2}} p_i \|)^2 (\| A^{-\frac{1}{2}} r_i \|^2 + \| A^{-\frac{1}{2}} r_{i+1} \|^2)$

Hestenes and Stiefel [3] gave a proof of (10) using a backward induction based on the fact that $x_{i+1} = \hat{x}$ for some $i < n$. Kammerer and Nashed [5] gave a proof by forward induction, that is also valid in the Hilbert space case.

2.2. The iscg algorithm

We now introduce the independent start conjugate gradient algorithm (iscg).

Algorithm 4. *The independent start conjugate gradient algorithm:*

take x_0 ; $r_0 := b - Ax_0$;
take $p_0 \neq 0$;
  $i := 0$;
  while $r_i \neq 0 \vee p_i \neq 0$ do
    begin calculate a_i, x_{i+1}, r_{i+1}, b_i, p_{i+1} from (2), (3), (4), (5) and (6) ;
      $i := i + 1$
    end .

Remark 5.

Apart from the start this method is exactly the same as the cg-method. Instead of the start $p_0 := r_0 := b - Ax_0$ we take $r_0 = b - Ax_0$ and $p_0 \neq 0$ may be chosen arbitrarily. The cg-method is a special case of iscg and consequently all the properties of iscg also hold for cg. □

Remark 6.

It is quite obvious from an induction argument that the residual vector corresponding with x_i is equal to r_i for all $i \geq 0$, i.e.

$$r_i = b - Ax_i \quad . \tag{11}$$

Since r_i is not calculated from this formula but from recursion (4) we call r_i *the recursive residual vector*. The vector $b - Ax_i$ is called *the true residual vector*. If exact arithmetic is in use the formulas would give exactly the same results. From (11) it immediately follows that

$$\hat{x} - x_i = A^{-1} r_i \quad . \tag{12}$$

This is called *the error vector*.
Relation (11) also immediately gives

$$A^{\frac{1}{2}}(\hat{x} - x_i) = A^{-\frac{1}{2}} r_i \quad . \tag{13}$$

This is called *the natural error vector*. This name will be explained in remark 9. *The natural relative error* is defined by $\| A^{\frac{1}{2}}(\hat{x} - x)\| / \| A^{\frac{1}{2}}\hat{x}\|$.
In the remaining part of this chapter we will concentrate on r_i and p_i but from the foregoing three relations the results can easily be interpreted for $\hat{x} - x_i$. Note that $r_i = 0$ implies $x_i = \hat{x}$ and $r_i \to 0$ implies $x_i \to \hat{x}$ $(i \to \infty)$. □

Remark 7.

The main purpose of introducing iscg is the fact that iscg is a *one-step method*: for every i, the step from x_i, r_i, p_i to x_{i+1}, r_{i+1}, p_{i+1} can be considered as the first step of iscg with start vectors x_i, r_i and p_i (x_i and r_i are coupled by (11)). □

Remark 8.

One could also consider iscg with the formulas (8) and (9) just like we did for cg. This gives algebraically different algorithms and these algorithms have different algebraic properties. We will discuss this in another paper. □

Remark 9.

The choice (2) for the formula for a_i is a natural choice from the following point of view.

The function

$$(14)\qquad f(a) := \| A^{\frac{1}{2}}(\hat{x} - x_i - ap_i)\|^2 = \| A^{\frac{1}{2}}(\hat{x} - x_i)\|^2 - 2a(A^{\frac{1}{2}}(\hat{x} - x_i), A^{\frac{1}{2}}p_i) + a^2\| A^{\frac{1}{2}}p_i\|^2 = \| A^{\frac{1}{2}}(\hat{x} - x_i)\|^2 - 2a(r_i, p_i) + a^2(p_i, Ap_i)$$

reaches its minimum value for $a = (r_i, p_i)/(p_i, Ap_i)$. Hence x_{i+1} minimizes $A^{\frac{1}{2}}(\hat{x} - x)$ along the line through x_i parallel to p_i, if a_i is computed from (2). This also means that $A^{\frac{1}{2}}(\hat{x} - x_i)$ seems to be the natural norm to measure the error of the approximate solution x_i. □

2.3. Algebraic properties of the iscg algorithm

We are now ready to prove two important theorems concerning the convergence of iscg. Most of the results were known already by Crowder and Wolfe [1] although they did not write them down explicitely. Our main reason to give the proofs here is because of the fact that we will use the same kind of argumentation to prove the convergence of iscg in the presence of round off.

Theorem 10.

Consider iscg and let $x_0, p_0 \in \mathbb{R}^n$, $p_0 \neq 0$. For $i \geq 0$ we have, if $r_i \neq 0 \wedge p_i \neq 0$:

$$(15)\qquad (r_{i+1}, p_i) = 0 \ ,$$

$$(16)\qquad (r_{i+1}, p_{i+1}) = (r_{i+1}, r_{i+1}) \ ,$$

$$\|A^{-\frac{1}{2}}r_{i+1}\|^2 + \|a_iA^{\frac{1}{2}}p_i\|^2 = \|A^{-\frac{1}{2}}r_i\|^2 , \tag{17}$$

$$(p_{i+1},Ap_i) = 0 , \tag{18}$$

$$\|A^{\frac{1}{2}}p_{i+1}\|^2 + \|b_iA^{\frac{1}{2}}p_i\|^2 = \|A^{\frac{1}{2}}r_{i+1}\|^2 . \tag{19}$$

Proof.

If $r_i \neq 0 \wedge p_i \neq 0$ then r_{i+1} and p_{i+1} are well-defined.
From (2) and (4) it immediately follows that

$$(r_{i+1},p_i) = (r_i,p_i) - a_i(p_i,Ap_i) = 0. \tag{20}$$

Together with (6) this yields

$$(r_{i+1},p_{i+1}) = (r_{i+1},r_{i+1}) + b_i(r_{i+1},p_i) = (r_{i+1},r_{i+1}) . \tag{21}$$

From (4) it follows that

$$A^{-\frac{1}{2}}r_{i+1} + a_iA^{\frac{1}{2}}p_i = A^{-\frac{1}{2}}r_i . \tag{22}$$

By taking squared norms of left and right hand sides and using (21) we get

$$\|A^{-\frac{1}{2}}r_{i+1}\|^2 + \|a_iA^{\frac{1}{2}}p_i\|^2 = \|A^{-\frac{1}{2}}r_i\|^2 . \tag{23}$$

From (5) and (6) it immediately follows that

$$(p_{i+1},Ap_i) = (r_{i+1},Ap_i) + b_i(p_i,Ap_i) = 0 . \tag{24}$$

From (6) it follows that

$$A^{\frac{1}{2}}p_{i+1} - b_iA^{\frac{1}{2}}p_i = A^{\frac{1}{2}}r_{i+1} . \tag{25}$$

By taking squared norms of left and right hand sides and using (24) we get

$$\|A^{\frac{1}{2}}p_{i+1}\|^2 + \|b_iA^{\frac{1}{2}}p_i\|^2 = \|A^{\frac{1}{2}}r_{i+1}\|^2 . \tag{26}$$

□

Remark 11.

Note that if follows from (16) that $r_{i+1} \neq 0$ implies $p_{i+1} \neq 0$. Therefore it follows that if iscg ends then it ends because of the fact that $r_{i+1} = 0$ as well as $p_{i+1} = 0$. Consequently the condition $p_{i+1} = 0$ could be left out in the stopping criterion. □

Theorem 12.

Consider iscg and let $x_0,p_0 \in \mathbb{R}^n$, $p_0 \neq 0$, $r_0 \neq 0$.

Then

$$\|A^{-\frac{1}{2}}r_i\| \leq \|A^{-\frac{1}{2}}r_0\| , \tag{27}$$

and, if $i \geq 1$ and $r_i \neq 0$, then

$$\|A^{-\frac{1}{2}}r_{i+1}\| \leq (\kappa - 1)/(\kappa + 1)\|A^{-\frac{1}{2}}r_i\| . \tag{28}$$

Consequently, either $r_i = 0$ for some $i \geq 1$ or $r_i \to 0$ $(i \to \infty)$.

Proof.

Inequalitly (27) follows immediately from (17).

Let $i \geq 0$ and $r_i \neq 0$. Using the definition of a_i, (17) may be written as

$$\|A^{-\frac{1}{2}}r_{i+1}\|^2 = \{1 - (r_i,p_i)^2/(\|A^{\frac{1}{2}}p_i\| \, \|A^{-\frac{1}{2}}r_i\|)^2\}\|A^{-\frac{1}{2}}r_i\|^2 . \tag{29}$$

Similarly, from the definition of b_i, (19) may be written as

$$\|A^{\frac{1}{2}}p_{i+1}\|^2 = \{1 - (r_{i+1},Ap_i)^2/(\|A^{\frac{1}{2}}p_i\| \, \|A^{\frac{1}{2}}r_{i+1}\|)^2\}\|A^{\frac{1}{2}}r_{i+1}\|^2 . \tag{30}$$

Hence, certainly for $i \geq 1$:

$$\|A^{\frac{1}{2}}p_i\| \leq \|A^{\frac{1}{2}}r_i\| . \tag{31}$$

Substitution of (16) in (29) and using (30) gives for $i \geq 1$:

$$\|A^{-\frac{1}{2}}r_{i+1}\|^2 \leq \{1 - \|r_i\|^4/(\|A^{\frac{1}{2}}r_i\| \, \|A^{-\frac{1}{2}}r_i\|)^2\}\|A^{-\frac{1}{2}}r_i\|^2 . \tag{32}$$

From the Kantorovich inequality (see [4], p. 83)

$$\frac{(r,r)^2}{(r,Ar)(r,A^{-1}r)} \geq \frac{4\kappa}{(\kappa + 1)^2} \tag{33}$$

we finally get for $i \geq 1$:

$$\|A^{-\frac{1}{2}}r_{i+1}\|^2 \leq (\kappa - 1)^2/(\kappa + 1)^2\|A^{-\frac{1}{2}}r_i\|^2 \tag{34}$$

which proves (28).

Since $(\kappa - 1)/(\kappa + 1) < 1$ this implies that if $r_i \neq 0$ for all $i \geq 1$ then $\|A^{-\frac{1}{2}}r_i\| \to 0$ $(i \to \infty)$ and since $\|r_i\| \leq \|A^{+\frac{1}{2}}\| \, \|A^{-\frac{1}{2}}r_i\|$ then also $r_i \to 0$ $(i \to \infty)$. □

Remark 13.

From (28) it follows that if no $r_i = 0$ then the convergence of $A^{-\frac{1}{2}}r_i$ is at worst linear. Crowder and Wolfe [1] gave an example of iscg in which

the ratio $\|A^{-\frac{1}{2}}r_{i+1}\|/\|A^{-\frac{1}{2}}r_i\|$ is constant for all $i \geq 0$ and hence the finite termination property of cg does not hold in all cases for iscg. Obviously there are initial vectors r_0, p_0 for which the convergence is only linear. □

Powell [6] proved the following stronger results for iscg.

Theorem 14.

If $r_i \neq 0$ for all $0 \leq i \leq n + 1$ then:

(35) There exists an ℓ satisfying $2 \leq \ell < n$ such that $p_1,\ldots,p_\ell$ are mutually conjugate and p_1 and $p_{\ell+1}$ are not conjugate.

(36) For all $i \geq 0$ the directions $p_{i+1},\ldots,p_{i+\ell}$ are mutually conjugate, but p_{i+1} and $p_{i+\ell+1}$ are not conjugate.

(37) Termination never occurs and convergence to the solution occurs at a linear rate.

Remark 15.

The condition $\ell \geq 2$ in theorem 14 immediately follows from (18) which states that always $(p_1,Ap_2) = 0$. If $p_1,\ldots,p_n$ are mutually conjugate then $r_{n+1} = 0$ since x_{n+1} then minimizes $\|A^{\frac{1}{2}}(\hat{x} - x)\|$ on the affine set passing through x_1 and spanned by the n independent vectors $p_1,\ldots,p_n$. Therefore $\ell < n$ in theorem 14. □

Remark 16.

The most important conclusion of theorem 14 is that iscg either terminates within (n + 1) iterations or convergence to the solution occurs at a linear rate. Powell also shows that in the general case, when both r_0 and p_0 are arbitrary, then the linear rate of convergence is usual. We think that this last fact has been overlooked in the literature. For instance, it means that if during the cg iterations r_i and p_i are computed exactly in all steps exept from one, then we may expect the convergence to be only linear. □

Remark 17.

Obviously iscg generally does not end and hence for practical implementation one needs an extra stopping criterion for the case where $r_i \neq 0$ and $p_i \neq 0$ for all i. We will not formulate a stopping criterion here. □

Remark 18.

We finally mention the fact that (14) does not hold for iscg and that there exist initial vectors r_0 and p_0 for which the error vector $\|\hat{x} - x_i\|$ does not decrease monotonically. □

3. Round-off analysis of one step of the iscg algorithm

3.1. Introduction

In the presence of rounding errors one of the most pleasant features of the conjugate gradient method, the finite termination property, does not hold anymore. For ill-conditioned linear systems the iterand x_n is not even a reasonable approximation of $\hat{x}$. For this reason cg became quite unpopular. It was Reid [7] who brought the method back to the attention of numerical analists. For reasonably well conditioned systems cg, when considered as an iterative method, appears to give very satisfactory results after less than n steps. The convergence rate of cg strongly depends on the condition number of the matrix involved. Therefore in practice one uses cg in combination with a preconditioning method. We will not discuss this here.

Although it turns out that for ill conditioned systems x_n may be a bad approximation of $\hat{x}$, continuing the iteration steps ultimately gives values of x_i that are reasonable approximations of $\hat{x}$ and the recursive residuals r_i even tend to zero.

Up to now, no literature has been published explaining this behaviour. In this report we will prove that in the presence of round-off r_i tends to zero, not only for cg but also for iscg.

Although cg, as a special case of iscg, has stronger algebraic properties than iscg itself, we believe that for ill conditioned systems the numerical behaviour of cg and iscg is very similar, except from the first few steps.

One effect of round-off is that orthogonality relations like $(r_i,p_j) = 0$ $(i > j)$ and $(p_i,Ap_j) = 0$ $(i \neq j)$ are no longer true and that the decay of orthogonality for increasing $|i - j|$ destroys the stronger algebraic properties that are based on induction arguments. However, neither of the relations of theorems 2.10 and 2.12 depend on any inductive hypothesis for their validity and therefore we may expect them to hold quite accurately even in the presence of round-off. Stated differently, the approximate validity of these relations is not affected by the loss of orthogonality and hence we may expect that the linear convergence of exact iscg is not disturbed drastically by rounding errors.

3.2. Round-off error analysis

In this section we will investigate the numerical counterparts of several of the algebraic relations of iscg, mentioned in section 2.3. Especially

we are interested in (2.16) , (2.29) and (2.30), since these are the key-points for the proof of theorem 2.12.
We will closely follow the lines of the proof of theorem 2.10. The capital characters D, E, F and G, appearing in the error analysis, will always refer to matrices describing particular computations as mentioned in section 1.3 . By a_i, b_i, r_i, r_{i+1}, p_i, p_{i+1} we will always indicate the numbers and vectors as they are computed and stored by iscg. For clearness' sake, (r_i,p_i) is the exact inner product of the stored vectors r_i and p_i, where as $fl((r_i,p_i))$ denotes the computed value of this inner product. In the formulation of the lemma's and theorems we will not always mention the restriction that r_i and p_i are supposed to be nonzero during the computations.
We are primarily interested in studying how the matrix condition number κ influences the various error estimates. We did not make much effort to determine the smallest possible numbers appearing as numerical factors in the various bounds. Surely many of them can easily be lowered.
In the whole error analysis that will be carried out in this section, we have not ignored terms of any order in ε.
We are now ready to prove

Theorem 1.

Consider iscg with arbitrary initial vectors $r_0,p_0 \neq 0$.
Suppose

(1) $\quad 16\varepsilon(C_1 + 2C_2 + 1)\kappa < 1$.

Then for $i \geq 0$:

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

(3) $\quad \|A^{\frac{1}{2}}p_{i+1}\|^2 = \{1 - (r_{i+1},Ap_i)^2/(\|A^{\frac{1}{2}}p_i\| \, \|A^{\frac{1}{2}}r_{i+1}\|)^2 + \rho_{i+1}\}\|A^{\frac{1}{2}}r_{i+1}\|^2$,

(4) $\quad (r_{i+1},p_{i+1}) = (1 + \lambda_{i+1})\|r_{i+1}\|^2$,

where

(5) $\quad |\mu_i| \leq \varepsilon(13C_1 + 3C_2 + 38)\kappa$,

(6) $\quad |\rho_{i+1}| \leq \varepsilon(C_1 + 2C_2 + 25)\kappa$,

(7) $$|\lambda_{i+1}| \leq \varepsilon(8C_1 + 12C_2 + 64)\kappa^{3/2}\|A^{-1/2}r_i\|/\|A^{-1/2}r_{i+1}\| .$$

Proof. The three combinations (2)(5), (3)(6) and (4)(7) will be proved in the three separate parts I, II and III.

Part I.

We first consider the computation of a_i.

(8) $$fl((r_i,p_i)) = ((I + D_i')r_i,p_i) = (r_i,p_i) + \alpha_i ,$$

where

(9) $$|\alpha_i| = |(D_i'r_i,p_i)| \leq \|D_i'\|\,\|r_i\|\,\|p_i\| \leq \varepsilon C_2\|r_i\|\,\|p_i\| \leq$$

$$\leq \varepsilon C_2\kappa^{\frac{1}{2}}\|A^{-\frac{1}{2}}r_i\|\,\|A^{\frac{1}{2}}p_i\| .$$

Further we have

(10) $$fl((p_i,Ap_i)) = ((I + D_i'')p_i,(A + E_i)p_i) = (p_i,Ap_i) + \beta_i ,$$

where

(11) $$|\beta_i| = |(D_i''p_i,Ap_i) + (p_i,E_ip_i) + (D_i''p_i,E_ip_i)| \leq$$

$$\leq \varepsilon C_2\|p_i\|\,\|Ap_i\| + \varepsilon C_1\|A\|\,\|p_i\|^2 + \varepsilon^2 C_1C_2\|A\|\,\|p_i\|^2 \leq$$

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

We used the fact that from (1) it follows that $\varepsilon C_1 \leq 1$.
So finally

(12) $$a_i = fl\left(\frac{fl(r_i,p_i)}{fl(p_i,Ap_i)}\right) = \left(\frac{(r_i,p_i) + \alpha_i}{(p_i,Ap_i) + \beta_i}\right)(1 + \gamma_i) ,$$

where

(13) $$|\gamma_i| \leq \varepsilon .$$

Hence

(14) $$a_i = (r_i,p_i)/(p_i,Ap_i) + \delta a_i ,$$

where δa_i satisfies

$$(15) \qquad (p_i,Ap_i)\delta a_i = \{\alpha_i+\gamma_i(r_i,p_i)-\beta_i(r_i,p_i)/(p_i,Ap_i)+\alpha_i\gamma_i\}/\{1+\beta_i/(p_i,Ap_i)\} .$$

From (1) and (11) it follows that

$$(16) \qquad |\beta_i/(p_i,Ap_i)| \le \varepsilon(C_1 + 2C_2)\kappa \le \tfrac{1}{2} .$$

Since $|(r_i,p_i)| \le \|A^{-\frac{1}{2}}r_i\| \, \|A^{\frac{1}{2}}p_i\|$ we find from (1), (9), (11), (13), (15) and (16)

$$(17) \qquad |(p_i,Ap_i)\delta a_i| \le 2\varepsilon(C_1 + 3C_2 + 2)\kappa\|A^{-\frac{1}{2}}r_i\| \, \|A^{\frac{1}{2}}p_i\| .$$

Note that by a similar derivation we can find the upperbound

$$(18) \qquad |(p_i,Ap_i)\delta a_i| \le 2\varepsilon(C_1 + 3C_2 + 2)\kappa\|r_i\| \, \|p_i\| .$$

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

$$(19) \qquad (I + G'_{i+1})r_{i+1} = r_i - (I + F'_i)a_i(A + E_i)p_i ,$$

Hence

$$(20) \qquad r_{i+1} = r_i - a_iAp_i + \delta r_{i+1} ,$$

where

$$(21) \qquad \|\delta r_{i+1}\| \le \|a_iE_ip_i\| + \|a_iF'_iAp_i\| + \|G'_{i+1}r_{i+1}\| + \|a_iF'_iE_ip_i\| \le$$

$$\le \varepsilon C_1\|A\| \, \|a_ip_i\| + \varepsilon\|a_iAp_i\| + \varepsilon\|r_{i+1}\| + \varepsilon^2C_1\|A\| \, \|a_ip_i\| \le$$

$$\le \varepsilon\|A^{\frac{1}{2}}\|(\|A^{-\frac{1}{2}}r_{i+1}\| + (C_1 + 2)\kappa^{\frac{1}{2}}\|a_iA^{\frac{1}{2}}p_i\|) \le$$

$$\le \varepsilon(\|A^{-\frac{1}{2}}r_{i+1}\| + (C_1 + 2)\|a_iA^{\frac{1}{2}}p_i\|)\kappa^{\frac{1}{2}}\|A^{\frac{1}{2}}\| .$$

We are now ready to prove (2)(5).
It follows from (20) that

$$(22) \qquad A^{-\frac{1}{2}}r_{i+1} + a_iA^{\frac{1}{2}}p_i - A^{-\frac{1}{2}}\delta r_{i+1} = A^{-\frac{1}{2}}r_i ,$$

and hence, by taking squared norms of left and right hand sides we get

$$(23) \qquad \|A^{-\frac{1}{2}}r_{i+1}\|^2 + \|a_iA^{\frac{1}{2}}p_i\|^2 = \|A^{-\frac{1}{2}}r_i\|^2 +$$

$$- 2a_i(r_{i+1},p_i) + 2(\delta r_{i+1},A^{-1}r_{i+1}) + 2a_i(\delta r_{i+1},p_i) - \|A^{-\frac{1}{2}}\delta r_{i+1}\|^2 .$$

From (14) and (20) we get

$$(24)\qquad a_i(r_{i+1},p_i) = a_i(r_i,p_i) - a_i^2(p_i,Ap_i) + a_i(\delta r_{i+1},p_i) =$$

$$= -(r_i,p_i)\delta a_i - (p_i,Ap_i)(\delta a_i)^2 + a_i(\delta r_{i+1},p_i) .$$

Substitution in (23) gives the counterpart of (2.17) in the presence of round off:

$$(25)\qquad \|A^{-\frac{1}{2}}r_{i+1}\|^2 + \|a_iA^{\frac{1}{2}}p_i\|^2 = \|A^{-\frac{1}{2}}r_i\|^2 +$$

$$+ 2\{(r_i,p_i)\delta a_i + (p_i,Ap_i)(\delta a_i)^2 + (\delta r_{i+1},A^{-1}r_{i+1})\} - \|A^{-\frac{1}{2}}\delta r_{i+1}\|^2.$$

Bringing $\|a_iA^{\frac{1}{2}}p_i\|^2$ to the right hand side and substituting (14) for a_i gives the counterpart of (2.29):

$$(26)\qquad \|A^{-\frac{1}{2}}r_{i+1}\|^2 = \{1 - (r_i,p_i)^2/(\|A^{\frac{1}{2}}p_i\|\,\|A^{-\frac{1}{2}}r_i\|)^2 + \mu_i\}\|A^{-\frac{1}{2}}r_i\|^2 ,$$

where

$$(27)\qquad \mu_i = [(p_i,Ap_i)(\delta a_i)^2 + 2(\delta r_{i+1},A^{-1}r_{i+1}) - \|A^{-\frac{1}{2}}\delta r_{i+1}\|^2]/\|A^{-\frac{1}{2}}r_i\|^2 .$$

From (17) we get

$$(28)\qquad (p_i,Ap_i)(\delta a_i)^2 = |(p_i,Ap_i)\delta a_i|^2/\|A^{\frac{1}{2}}p_i\|^2 \le$$

$$\le 4\varepsilon^2(C_1 + 3C_2 + 2)^2\kappa^2\|A^{-\frac{1}{2}}r_i\|^2 \le \varepsilon(C_1 + 3C_2 + 2)\kappa\|A^{-\frac{1}{2}}r_i\|^2 ,$$

since it follows form (1) that $4\varepsilon(C_1 + 3C_2 + 2)\kappa < 1$.
From (21) we get

$$(29)\qquad (\delta r_{i+1},A^{-1}r_{i+1}) \le \|A^{-\frac{1}{2}}\|\,\|\delta r_{i+1}\|\,\|A^{-\frac{1}{2}}r_{i+1}\| \le$$

$$\le \varepsilon(\|A^{-\frac{1}{2}}r_{i+1}\|^2 + (C_1 + 2)\|a_iA^{\frac{1}{2}}p_i\|\,\|A^{-\frac{1}{2}}r_{i+1}\|)\kappa$$

and

$$(30)\qquad \|A^{-\frac{1}{2}}\delta r_{i+1}\|^2 \le \varepsilon^2(\|A^{-\frac{1}{2}}r_{i+1}\| + (C_1 + 2)\|a_iA^{\frac{1}{2}}p_i\|)^2\kappa^2 \le$$

$$\le 2\varepsilon^2(\|A^{-\frac{1}{2}}r_{i+1}\|^2 + (C_1 + 2)^2\|a_iA^{\frac{1}{2}}p_i\|^2)\kappa^2 \le$$

$$\le \varepsilon(\|A^{-\frac{1}{2}}r_{i+1}\|^2 + (C_1 + 2)\|a_iA^{\frac{1}{2}}p_i\|^2)\kappa ,$$

since it follows form (1) that $2\varepsilon\kappa < 1$ and $2\varepsilon(C_1 + 2)\kappa < 1$. In order to determine a bound for $|\mu_i|$ we need to bound $\| A^{-\frac{1}{2}} r_{i+1} \|$ and $\| a_i A^{\frac{1}{2}} p_i \|$ in terms of $\| A^{-\frac{1}{2}} r_i \|$.

From (17) it follows that

$$(31) \qquad |(r_i,p_i)\delta a_i| = |(A^{-\frac{1}{2}} r_i, A^{\frac{1}{2}} p_i)|\,|(p_i,Ap_i)\delta a_i|/\| A^{\frac{1}{2}} p_i \|^2 \leq$$

$$\leq 2\varepsilon(C_1 + 3C_2 + 2)\kappa \| A^{-\frac{1}{2}} r_i \|^2 .$$

Substitution of (3.35), (3.36) and (3.38) in (3.32) gives

$$(32) \qquad \| A^{-\frac{1}{2}} r_{i+1} \|^2 + \| a_i A^{\frac{1}{2}} p_i \|^2 \leq \| A^{-\frac{1}{2}} r_i \|^2 + 6\varepsilon(C_1 + 3C_2 + 2)\kappa \| A^{-\frac{1}{2}} r_i \|^2 +$$

$$+ 2\varepsilon\kappa \| A^{-\frac{1}{2}} r_{i+1} \|^2 + 2\varepsilon(C_1 + 2)\kappa \| a_i A^{\frac{1}{2}} p_i \|\, \| A^{-\frac{1}{2}} r_{i+1} \|$$

or

$$(33) \qquad \| A^{-\frac{1}{2}} r_{i+1} \|^2 (1 - 2\varepsilon\kappa) - 2\varepsilon(C_1 + 2)\kappa \| a_i A^{\frac{1}{2}} p_i \|\, \| A^{-\frac{1}{2}} r_{i+1} \| + \| a_i A^{\frac{1}{2}} p_i \|^2 \leq$$

$$\leq \{1 + 6\varepsilon(C_1 + 3C_2 + 2)\kappa\} \| A^{-\frac{1}{2}} r_i \|^2$$

Since, from (1) we have $2\varepsilon\kappa < \frac{1}{4}$, $2\varepsilon(C_1 + 2)\kappa < \frac{1}{2}$ and $6\varepsilon(C_1 + 3C_2 + 2)\kappa < 1$ it follows that

$$(34) \qquad 3\| A^{-\frac{1}{2}} r_{i+1} \|^2 - 2\| a_i A^{\frac{1}{2}} p_i \|\, \| A^{-\frac{1}{2}} r_{i+1} \| + 3\| a_i A^{\frac{1}{2}} p_i \|^2 \leq 8\| A^{-\frac{1}{2}} r_i \|^2$$

and from this quadratic inequality it easily follows that

$$(35) \qquad \| a_i A^{\frac{1}{2}} p_i \| \leq 2\| A^{-\frac{1}{2}} r_i \|$$

and

$$(36) \qquad \| A^{-\frac{1}{2}} r_{i+1} \| \leq 2\| A^{-\frac{1}{2}} r_i \| .$$

From (28), (29), (30), (35) and (36) we find for (27):

$$(37) \qquad |\mu_i| \leq \varepsilon\{(C_1 + 3C_2 + 2) + 2(4 + 4(C_1 + 2)) + (4 + 4(C_1 + 2))\}\kappa \leq$$

$$\leq \varepsilon(13C_1 + 3C_2 + 38)\kappa ,$$

which proves (5).

Part II.

The now following proof of (3)(6) is entirely analogous to the proof of (2)(5) given above.

We first consider the computation of b_i.

$$(38) \qquad fl((r_{i+1},Ap_i)) = ((I + D_i''')r_{i+1},(A + E_i)p_i) = (r_{i+1},Ap_i) + \tau_i$$

where

$$(39) \qquad |\tau_i| = |(D_i''' r_{i+1},Ap_i) + (r_{i+1},E_ip_i) + (D_i''' r_{i+1},E_ip_i)| \leq$$

$$\leq \varepsilon C_2 \| r_{i+1} \| \| Ap_i \| + \varepsilon C_1 \| A \| \| r_{i+1} \| \| p_i \| + \varepsilon^2 C_1 C_2 \| A \| \| r_{i+1} \| \| p_i \| \leq$$

$$\leq \varepsilon(C_1 + 2C_2)\| A \| \| r_{i+1} \| \| p_i \| \leq$$

$$\leq \varepsilon(C_1 + 2C_2)\kappa \| A^{\frac{1}{2}} r_{i+1} \| \| A^{\frac{1}{2}} p_i \| .$$

Hence, with (10)

$$(40) \qquad b_i = - fl\left(\frac{fl(r_{i+1},Ap_i)}{fl(p_i,Ap_i)}\right) = -\left(\frac{(r_{i+1},Ap_i) + \tau_i}{(p_i,Ap_i) + \beta_i}\right)(1 + \nu_i)$$

where

$$(41) \qquad |\nu_i| \leq \varepsilon$$

From this it follows that

$$(42) \qquad b_i = -(r_{i+1},Ap_i)/(p_i,Ap_i) - \delta b_i ,$$

where δb_i satisfies

$$(43) \qquad (p_i,Ap_i)\delta b_i = \{\tau_i + \nu_i(r_{i+1},Ap_i)\tau_i\nu_i +$$

$$- \beta_i(r_{i+1},Ap_i)/(p_i,Ap_i)\}/\{1 + \beta_i/(p_i,Ap_i)\} .$$

Since $|(r_{i+1},Ap_i)| \leq \| A^{\frac{1}{2}} r_{i+1} \| \| A^{\frac{1}{2}} p_i \|$ we find from (1), (11), (16), (39), (41) and (43)

$$(44) \qquad |(p_i,Ap_i)\delta b_i| \leq 4\varepsilon(C_1 + 2C_2 + 1)\kappa \| A^{\frac{1}{2}} r_{i+1} \| \| A^{\frac{1}{2}} p_i \| .$$

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

$$(45) \qquad (I + G_{i+1}'')p_{i+1} = r_{i+1} + (I + F_i'')b_ip_i .$$

Hence

$$p_{i+1} = r_{i+1} + b_i p_i + \delta p_{i+1} \,, \tag{46}$$

where

$$\|\delta p_{i+1}\| \le \|b_i F_i'' p_i\| + \|G_{i+1}'' p_{i+1}\| \le \varepsilon\|b_i p_i\| + \varepsilon\|p_{i+1}\| \le \tag{47}$$

$$\le \varepsilon(\|b_i A^{\frac{1}{2}} p_i\| + \|A^{\frac{1}{2}} p_{i+1}\|)\|A^{-\frac{1}{2}}\| \,.$$

We are now ready to prove (3) and (6).
It follows from (46) that

$$A^{\frac{1}{2}} p_{i+1} - b_i A^{\frac{1}{2}} p_i - A^{\frac{1}{2}} \delta p_{i+1} = A^{\frac{1}{2}} r_{i+1} \,, \tag{48}$$

and hence, by taking squared norms of left and right hand sides we get

$$\|A^{\frac{1}{2}} p_{i+1}\|^2 + \|b_i A^{\frac{1}{2}} p_i\|^2 = \|A^{\frac{1}{2}} r_{i+1}\|^2 + 2b_i (p_{i+1}, Ap_i) + \tag{49}$$

$$+ 2(p_{i+1}, A\delta p_{i+1}) - 2b_i (\delta p_{i+1}, Ap_i) - \|A^{\frac{1}{2}} \delta p_{i+1}\|^2 \,.$$

From (42) and (46) we get

$$b_i (p_{i+1}, Ap_i) = b_i (r_{i+1}, Ap_i) + b_i^2 (p_i, Ap_i) + b_i (\delta p_{i+1}, Ap_i) = \tag{50}$$

$$= (r_{i+1}, Ap_i)\delta b_i + (p_i, Ap_i)(\delta b_i)^2 + b_i (\delta p_{i+1}, Ap_i) \,.$$

Substitution in (49) gives the counterpart of (2.19) in the presence of round-off:

$$\|A^{\frac{1}{2}} p_{i+1}\|^2 + \|b_i A^{\frac{1}{2}} p_i\|^2 = \|A^{\frac{1}{2}} r_{i+1}\|^2 + \tag{51}$$

$$+ 2\{(r_{i+1}, Ap_i)\delta b_i + (p_i, Ap_i)(\delta b_i)^2 + (\delta p_{i+1}, Ap_{i+1})\} +$$

$$- \|A^{\frac{1}{2}} \delta p_{i+1}\|^2 \,.$$

Bringing $\|b_i A^{\frac{1}{2}} p_i\|^2$ to the right hand side and substituting (42) for b_i gives the counterpart of (2.30):

$$\|A^{\frac{1}{2}} p_{i+1}\|^2 = \{1 - (r_{i+1}, Ap_i)^2 / (\|A^{\frac{1}{2}} p_i\| \, \|A^{\frac{1}{2}} r_{i+1}\|)^2 + \rho_{i+1}\} \|A^{\frac{1}{2}} r_{i+1}\|^2 \tag{52}$$

where

$$\rho_{i+1} = [(p_i, Ap_i)(\delta b_i)^2 + 2(\delta p_{i+1}, Ap_{i+1}) - \|A^{\frac{1}{2}} \delta p_{i+1}\|^2] / \|A^{\frac{1}{2}} r_{i+1}\|^2 \,. \tag{53}$$

From (44) we get

$$(54)\qquad (p_i,Ap_i)(\delta b_i)^2 = |(p_i,Ap_i)\delta b_i|^2/\|A^{\frac{1}{2}}p_i\|^2 \le$$

$$\le 16\varepsilon^2(C_1 + 2C_2 + 1)^2\kappa^2\|A^{\frac{1}{2}}r_{i+1}\|^2 \le$$

$$\le \varepsilon(C_1 + 2C_2 + 1)\kappa\|A^{\frac{1}{2}}r_{i+1}\|^2 ,$$

since it follows from (1) that $16\varepsilon(C_1 + 2C_2 + 1)\kappa < 1$.

From (47) we get

$$(55)\qquad |(\delta p_{i+1},Ap_{i+1})| \le \|A^{\frac{1}{2}}\|\,\|\delta p_{i+1}\|\,\|A^{\frac{1}{2}}p_{i+1}\| \le$$

$$\le \varepsilon(\|A^{\frac{1}{2}}p_{i+1}\|^2 + \|b_iA^{\frac{1}{2}}p_i\|\,\|A^{\frac{1}{2}}p_{i+1}\|)\kappa^{\frac{1}{2}}$$

and

$$(56)\qquad \|A^{\frac{1}{2}}\delta p_{i+1}\|^2 \le \varepsilon^2(\|b_iA^{\frac{1}{2}}p_i\| + \|A^{\frac{1}{2}}p_{i+1}\|)^2\kappa \le$$

$$\le 2\varepsilon^2(\|b_iA^{\frac{1}{2}}p_i\|^2 + \|A^{\frac{1}{2}}p_{i+1}\|^2)\kappa \le$$

$$\le \varepsilon(\|b_iA^{\frac{1}{2}}p_i\|^2 + \|A^{\frac{1}{2}}p_{i+1}\|^2)$$

since it follows from (1) that $2\varepsilon\kappa < 1$.

In order to determine a bound for $|\rho_{i+1}|$ we need to bound $\|b_iA^{\frac{1}{2}}p_i\|$ and $\|A^{\frac{1}{2}}p_{i+1}\|$ in terms of $\|A^{\frac{1}{2}}r_{i+1}\|$.

From (44) it follows that

$$(57)\qquad |(r_{i+1},Ap_i)\delta b_i| = |(A^{\frac{1}{2}}r_{i+1},A^{\frac{1}{2}}p_i)|\,|(p_i,Ap_i)\delta b_i|/\|A^{\frac{1}{2}}p_i\|^2 \le$$

$$\le 4\varepsilon(C_1 + 2C_2 + 1)\kappa\|A^{\frac{1}{2}}r_{i+1}\|^2 .$$

Substitution of (54), (55) and (57) in (51) gives

$$(58)\qquad \|A^{\frac{1}{2}}p_{i+1}\|^2 + \|b_iA^{\frac{1}{2}}p_i\|^2 \le \|A^{\frac{1}{2}}r_{i+1}\|^2 + 8\varepsilon(C_1 + 2C_2 + 1)\kappa\|A^{\frac{1}{2}}r_{i+1}\|^2 +$$

$$+ 2\varepsilon(C_1 + 2C_2 + 1)\kappa\|A^{\frac{1}{2}}r_{i+1}\|^2 + 2\varepsilon\kappa^{\frac{1}{2}}\|A^{\frac{1}{2}}p_{i+1}\|^2 +$$

$$+ 2\varepsilon\kappa^{\frac{1}{2}}\|b_iA^{\frac{1}{2}}p_i\|\,\|A^{\frac{1}{2}}p_{i+1}\|\kappa^{\frac{1}{2}}$$

or

$$(59)\qquad \|A^{\frac{1}{2}}p_{i+1}\|^2(1 - 2\varepsilon\kappa^{\frac{1}{2}}) - 2\varepsilon\kappa^{\frac{1}{2}}\|A^{\frac{1}{2}}p_{i+1}\|\,\|b_iA^{\frac{1}{2}}p_i\| + \|b_iA^{\frac{1}{2}}p_i\|^2 \le$$

$$\le \{1 + 10\varepsilon(C_1 + 2C_2 + 1)\kappa\}\|A^{\frac{1}{2}}r_{i+1}\|^2 ,$$

and just like for (33) it follows that

$$(60)\qquad \|A^{\frac{1}{2}}p_{i+1}\| \le 2\|A^{\frac{1}{2}}r_{i+1}\|$$

and

$$(61)\qquad \|b_iA^{\frac{1}{2}}p_i\| \le 2\|A^{\frac{1}{2}}r_{i+1}\| .$$

From (54), (56), (60) and (61) we find for (53)

$$(62)\qquad |\rho_{i+1}| \le \varepsilon\{(C_1 + 2C_2 + 1) + 16 + 8\}\kappa \le \varepsilon(C_1 + 2C_2 + 25)\kappa ,$$

which proves (6).

Part III.

Now we finally prove (4)(7).

From (46) we get

$$(63)\qquad (r_{i+1},p_{i+1}) = (r_{i+1},r_{i+1}) + b_i(r_{i+1},p_i) + (r_{i+1},\delta p_{i+1}) .$$

From (14) and (20) we get

$$(64)\qquad b_i(r_{i+1},p_i) = b_i\{(r_i,p_i) - a_i(p_i,Ap_i) + (\delta r_{i+1},p_i)\}$$

$$= -b_i(p_i,Ap_i)\delta a_i + b_i(\delta r_{i+1},p_i) .$$

Substitution in (63) gives

$$(65)\qquad (r_{i+1},p_{i+1}) = (r_{i+1},r_{i+1}) - b_i(p_i,Ap_i)\delta a_i + b_i(\delta r_{i+1},p_i) + (r_{i+1},\delta p_{i+1})$$

$$= (1 + \lambda_{i+1})(r_{i+1},r_{i+1})$$

where

$$(66)\qquad \lambda_{i+1} = \{-b_i(p_i,Ap_i)\delta a_i + b_i(\delta r_{i+1},p_i) + (r_{i+1},\delta p_{i+1})\}/\|r_{i+1}\|^2 .$$

From (17) and (61) we get

$$(67)\qquad |b_i(p_i,Ap_i)\delta a_i| \le 2\varepsilon(C_1 + 3C_2 + 2)\kappa\|A^{-\frac{1}{2}}r_i\|\,\|b_iA^{\frac{1}{2}}p_i\| \le$$

$$\le 4\varepsilon(C_1 + 3C_2 + 2)\kappa\|A^{-\frac{1}{2}}r_i\|\,\|A^{\frac{1}{2}}r_{i+1}\| .$$

From (21), (35), (36) and (61) we get

$$(68) \qquad |b_i(\delta r_{i+1}, p_i)| \le \|A^{-\frac{1}{2}}\| \, \|\delta r_{i+1}\| \, \|b_i A^{\frac{1}{2}} p_i\| \le$$

$$\le 4\varepsilon(C_1 + 12)\kappa \|A^{-\frac{1}{2}} r_i\| \, \|A^{\frac{1}{2}} r_{i+1}\| \, .$$

From (36), (47), (60) and (61) we get

$$(69) \qquad |(r_{i+1}, \delta p_{i+1})| \le \|r_{i+1}\| \, \|\delta p_{i+1}\| \le 4\varepsilon \|A^{-\frac{1}{2}}\| \, \|A^{\frac{1}{2}} r_{i+1}\| \, \|r_{i+1}\| \le$$

$$\le 4\varepsilon\kappa^{\frac{1}{2}} \|A^{\frac{1}{2}} r_{i+1}\| \, \|A^{-\frac{1}{2}} r_{i+1}\| \le 8\varepsilon\kappa^{\frac{1}{2}} \|A^{\frac{1}{2}} r_{i+1}\| \, \|A^{-\frac{1}{2}} r_i\| \, .$$

Substitution of (67), (68) and (69) in (66) gives

$$(70) \qquad |\lambda_{i+1}| \le \varepsilon(8C_1 + 12C_2 + 64)\kappa \|A^{\frac{1}{2}} r_{i+1}\| \, \|A^{-\frac{1}{2}} r_i\| / \|r_{i+1}\|^2 \le$$

$$\le \varepsilon(8C_1 + 12C_2 + 64)\kappa^{3/2} \|A^{-\frac{1}{2}} r_i\| / \|A^{-\frac{1}{2}} r_{i+1}\| \, . \qquad \square$$

Remark 2.

If we would have ignored terms of order ε^2 in the presence of terms of order ε then, instead of (53) we would have taken

$$(71) \qquad \rho_{i+1} = 2(\delta p_{i+1}, Ap_{i+1}) / \|A^{\frac{1}{2}} r_{i+1}\|^2$$

and from (55), (60) and (61) it then follows that

$$(72) \qquad |\rho_{i+1}| \le 16\varepsilon\kappa^{\frac{1}{2}} \, .$$

This upperbound differs from (6) by a factor of order $\kappa^{\frac{1}{2}}$. The difference is caused by the fact that the second order terms in (53) are of order $\varepsilon^2\kappa^2$ which give order $\varepsilon\kappa$ under the assumption $\varepsilon\kappa < 1$. $\square$

Remark 3.

It follows from (2)(5) and (3)(6) that (2.29) and (2.30) hold quite well in the presence of round-off. Especially it follows from (2)(5) that $\|A^{-\frac{1}{2}} r_{i+1}\| / \|A^{-\frac{1}{2}} r_i\|$ never exceeds $1 + \varepsilon(13C_1 + 3C_2 + 38)\kappa$.
From (4)(7) it follows that in the presence of round-off (2.16) may be seriously perturbed if $\|A^{-\frac{1}{2}} r_i\| / \|A^{-\frac{1}{2}} r_{i+1}\|$ is large. Stated differently, relation (2.16) holds reasonably well unless $\|A^{-\frac{1}{2}} r_{i+1}\| << \|A^{-\frac{1}{2}} r_i\|$. $\square$

4. The convergence of r_i

4.1. Introduction

Considering the proof of theorem (2.12) we may expect that in the presence of round-off the approximate validity of (2.16) expressed in terms of r_i and p_i instead of r_{i+1} and p_{i+1}, and the approximate validity of (2.29 and (2.30) will imply the approximate validity of (2.32) and consequently the approximate validity of (2.34) . Together with remark (3.3) this gives the basic idea for the proof of the convergence of r_i: $\|A^{-\frac{1}{2}}r_{i+1}\|/\|A^{-\frac{1}{2}}r_i\|$ is less than $(\kappa - 1)/(\kappa + 1)$ unless $\|A^{-\frac{1}{2}}r_i\|/\|A^{-\frac{1}{2}}r_{i-1}\|$ is very small, or, stated more precisely, if the natural error vector does not decrease by the expected rate in step i+1 then it did decrease at least by the square of the expected rate in step i and i+1 together.
Since, clearly, $\|A^{-\frac{1}{2}}r_{i+1}\|/\|A^{-\frac{1}{2}}r_i\|$ depends on $\|A^{-\frac{1}{2}}r_i\|/\|A^{-\frac{1}{2}}r_{i-1}\|$, we first consider two successive steps of iscg.
From this we will prove the linear convergence of iscg. Once more we mention the fact that for simplicity we often estimate rather roughly the factor appearing in the various bounds and that we ignore the possibility of underflow and overflow. Again in lemma's and theorems we suppress mentioning the fact that we suppose that no $r_i = 0$ or $p_i = 0$.
The capital characters D, E, F and G and the symbols a_i, b_i, r_i, r_{i+1}, p_i and p_{i+1} are used under the same conventions as mentioned in section (3.2). No terms in ε are ignored.

4.2. Two steps of iscg

The influence of the rate $\|A^{-\frac{1}{2}}r_1\|/\|A^{-\frac{1}{2}}r_0\|$ on the rate $\|A^{-\frac{1}{2}}r_2\|/\|A^{-\frac{1}{2}}r_1\|$ is expressed by the following lemma.

Lemma 1. Let $16\varepsilon(C_1 + 2C_2 + 1)\kappa < 1$ and consider two steps of iscg with arbitrary initial vectors $r_0, p_0 \neq 0$. Then we have

$$(1) \qquad \|A^{-\frac{1}{2}}r_2\|^2 \leq \{(\kappa - 1)^2/(\kappa + 1)^2 + \gamma_1\}\|A^{-\frac{1}{2}}r_1\|^2 ,$$

where

$$(2) \qquad |\gamma_1| \leq \varepsilon(C_1 + 2C_2 + 8)(64\kappa^{\frac{1}{2}}\|A^{-\frac{1}{2}}r_0\|/\|A^{-\frac{1}{2}}r_1\| + 14\kappa) .$$

Proof. From (3.2) we know

$$\|A^{-\frac{1}{2}}r_2\|^2/\|A^{-\frac{1}{2}}r_1\|^2 = 1 - (r_1,p_1)^2/(\|A^{\frac{1}{2}}p_1\|\,\|A^{-\frac{1}{2}}r_1\|)^2 + \mu_1 \,, \tag{3}$$

and

$$|\mu_1| \le \varepsilon(13C_1 + 3C_2 + 38)\kappa \,. \tag{4}$$

From (3.3) we get

$$\|A^{\frac{1}{2}}p_1\|^2 \le \|A^{\frac{1}{2}}r_1\|^2(1 + \rho_1) \,, \tag{5}$$

and

$$|\rho_1| \le \varepsilon(C_1 + 2C_2 + 25)\kappa \,. \tag{6}$$

Together with (3.4 . this gives

$$\left(\frac{(r_1,p_1)}{\|A^{\frac{1}{2}}p_1\|\,\|A^{-\frac{1}{2}}r_1\|}\right)^2 \ge \frac{\|r_1\|^4(1 + \lambda_1)^2}{\|A^{\frac{1}{2}}r_1\|^2\|A^{-\frac{1}{2}}r_1\|^2(1 + \rho_1)} \,. \tag{7}$$

Hence, using the Kantorovich inequality (2.33)

$$\left(\frac{(r_1,p_1)}{\|A^{\frac{1}{2}}p_1\|\,\|A^{-\frac{1}{2}}r_1\|}\right)^2 \ge \frac{4\kappa(1 + \lambda_1)^2}{(\kappa + 1)^2(1 + \rho_1)} = \frac{4\kappa(1 + \lambda_1)^2}{(\kappa + 1)^2} + \varphi_1 \,, \tag{8}$$

where

$$\varphi_1 := \frac{-4\kappa(1 + \lambda_1)^2\rho_1}{(\kappa + 1)^2(1 + \rho_1)} \,. \tag{9}$$

But since $\left(\frac{(r_1,p_1)}{\|A^{\frac{1}{2}}p_1\|\,\|A^{-\frac{1}{2}}r_1\|}\right)^2 \le 1$, also $\frac{4\kappa(1 + \lambda_1)^2}{(\kappa + 1)^2(1 + \rho_1)} \le 1$ and consequently

$$|\varphi_1| \le |\rho_1| \le \varepsilon(C_1 + 2C_2 + 25)\kappa \,. \tag{10}$$

Substitution of (8) in (3) yields

$$\frac{\|A^{-\frac{1}{2}}r_2\|^2}{\|A^{-\frac{1}{2}}r_1\|^2} \le 1 - \frac{4\kappa(1 + \lambda_1)^2}{(\kappa + 1)^2} + |\psi_1| \le (\kappa - 1)^2/(\kappa + 1)^2 + 8|\lambda_1|/\kappa + |\psi_1| \tag{11}$$

where

$$|\psi_1| = |\mu_1 - \varphi_1| \le |\mu_1| + |\varphi_1| \le 7\varepsilon(2C_1 + C_2 + 9)\kappa \ . \tag{12}$$

From (3.7) we have

$$|\lambda_1| \le \varepsilon(8C_1 + 12C_2 + 64)\kappa^{3/2}\|A^{-\frac{1}{2}}r_0\|/\|A^{-\frac{1}{2}}r_1\| \ , \tag{13}$$

and consequently

$$8|\lambda_1|\kappa^{-1} \le 32\varepsilon(2C_1 + 4C_2 + 16)\kappa^{\frac{1}{2}}\|A^{-\frac{1}{2}}r_0\|/\|A^{-\frac{1}{2}}r_1\| \ . \tag{14}$$

Now (1), (2) follows from (11), (12) and (14). □

Theorem 2.

Consider two steps of iscg with arbitrary initial vectors $r_0, p_0 \neq 0$. Let $0 < \theta < 1$ and

$$L_\theta := \{\theta + (1 - \theta)((\kappa - 1)/(\kappa + 1))^2\}^{\frac{1}{2}} \ . \tag{15}$$

If

$$106\varepsilon(C_1 + 2C_2 + 8)\kappa^2 \le \theta L_\theta^2 \tag{16}$$

then at least one of the following two inequalities is true:

$$\|A^{-\frac{1}{2}}r_2\| \le L_\theta\|A^{-\frac{1}{2}}r_1\| \ , \tag{17}$$

$$\|A^{-\frac{1}{2}}r_2\| \le L_\theta^2\|A^{-\frac{1}{2}}r_0\| \ . \tag{18}$$

Proof. Since $\theta L_\theta^2 < 1$ the restriction of lemma (1) certainly is satisfied and from (3.2) it follows that

$$\|A^{-\frac{1}{2}}r_2\|^2/\|A^{-\frac{1}{2}}r_1\|^2 \le 1 + |\delta_1| \le 2 \ . \tag{19}$$

If $\|A^{-\frac{1}{2}}r_2\|^2/\|A^{-\frac{1}{2}}r_0\|^2 \le L_\theta^4$ then we are ready. If $\|A^{-\frac{1}{2}}r_2\|^2/\|A^{-\frac{1}{2}}r_0\|^2 > L_\theta^4$ then it follows from (19) and the fact $\|A^{-\frac{1}{2}}r_2\|^2/\|A^{-\frac{1}{2}}r_0\|^2 = (\|A^{-\frac{1}{2}}r_2\|^2/\|A^{-\frac{1}{2}}r_1\|^2)(\|A^{-\frac{1}{2}}r_1\|^2/\|A^{-\frac{1}{2}}r_0\|^2)$ that

$$\|A^{-\frac{1}{2}}r_1\|^2/\|A^{-\frac{1}{2}}r_0\|^2 > \tfrac{1}{2}L_\theta^4 \ . \tag{20}$$

Substitution in (2) gives

$$(21) \qquad |\gamma_1| \le \varepsilon(C_1 + 2C_2 + 8)(92\kappa^{\frac{1}{2}}L_\theta^{-2} + 14\kappa) \le 106\varepsilon(C_1 + 2C_2 + 8)\kappa L_\theta^{-2} ,$$

since $L_\theta^{-2} > 1$.
Hence, form (16), $|\gamma_1| \le \theta\kappa^{-1}$ and then it finally follows from (1) that

$$(22) \qquad \|A^{-\frac{1}{2}}r_2\|^2 \le \{(\kappa - 1)^2/(\kappa + 1)^2 + \theta\kappa^{-1}\}\|A^{-\frac{1}{2}}r_1\|^2 \le$$

$$\le \{(\kappa - 1)^2/(\kappa + 1)^2 + 4\theta\kappa/(\kappa + 1)^2\}\|A^{-\frac{1}{2}}r_1\|^2$$

$$= L_\theta^2\|A^{-\frac{1}{2}}r_1\|^2 . \qquad \square$$

Remark 3.

Since iscg is a one-step-method we also may conclude from theorem 2 that if $106\varepsilon(C_1 + 2C_2 + 8)\kappa^2 < \theta L_\theta$ then for any $k \ge 1$ at least one of the following two inequalities is true:

$$(23) \qquad \|A^{-\frac{1}{2}}r_{k+1}\| \le L_\theta\|A^{-\frac{1}{2}}r_k\| ,$$

$$(24) \qquad \|A^{-\frac{1}{2}}r_{k+1}\| \le L_\theta^2\|A^{-\frac{1}{2}}r_{k-1}\| .$$

This means that if in a certain step the natural error vector does not decrease by a factor L_θ, then still it did decrease by a factor L_θ^2 in the last two steps together. It is easily seen that the assertion given by (23), (24) is equivalent with the assertion that for every $k \ge 2$

$$\frac{\|A^{-\frac{1}{2}}r_k\|}{\|A^{-\frac{1}{2}}r_{k-2}\|} \le L_\theta^2 \text{ or } \frac{\|A^{-\frac{1}{2}}r_{k+1}\|}{\|A^{-\frac{1}{2}}r_{k-1}\|} \le L_\theta^2 \qquad \square$$

Remark 4.

Note that $L_\theta \to (\kappa - 1)/(\kappa + 1)$ if $\theta \to 0$. This is the algebraic convergence rate of (2.28) Relation (16) shows that it depends on the value of $\varepsilon\kappa^2$ how nearly this theoretical rate of convergence can be reached. $\square$

4.3. The linear convergence of r_i

The linear convergence of r_i is expressed by the following theorem.

Theorem 5.

Consider iscg with arbitrary initial vectors $r_0, p_0 \neq 0$. Let $0 < \theta < 1$ and let L_θ be defined by (15). If

$$106\varepsilon(C_1 + 2C_2 + 8)\kappa^2 \leq \theta L_\theta^2 \tag{25}$$

then we have for $i \geq 0$:

$$\|A^{-\frac{1}{2}} r_{i+1}\| \leq (1 + \varepsilon(13C_1 + 3C_2 + 38)\kappa) L_\theta^i \|A^{-\frac{1}{2}} r_0\| . \tag{26}$$

Proof.

For $i = -1$ inequality (26) is trivially satisfied since $L_\theta < 1$. For $i = 0$ it follows immediately from (3.2)

Now let $k \geq 1$ and suppose (26) holds for all $-1 \leq i \leq k - 1$.

If $\|A^{-\frac{1}{2}} r_{k+1}\| / \|A^{-\frac{1}{2}} r_k\| \leq L_\theta$ then

$$\begin{aligned} \|A^{-\frac{1}{2}} r_{k+1}\| / \|A^{-\frac{1}{2}} r_0\| &= (\|A^{-\frac{1}{2}} r_{k+1}\| / \|A^{-\frac{1}{2}} r_k\|)(\|A^{-\frac{1}{2}} r_k\| / \|A^{-\frac{1}{2}} r_0\|) \leq \\ &\leq L_\theta (1 + \varepsilon(13C_1 + 3C_2 + 38)\kappa) L_\theta^{k-1} \|A^{-\frac{1}{2}} r_0\| = \\ &= (1 + \varepsilon(13C_1 + 3C_2 + 38)) L_\theta^k \|A^{-\frac{1}{2}} r_0\| . \end{aligned} \tag{27}$$

If $\|A^{-\frac{1}{2}} r_{k+1}\| / \|A^{-\frac{1}{2}} r_k\| > L_\theta$ then, from (24), certainly $\|A^{-\frac{1}{2}} r_{k+1}\| / \|A^{-\frac{1}{2}} r_{k-1}\| \leq L_\theta^2$ and therefore

$$\begin{aligned} \|A^{-\frac{1}{2}} r_{k+1}\| / \|A^{-\frac{1}{2}} r_0\| &= (\|A^{-\frac{1}{2}} r_{k+1}\| / \|A^{-\frac{1}{2}} r_{k-1}\|)(\|A^{-\frac{1}{2}} r_{k-1}\| / \|A^{-\frac{1}{2}} r_0\|) \leq \\ &\leq L_\theta^2 (1 + \varepsilon(13C_1 + 3C_2 + 38)\kappa) L_\theta^{k-2} \|A^{-\frac{1}{2}} r_0\| = \\ &= (1 + \varepsilon(13C_1 + 3C_2 + 38)\kappa) L_\theta^k \|A^{-\frac{1}{2}} r_0\| . \end{aligned} \tag{28}$$

Hence, in both cases, (26) also holds for $i = k$ and (26) follows by induction. □

Remark 6.

If $\kappa \geq 2$ then $L_\theta \geq 1/9$. Hence, if we are willing to make the additional assumption that $\kappa \geq 2$ then (16) and (25) certainly are satisfied if

$$954\varepsilon(C_1 + 2C_2 + 8)\kappa^2 \leq \theta \tag{29}$$ □

We now come to the most important result of this section.

Theorem 7.

Consider iscg with arbitrary initial vectors $r_0, p_0 \neq 0$ and let

$$106\varepsilon(C_1 + 2C_2 + 8)\kappa^2 < 1 \tag{30}$$

then

$$r_i \to 0 \quad \text{and} \quad p_i \to 0 \quad (i \to \infty) . \tag{31}$$

Proof.

In theorem 5 take $\theta \in (0,1)$ such that $106\varepsilon(C_1 + 2C_2 + 8)\kappa^2 \leq \theta L_\theta^2$. This is possible since θL_θ^2 is a continuous function varying from 0 to 1. Then, since $L_\theta < 1$ it follows from (26) that $\| A^{-\frac{1}{2}} r_{i+1} \| \to 0$ $(i \to \infty)$ and consequently $r_i \to 0$ $(i \to \infty)$.

From (33), (36) and (30) it follows that

$$\| A^{\frac{1}{2}} p_{i+1} \|^2 \leq (1 + |\rho_{i+1}|) \| A^{\frac{1}{2}} r_{i+1} \|^2 \leq 2 \| A^{\frac{1}{2}} r_{i+1} \|^2 , \tag{32}$$

and consequently $p_i \to 0$ $(i \to \infty)$. □

5. The convergence of x_i

5.1. Introduction

In the two foregoing chapters we disregarded the computation of x_i. The convergence of the computed recursive residual to zero does not guarantee the convergence of x_i to $\hat{x}$ since in the presence of round-off r_i is different from the true residual $\hat{r}_i := b - Ax_i$, since the computational errors occuring in the implementation of (2.3) and (2.4) are rather independent. Especially, a perturbation on x_{i+1} does not effect r_{i+1}. From this one can see that assuming only that the machine has *strong arithmetic* in the sense of Dekker [2], i.e. multiplication, division, addition and subtraction have a low relative error, bounded by ε times the magnitude of the exact result (see (1.4) and (1.5)), is not sufficient to guarantee the uniform boundedness of $\hat{r}_i - r_i$. For in that case the error in the computation of x_{i+1} can be of order $\varepsilon \| x_i \|$ at each step. Then the difference $\hat{r}_{i+1} - r_{i+1}$ can increase by $\varepsilon \| A \| \| x_i \|$ at each step and this ultimately equals $\varepsilon \| A \| \| \hat{x} \|$.

From his experiments Reid [7] found that $\hat{r}_i$ and r_i depart from each other very slowly. He showed that any errors that occur in the evaluation of p_i and a_i do not make a direct contribution to the difference between the computed recursive residual r_i and the actual value of $b - Ax_{i+1}$.

In the next section we will examine how much the exact true residual of x_{i+1} and the computed recursive residual r_{i+1} can differ in order to obtain an estimate for the asymptotic behaviour of the natural error vector $A^{\frac{1}{2}}(\hat{x} - x_{i+1})$.

5.2. An estimate for the natural relative error

In this section we use the results of chapter 4 with $\theta := \frac{1}{2}$ and $L := L_{\frac{1}{2}}$ as defined by (4.15) Taking another θ would give similar results. The error analysis is carried out under the same conventions as mentioned in section 3.2 and no terms in ε are ignored.

Theorem 1.

Consider iscg with initial vector $x_0 := 0$ and arbitrary initial vector $p_0 \neq 0$.

If

$$(1) \qquad 424\varepsilon(C_1 + 2C_2 + 8)\kappa^2 \le 1$$

then there exists an $i_0 > 0$ such that we have for all $i \ge i_0$:

$$(2) \qquad \frac{\| A^{\frac{1}{2}}(\hat{x} - x_i) \|}{\| A^{\frac{1}{2}}\hat{x} \|} \le 6\varepsilon\{(119 \log 1/\varepsilon + 17\beta)\kappa^{3/2} + 25(C_1 + 3)\kappa^2\} .$$

Proof.

Let $\hat{r}_i$ be the exact true residual of x_i:

$$(3) \qquad \hat{r}_i := b - Ax_i .$$

Note that $x_0 = 0$ implies $r_0 = b = \hat{r}_0$ and $A^{-\frac{1}{2}}r_0 = A^{\frac{1}{2}}\hat{x} = A^{-\frac{1}{2}}\hat{r}_0$.
We have

$$(4) \qquad \| A^{\frac{1}{2}}(\hat{x} - x_{i+1}) \| = \| A^{-\frac{1}{2}}\hat{r}_{i+1} \| \le \| A^{-\frac{1}{2}}(\hat{r}_{i+1} - r_{i+1}) \| + \| A^{-\frac{1}{2}}r_{i+1} \| .$$

Since $\theta L_\theta^2 \ge \frac{1}{4}$ if $\theta = \frac{1}{2}$, inequality (1) implies the validity of (4.25) and consequently we may use the result (4.26). Rather than (4.26) we will use the weaker result:

$$(5) \qquad \| A^{-\frac{1}{2}}r_{i+1} \| \le 2L^i \| A^{\frac{1}{2}}\hat{x} \| , \quad (i \ge 0) .$$

Hence, in (4), $\| A^{-\frac{1}{2}}r_{i+1} \|$ tends to zero and therefore we will concentrate on $y_{i+1} := A^{-\frac{1}{2}}(\hat{r}_{i+1} - r_{i+1})$.

The computed vector x_{i+1} satisfies (see (1.8) and (1.7)):

$$(6) \qquad (I + G_{i+1})x_{i+1} = x_i + (I + F_i)a_i p_i .$$

Hence

$$(7) \qquad x_{i+1} = x_i + a_i p_i + \delta x_{i+1} ,$$

where

$$(8) \qquad \| \delta x_{i+1} \| = \| -G_{i+1}x_{i+1} + F_i a_i p_i \| \le \varepsilon \| x_{i+1} \| + \varepsilon \| a_i p_i \| .$$

Further

$$\hat{r}_{i+1} = b - Ax_i - a_i Ap_i - A\delta x_{i+1} = \hat{r}_i - a_i Ap_i - A\delta x_{i+1} . \tag{9}$$

Together with recursion (3.20) this yields

$$\hat{r}_{i+1} - r_{i+1} = \hat{r}_i - r_i - \delta r_{i+1} - A\delta x_{i+1} \tag{10}$$

or

$$y_{i+1} = y_i - A^{-\frac{1}{2}}\delta r_{i+1} - A^{\frac{1}{2}}\delta x_{i+1} . \tag{11}$$

Hence, since $r_0 = \hat{r}_0$, we obtain the basic formula

$$y_{i+1} = - \sum_{\ell=1}^{i+1} A^{-\frac{1}{2}}\delta r_\ell - \sum_{\ell=1}^{i+1} A^{\frac{1}{2}}\delta x_\ell . \tag{12}$$

The convergence of the first sum follows from the following consideration.

From (3.21), (3.35) and (3.36) we may conclude for $\ell \geq 1$:

$$\begin{aligned} \| \delta r_\ell \| &\leq \varepsilon(\| A^{-\frac{1}{2}} r_\ell \| + (C_1 + 2)\| a_{\ell-1} A^{\frac{1}{2}} p_{\ell-1} \|)\kappa^{\frac{1}{2}} \| A^{\frac{1}{2}} \| \leq \\ &\leq 2\varepsilon(C_1 + 3)\kappa^{\frac{1}{2}} \| A^{\frac{1}{2}} \| \, \| A^{-\frac{1}{2}} r_{\ell-1} \| . \end{aligned} \tag{13}$$

This immediately gives

$$\| \delta r_1 \| \leq 2\varepsilon(C_1 + 3)\kappa^{\frac{1}{2}} \| A^{\frac{1}{2}} \| \, \| A^{\frac{1}{2}} \hat{x} \| \tag{14}$$

and together with (5) it yields for $\ell \geq 2$:

$$\| \delta r_\ell \| \leq 4\varepsilon(C_1 + 3)\kappa^{\frac{1}{2}} \| A^{\frac{1}{2}} \| L^{\ell-2} \| A^{\frac{1}{2}} \hat{x} \| . \tag{15}$$

Hence

$$\begin{aligned} \sum_{\ell=1}^{i+1} \| A^{-\frac{1}{2}}\delta r_\ell \| &\leq 2\varepsilon(C_1 + 3)\kappa \| A^{\frac{1}{2}} \hat{x} \| (1 + 2 \sum_{\ell=2}^{\infty} L^{\ell-2}) \leq \\ &\leq 6\varepsilon(C_1 + 3)\kappa \| A^{\frac{1}{2}} \hat{x} \| (1 - L)^{-1} . \end{aligned} \tag{16}$$

Since generally x_{i+1} will not tend to zero, the convergence of the second sum at the right-hand side of (12) will not follow from (8). However, from lemma 1.2 and formula 1.7 we may conclude that x_{i+1} also satisfies

(17) $$x_{i+1} = x_i + (I + H_i)(I + F_i)a_i p_i$$

and therefore also

(18) $$\| \delta x_{i+1} \| = \| (H_i + F_i + H_i F_i) a_i p_i \| \leq$$

$$\leq (\beta + \varepsilon + (\beta + \varepsilon)\varepsilon) \| a_i p_i \| \leq 3\beta \| a_i p_i \| .$$

Since it follows from (3.35) and (5) that for $i \geq 1$

(19) $$\| a_i p_i \| \leq \| A^{-\frac{1}{2}} \| \, \| a_i A^{\frac{1}{2}} p_i \| \leq 2 \| A^{-\frac{1}{2}} \| \, \| A^{-\frac{1}{2}} r_i \| \leq 4 \| A^{-\frac{1}{2}} \| L^{i-1} \| A^{\frac{1}{2}} \hat{x} \| ,$$

where $L < 1$, the sum $\sum_{\ell=1}^{\infty} \| a_\ell p_\ell \|$ converges. Consequently, from (18) we may conclude the convergence of the second sum in (12).

We now have come to the basic idea for estimating y_{i+1} : in $\sum_{\ell=1}^{i+1} \| A^{\frac{1}{2}} \delta x_\ell \|$ use (8) for small ℓ and use (18) as soon as $\| a_\ell p_\ell \| / \| A^{\frac{1}{2}} \hat{x} \|$ is of order ε.

The last index for which we use (8) will be denoted by N. Let $N \geq 1$ first be arbitrary. Then it follows from (12) that for $i \geq N$:

(20) $$\| y_{i+1} \| \leq \| y_N \| + \sum_{\ell=N+1}^{i+1} \| A^{\frac{1}{2}} \delta x_l \| + \sum_{\ell=N+1}^{i+1} \| A^{-\frac{1}{2}} \delta r_l \| .$$

First we estimate $\| y_N \|$ using (8). To do so we derive a recursion for y_i in terms of $\hat{x}$ and δr_i.
We have

(21) $$\| x_{i+1} \| \leq \| x_{i+1} - \hat{x} \| + \| \hat{x} \| \leq \| A^{-\frac{1}{2}} \| \, \| A^{\frac{1}{2}} (x_{i+1} - \hat{x}) \| + \| \hat{x} \| \leq$$

$$\leq \| A^{-\frac{1}{2}} \| \, \| A^{-\frac{1}{2}} \hat{r}_{i+1} \| + \| \hat{x} \| \leq \| A^{-\frac{1}{2}} \| (\| y_{i+1} \| + \| A^{-\frac{1}{2}} r_{i+1} \|) + \| \hat{x} \| .$$

Since, from (5), $\| A^{-\frac{1}{2}} r_{i+1} \| \leq 2 \| A^{\frac{1}{2}} \hat{x} \|$, we find

(22) $$\| x_{i+1} \| \leq \| A^{-\frac{1}{2}} \| (\| y_{i+1} \| + 3 \| A^{\frac{1}{2}} \hat{x} \|) , \quad (i \geq 0) .$$

From (19)

(23) $$\| a_i p_i \| \leq 4 \| A^{-\frac{1}{2}} \| \, \| A^{\frac{1}{2}} \hat{x} \| , \quad (i \geq 1) .$$

For $i = 0$ one even has $\| a_0 p_0 \| \leq 2 \| A^{-\frac{1}{2}} \| \, \| A^{-\frac{1}{2}} r_0 \| = 2 \| A^{-\frac{1}{2}} \| \, \| A^{\frac{1}{2}} \hat{x} \|$.

Substitution of (22) and (23) in (8) yields

$$\| \delta x_{i+1} \| \leq \varepsilon \| A^{-\frac{1}{2}} \| (\| y_{i+1} \| + 7 \| A^{\frac{1}{2}} \hat{x} \|) \quad . \tag{24}$$

Hence it follows with (11)

$$\| y_{i+1} \| \leq \| y_i \| + \varepsilon \kappa^{\frac{1}{2}} (\| y_{i+1} \| + 7 \| A^{\frac{1}{2}} \hat{x} \|) + \| A^{-\frac{1}{2}} \delta r_{i+1} \| \tag{25}$$

or

$$(1 - \varepsilon \kappa^{\frac{1}{2}}) \| y_{i+1} \| \leq \| y_i \| + \| A^{-\frac{1}{2}} \delta r_{i+1} \| + 7 \varepsilon \kappa^{\frac{1}{2}} \| A^{\frac{1}{2}} \hat{x} \| \quad . \tag{26}$$

Backward repetition from N to 0 of this recursion gives, since $y_0 = 0$,

$$\| y_N \| \leq \sum_{\ell=1}^{N} (1 - \varepsilon \kappa^{\frac{1}{2}})^{\ell - N - 1} (\| A^{-\frac{1}{2}} \delta r_\ell \| + 7 \varepsilon \kappa^{\frac{1}{2}} \| A^{\frac{1}{2}} \hat{x} \|) \quad . \tag{27}$$

Since $\varepsilon \kappa^{\frac{1}{2}} < \frac{1}{2}$ we know that $(1 - \varepsilon \kappa^{\frac{1}{2}})^{\ell - N - 1} \leq (1 - \varepsilon \kappa^{\frac{1}{2}})^{-N} < e^{2N\varepsilon\kappa^{\frac{1}{2}}}$ and hence we have

$$\| y_N \| \leq e^{2N\varepsilon\kappa^{\frac{1}{2}}} (7N\varepsilon\kappa^{\frac{1}{2}} \| A^{\frac{1}{2}} \hat{x} \| + \sum_{\ell=1}^{N} \| A^{-\frac{1}{2}} \delta r_\ell \|) \quad . \tag{28}$$

We now return to (20).

Since $e^{2N\varepsilon\kappa^{\frac{1}{2}}} > 1$ we may conclude from (20) and (28) that for $i \geq N \geq 1$:

$$\| y_{i+1} \| \leq e^{2N\varepsilon\kappa^{\frac{1}{2}}} (7N\varepsilon\kappa^{\frac{1}{2}} \| A^{\frac{1}{2}} \hat{x} \| + \sum_{\ell=1}^{i+1} \| A^{-\frac{1}{2}} \delta r_\ell \|) + \sum_{\ell=N+1}^{i+1} \| A^{\frac{1}{2}} \delta x_\ell \| \quad . \tag{29}$$

We now use (18) to estimate the last sum in (29).

From (18) and (19) we find

$$\sum_{\ell=N+1}^{i+1} \| A^{\frac{1}{2}} \delta x_\ell \| \leq 12 \kappa^{\frac{1}{2}} \beta \| A^{\frac{1}{2}} \hat{x} \| \sum_{\ell=N+1}^{\infty} L^{\ell - 2} \leq 12 \kappa^{\frac{1}{2}} \beta \| A^{\frac{1}{2}} \hat{x} \| L^{N-1} (1 - L)^{-1} \quad . \tag{30}$$

Substitution of (16) and (30) in (29) yields

$$\| y_{i+1} \| \leq e^{2N\varepsilon\kappa^{\frac{1}{2}}} (7N\varepsilon\kappa^{\frac{1}{2}} + 6\varepsilon(C_1 + 3)\kappa(1 - L)^{-1}) \| A^{\frac{1}{2}} \hat{x} \| + 12 \kappa^{\frac{1}{2}} \beta L^{N-1} (1 - L)^{-1} \| A^{\frac{1}{2}} \hat{x} \| \quad . \tag{31}$$

Now let N be the smallest integer such that $L^{N-1} \leq \varepsilon$.
Then

$$N \leq (\log 1/\varepsilon)/(\log 1/L) + 2 \ . \tag{32}$$

Since $L = (1 - 2\kappa/(\kappa + 1)^2)^{\frac{1}{2}}$ and $\kappa \geq 1$ it follows

$$L < 1 - \kappa/(\kappa + 1)^2 \ , \quad (\log 1/L) > (16\kappa)^{-1} \tag{33}$$

and consequently

$$N \leq 16\kappa \log 1/\varepsilon + 2 \leq 17\kappa \log 1/\varepsilon \ , \tag{34}$$

$$N\varepsilon\kappa^{1/2} \leq 17\kappa^{3/2}\varepsilon \log 1/\varepsilon \ . \tag{35}$$

Note that $\varepsilon \log 1/\varepsilon \leq 4/e$ so that, with (1), $2N\varepsilon\kappa^{\frac{1}{2}} < 1$ and $e^{2N\varepsilon\kappa^{\frac{1}{2}}} < 3$. Also $(1 - L)^{-1} \leq 4\kappa$.
Substituting the various inequalities in (31) yields for $i \geq N$:

$$\|y_{i+1}\| \leq 3\varepsilon\{(119 \log 1/\varepsilon + 16\beta)\kappa^{3/2} + 24(C_1 + 3)\kappa^2\}\|A^{\frac{1}{2}}\hat{x}\| \ . \tag{36}$$

From (5) and our choice of N we find for $i \geq N$:

$$\|A^{-\frac{1}{2}}r_{i+1}\| \leq 2\varepsilon\|A^{\frac{1}{2}}\hat{x}\| \ . \tag{37}$$

Substitution of (36) and (37) in (4) proves (2). □

Remark 2.

Wozniakowski [8] proves, neglecting terms of order ε^2, that his version of the conjugate gradient algorithm (wcg) produces vectors x_i such that ultimately

$$\|A^{\frac{1}{2}}(\hat{x} - x_i)\| \leq C\varepsilon\kappa\|A^{\frac{1}{2}}\|\,\|x_i\| \ , \tag{38}$$

where C is a constant depending on C_1 and C_2.
From (38) it follows that

$$\frac{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|}{\|A^{\frac{1}{2}}\hat{x}\|} \leq C'\varepsilon\kappa^{3/2} \tag{39}$$

This result essentially differs from our result (2), (50) by a factor $\max(\kappa^{\frac{1}{2}}, \log 1/\varepsilon)$. From our assumption (1) it does not follow which of the two constants $\kappa^{\frac{1}{2}}$ and $\log 1/\varepsilon$ is the largest.

Analytically the factor $\kappa^{\frac{1}{2}}$ is caused by the fact that our estimate for $\sum_{\ell=1}^{\infty} A^{\frac{1}{2}}\delta r_\ell$ contains a factor $(1 - L)^{-1}$, which is of order κ^2 (see (16)).
This factor is not a consequence of the rather complicated way we bounded $\sum_{\ell=1}^{\infty} A^{-\frac{1}{2}}\delta x_\ell$.
The factor $\log 1/\varepsilon$ comes from the first N terms of the sum $\sum_{\ell=1}^{\infty} A^{-\frac{1}{2}}\delta x_\ell$.
We think that it will be difficult to find a set of data that confirm the difference between the estimates (2) and (39) for respectively iscg and wcg. □

Remark 3.

We may expect that ultimately the computed true residual $\tilde{r}_i := fl(\hat{r}_i)$ is at least of order $\varepsilon\|A\|\,\|\hat{x}\|$. Since r_i tends to zero as i tends to infinity, the difference between the computed true residual and the computed recursive residual ultimately will be at least of order $\varepsilon\|A\|\,\|\hat{x}\|$. □

Remark 4.

Let $1 \le j \le n$ and let x^j denote the j^{th} component of the vector x. Suppose there exists an i_0 and a positive real number α such that $|x_i^j| \ge \alpha$ for all $i \ge i_0$. Since $fl(a_i p_i) \to 0$ $(i \to \infty)$ then certainly there exists an i_1 such that $|(fl(a_i p_i))^j| < (\varepsilon/\beta)|x_i^j|$ for all $i \ge i_1$. From (1.16) we then may conclude that $x_{i+1}^j = x_i^j$ for all $i \ge i_1$. Consequently, if all components of $\hat{x}$ are nonzero and if ε is small enough, then after a certain number of iterations the vectors x_i do not change anymore. □

Remark 5.

If we take an arbitrary $x_0 \neq 0$ in theorem 1 then $r_0 \neq \hat{r}_0$ and consequently $A^{-\frac{1}{2}}r_0 \neq A^{-\frac{1}{2}}\hat{r}_0$, $y_0 \neq 0$. We will study what difference it makes for the proof of theorem 1 given earlier if $x_0 \neq 0$.
Instead of (5) we have in that case

$$\|A^{-\frac{1}{2}}r_{i+1}\| \le 2L^i\|A^{-\frac{1}{2}}r_0\|\,, \quad (i \ge 0) \tag{40}$$

and (12) becomes

$$y_{i+1} = y_0 - \sum_{\ell=1}^{i+1} A^{-\frac{1}{2}}\delta r_\ell - \sum_{\ell=1}^{i+1} A^{\frac{1}{2}}\delta x_\ell\,. \tag{41}$$

Following the lines of the proof of theorem 1 we obtain:
instead of (24) :

$$\|\delta x_{i+1}\| \le \varepsilon\|A^{-\frac{1}{2}}\|\,(\|y_{i+1}\| + 6\|A^{-\frac{1}{2}}r_0\| + \|A^{\frac{1}{2}}\hat{x}\|)\,, \tag{42}$$

instead of (27) :

$$\|y_N\| \le (1-\varepsilon\kappa^{\frac12})^{-N}\|y_0\| + \tag{43}$$

$$+ \sum_{\ell=1}^{N} (1-\varepsilon\kappa^{\frac12})^{\ell-N-1}\{\|A^{-\frac12}\delta r_\ell\| + \varepsilon\kappa^{\frac12}(6\|A^{-\frac12}r_0\| + \|A^{\frac12}\hat{x}\|)\} .$$

instead of (36) :

$$\|y_{i+1}\| \le 3\varepsilon\{\|y_0\|/\varepsilon + 17\log 1/\varepsilon\kappa^{3/2}(6\|A^{-\frac12}r_0\| + \|A^{\frac12}\hat{x}\|) + \tag{44}$$

$$+ 24(C_1+3)\kappa^2\|A^{-\frac12}r_0\| + 16\beta\kappa^{3/2}\|A^{-\frac12}r_0\|\} .$$

For the computation of r_0 we have

$$r_0 = (I+F)(b-(A+E)x_0) = \hat{r}_0 + F(b-Ax_0) - (I+F)Ex_0 . \tag{45}$$

Hence, since $\|A^{\frac12}x_0\| \le \|A^{-\frac12}\hat{r}_0\| + \|A^{\frac12}\hat{x}\|$ we obtain using (1)

$$\|y_0\| = \|A^{-\frac12}(r_0-\hat{r}_0)\| \le \varepsilon(\|A^{-\frac12}\|\,\|\hat{r}_0\| + 2C_1\kappa^{\frac12}\|A^{\frac12}\|\,\|x_0\|) \le \tag{46}$$

$$\le 2\varepsilon(C_1+1)\kappa(\|A^{-\frac12}\hat{r}_0\| + \|A^{\frac12}\hat{x}\|) .$$

Consequently, again using (1),

$$\|A^{-\frac12}r_0\| \le \|y_0\| + \|A^{-\frac12}\hat{r}_0\| \le \kappa^{-1}\|A^{\frac12}\hat{x}\| + 2\|A^{-\frac12}\hat{r}_0\| . \tag{47}$$

Hence,certainly

$$\|A^{-\frac12}r_0\| \le \|A^{\frac12}\hat{x}\| + 2\|A^{-\frac12}\hat{r}_0\| \tag{48}$$

and from (40) for $i \ge N$

$$\|A^{-\frac12}r_{i+1}\| \le 2\varepsilon(\|A^{\frac12}\hat{x}\| + 2\|A^{-\frac12}\hat{r}_0\|) . \tag{49}$$

So finally from (4), (44), (46), (47), (48) and (49) we get

$$\|A^{\frac12}(\hat{x}-x_i)\| \le 48\varepsilon\{[\beta\kappa^{\frac12} + 2(C_1+3)\kappa + 8\log 1/\varepsilon\kappa^{3/2}]\|A^{\frac12}\hat{x}\| + \tag{50}$$

$$+ [(13\log 1/\varepsilon + 2\beta)\kappa^{3/2} + 4(C_1+3)\kappa^2]\,\|A^{\frac12}(\hat{x}-x_0)\|\} .$$

Hence, if $\|A^{\frac12}(\hat{x}-x_0)\| \le C\|A^{\frac12}\hat{x}\|$ for some reasonably small $C > 0$ then (50) essentially is the same as (2). This certainly is the case if $\|A^{\frac12}x_0\| \le C\|A^{\frac12}\hat{x}\|$ (and especially if $x_0 = 0$).

If $\|A^{\frac12}(\hat{x}-x_0)\| \le C\kappa^{-\frac12}\|A^{\frac12}\hat{x}\|$ we even have

$$\frac{\|A^{\frac12}(\hat{x}-x_0)\|}{\|A^{\frac12}\hat{x}\|} \le 48\varepsilon\{2[C_1+3+(C+1)\beta]\kappa + [(8+13C)]\log 1/\varepsilon + \tag{51}$$

$$+ 4C(C_1+3)]\,\kappa^{3/2}\}$$

6. Final comments

6.1. Comparison with Wozniakowski's results

In this section we compare Wozniakowski's [8] results and our results. In order to be in a position to ignore factors of the type $(1 + O(\varepsilon))$ Wozniakowski uses inequalities of the type $f(\varepsilon) \dot{\leq} g(\varepsilon)$, which means that $f(\varepsilon) \leq g(\varepsilon)(1 + O(\varepsilon))$. Most of his results are expressed in terms of this sort of inequalities.
We will use the same notation in this section. In order to be able to report on Wozniakowski's results and to discuss the relation to our results, we define the following two *gradient algorithms*.

Algorithm 1.

take x_0 ;
$r_0 := b - Ax_0$;
$i := 0$;
while $r_i \neq 0$ **do**
begin

(1) $a_i := (r_i, r_i)/(r_i, Ar_i)$;

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

(3) $r_{i+1} := \begin{cases} \text{either } b - Ax_{i+1} \text{ ;} \\ \text{or} \quad r_i - a_i Ar_i \text{ ;} \end{cases}$
(4)

(5) $i := i + 1$

end .

If the true residual formula (3) is in use then this algorithm will be referred to as *true residual gradient algorithm* (trg) and if the recursive formula (4) is in use, then this algorithm will be referred to as *recursive residual gradient algorithm* (rrg). Ofcourse, algebraically there is no difference between these two versions.

Wozniakowski considers first trg with round-off and then uses the results for the analysis of his version of the conjugate gradient algorithm (wcg).
For trg he proves the following basic result

Theorem 2.

If

$$2\varepsilon(C_1 + 2C_2 + 8)\kappa < 1 \tag{6}$$

then the sequence $\{x_i\}$ computed by trg satisfies

$$\overline{\lim_{i\to\infty}} \|A^{\frac{1}{2}}(\hat{x} - x_i)\| \dot{\leq} 3\varepsilon(5C_1 + 1)\kappa\|A^{\frac{1}{2}}\| \overline{\lim_{i\to\infty}} \|x_i\| . \tag{7}$$

However, from this theorem it does not even follow that $\overline{\lim}\, \|x_i\|$ is bounded.

Since

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

we may conclude from (7) that

$$\overline{\lim}\, \|A^{\frac{1}{2}}(\hat{x} - x_i)\| \dot{\leq} 3\varepsilon(5C_1 + 1)\{\kappa\|A^{\frac{1}{2}}\| \|\hat{x}\| + \kappa^{3/2} \overline{\lim}\, \|A^{\frac{1}{2}}(\hat{x} - x_i)\|\} . \tag{9}$$

Consequently, if additionally to (6) also

$$3\varepsilon(5C_1 + 1)\kappa^{3/2} < 1 \tag{10}$$

is satisfied, then it follows that

$$\overline{\lim}\, \|A^{\frac{1}{2}}(\hat{x} - x_i)\| \dot{\leq} \frac{3\varepsilon(5C_1 + 1)\kappa}{1 - 3\varepsilon(5C_1 + 1)\kappa^{3/2}} \|A^{\frac{1}{2}}\| \|\hat{x}\| \dot{=} \tag{11}$$

$$\dot{=} 3\varepsilon(5C_1 + 1)\kappa\|A^{\frac{1}{2}}\| \|\hat{x}\| .$$

Hence in that case $\overline{\lim}\ x_i$ is bounded.

From (11) it follows that ultimately

$$\frac{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|}{\|A^{\frac{1}{2}}\hat{x}\|} \dot{\leq} 3\varepsilon(5C_1 + 1)\kappa^{3/2} , \tag{12}$$

if (6) and (10) are satisfied.

Then under the same conditions one can prove that ultimately for the computed true residuals

$$\|A^{-\frac{1}{2}}r_i\| \dot{\leq} \varepsilon(15C_1 + 4)\kappa^{3/2} \|A^{\frac{1}{2}}\hat{x}\| . \tag{13}$$

Wozniakowski [8] does not contain results on rrg.
We made an error analysis of rrg, carried out under the same conventions as used in chapters 3 and 4. Note that part I of the proof of theorem 3.1 also holds for rrg replacing all p_i by r_i.
We obtained the following result.

Theorem 3.
Let $0 < \theta < 1$ and let

$$16\varepsilon(C_1 + 2C_2 + 2)\kappa^{3/2} \leq \theta \ , \tag{14}$$

then the sequence $\{r_i\}$ computed by rrg satisfies

$$\|A^{-\frac{1}{2}} r_{i+1}\| \leq L_\theta \|A^{-\frac{1}{2}} r_i\| \ , \tag{15}$$

where L_θ is defined by (4.15) .
Consequently $r_i \to 0$ $(i \to \infty)$.

From this it follows, using similar arguments as in chapter 5, that ultimately for rrg , with $x_0 = 0$,

$$\frac{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|}{\|A^{\frac{1}{2}}\hat{x}\|} \leq 3\varepsilon\{(119 \log 1/\varepsilon + 17\beta)\kappa^{3/2} + 25(C_1 + 3)\kappa^2\} \ , \tag{16}$$

if

$$16\varepsilon(C_1 + 2C_2 + 2)\kappa^{3/2} < 1 \ . \tag{17}$$

Remark 4.
Note that restriction (10) and (17) are of the same order in ε and κ but the estimates (12) and (16) differ by a factor $\kappa^{\frac{1}{2}}$. This is due to the fact that, just like in (3.21) the estimate for δr_i of the recursive residual r_i contains a factor $\kappa^{\frac{1}{2}}$ for rrg (see also remark 5.2) . □

Remark 5.
If we compare the restriction (5.1) for iscg and the restriction (17) for rrg, then we see that (17) is weaker in the sense that it contains a factor $\kappa^{3/2}$ instead of κ^2. Restriction (5.1) for iscg is a direct consequence of restriction (4.16) in theorem 4.2 which contains a factor κ^2. Tracing the proof of theorem 4.2 we observe that this factor κ^2 is caused by the fact that the estimate for γ_1 contains a factor κ.

Considering the proof of lemma 4.1 we notice that the factor κ appearing in the estimate for γ_1 is a consequence of the factor κ in the estimates for μ_1 and ρ_1.
However, for the proof of theorem 3 we did not use the estimate (3.5) for μ_i, but from (3.26) and (3.27) in part I of the proof of theorem 3.1 we proved for the rrg-case the validity of the inequality

$$(18) \qquad \frac{\|A^{-\frac{1}{2}}r_{i+1}\|^2}{\|A^{-\frac{1}{2}}r_i\|^2} \leq 1 - \frac{\|r_i\|^4}{\|A^{-\frac{1}{2}}r_i\|^2\|A^{\frac{1}{2}}r_i\|^2}\left(1 - \eta_i\frac{\|A^{-\frac{1}{2}}r_i\|^2}{\|r_i\|^2}\right),$$

under the restriction $16\varepsilon(C_1 + 2C_2 + 2)\kappa < 1$,
where

$$(19) \qquad \eta_i = \varepsilon\|A\|(8\kappa^{\frac{1}{2}}(C_1 + 1) + C_1 + 3C_2 + 18) .$$

From this theorem 3 easily follows.
Since in rrg the vectors p_i do not occur, the constant ρ_i is irrelevant in this case. □

Remark 6.
From (15) it follows that $A^{-\frac{1}{2}}r_i$ decreases at least by a factor L at each step for rrg, whereas for iscg we were only able to prove the weaker result expressed in remark 4.3 . □

Having indicated the difference between the use of true and recursive residuals for the gradient algorithm, we now come to the difference between Wozniakowski's version of the conjugate gradient algorithm (wcg) and our version.

The algorithm wcg is closely related to trg.
In wcg each step consists of two parts:
First the algorithm computes z_{i+1} from x_i by one step trg, i.e.

$$(20) \qquad r_i := b - Ax_i ;$$

$$(21) \qquad a_i := (r_i, r_i)/(r_i, Ar_i) ;$$

$$(22) \qquad z_{i+1} := x_i + a_i r_i .$$

Hence z_{i+1} minimizes $\|A^{\frac{1}{2}}(\hat{x} - z)\|$ along the line $z = x_i + ar_i$.
Secondly, the next approximant x_{i+1} is computed from

(23) $\quad y_{i+1} := z_{i+1} - x_{i-1}$;

(24) $\quad b_{i+1} := (y_{i+1}, b - Az_{i+1})/(y_{i+1}, Ay_{i+1})$;

(25) $\quad x_{i+1} := z_{i+1} - b_{i+1}y_{i+1}$.

Hence x_{i+1} minimizes $\|A^{\frac{1}{2}}(\hat{x} - x)\|$ along the line $x = z_{i+1} - by_{i+1}$. Note that in (18) as well as in (22) true residuals are taken. If no round-off occurs then wcg and cg give the same sequence $\{x_i\}$. Algebraically it is trivial that x_{i+1} is a better approximant to $\hat{x}$ than z_{i+1}, whatever y_{i+1} might be. Wozniakowski shows that, numerically, x_{i+1} is nearly (apart from terms of order ε) as good as z_{i+1}.
For z_{i+1} he uses the results for trg. Then he obtains

Theorem 7.

If

(26) $\quad 2\varepsilon(C_1 + 2C_2 + 8)\kappa < 1$

then the sequence $\{x_i\}$ computed by wcg satisfies

(27) $\quad \overline{\lim}\, \|A^{\frac{1}{2}}(\hat{x} - x_i)\| \leq 3\varepsilon(5C_1 + 2)\kappa \|A^{\frac{1}{2}}\| \, \overline{\lim}\, \|x_i\|$.

In a similar way as we concluded from theorem 2 the validity of (12), we conclude from theorem 6 that wcg produces x_i that ultimately satisfy

(28) $\quad \dfrac{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|}{\|A^{\frac{1}{2}}\hat{x}\|} \leq 3\varepsilon(5C_1 + 2)\kappa^{3/2}$

if (26) holds and if, moreover,

(29) $\quad 3\varepsilon(5C_1 + 2)\kappa^{3/2} < 1$.

Looking only at factors $\kappa^{\frac{1}{2}}$ appearing in the convergence results for the estimates for the natural relative error and in the restrictions, we come to the following brief comparison for the various algorithms

trg and rrg : the same restriction; the natural relative error is a factor $\kappa^{\frac{1}{2}}$ larger for rrg,

rrg and iscg: the restriction is a factor $\kappa^{\frac{1}{2}}$ stronger for iscg; the same natural relative error,

wcg and trg : the same restriction; the same natural relative error,

wcg and iscg: the restriction is a factor $\kappa^{\frac{1}{2}}$ stronger for iscg; the natural relative error is a factor $\kappa^{\frac{1}{2}}$ larger for iscg.

6.2. A class of iscg algorithms

Wozniakowski [8] considers a class of conjugate gradient algorithms. This class consists of algorithms that are the same as his original version of the conjugate gradient algorithm except from the computation of the constant b_i appearing in the original algorithm (see (2.4)). Instead of taking the value of the expression (24) the algorithms compute and use constants $\tilde{b}_i$ satisfying

$$\tilde{b}_i = b_i(1 + \delta b_i) ,$$

where $|\delta b_i| \leq 1$ and b_i is given by (24). He shows that his convergence results for wcg are valid for this whole class.
In imitation of Wozniakowski we define for every $M \geq 0$ a class Φ_M of iscg-algorithms. An algorithm φ belongs to Φ_M if φ is defined by the statements of iscg (see section 2.2) except from the computation of b_i. The actual values of b_i now only must satisfy

$$b_i = - \frac{(r_{i+1},Ap_i)}{(p_i,Ap_i)} (1 + \delta b_i) ,$$

with $|\delta b_i| \leq M$. It is quite obvious that (3.2) and (3.5) also hold for every $\varphi \in \Phi_M$ since part I of the proof of theorem is only based on the formula (2.2) and (2.4) It is easily seen that the conjugacy relation $(p_{i+1},Ap_i) = 0$ does not hold if $\delta b_i \neq 0$ and therefore the name conjugate gradient algorithm in fact is incorrect. All the results for iscg, given in the chapters 3, 4 and 5, can be generalized for Φ_M. We only formulate the following generalization of theorem 4.5 .

Theorem 4.

Let $M \geq 0$, $\varphi \in \Phi_M$ and let $r_0,p_0 \neq 0$ be arbitrary initial vectors.
Let $0 < \theta < 1$ and

$$L_{M,\theta} := \left\{1 - \frac{4(1-\theta)\kappa}{(\kappa+1)^2(M^2 + (1+M)^2)}\right\}^{\frac{1}{2}} .$$

If

$$208\varepsilon(C_1 + 2C_2 + 8)\kappa^2(M+1)^2 \leq \theta L_{M,\theta}^2 ,$$

then the sequence $\{r_i\}$ computed by φ satisfies

$$\|A^{-\frac{1}{2}} r_{i+1}\| \leq (1 + \varepsilon(13C_1 + 3C_2 + 38)\kappa) L_{M,\theta}^i \|A^{-\frac{1}{2}} r_0\| . \qquad \square$$

References

[1] Crowder, H. and P. Wolfe: Linear convergence of the conjugate gradient method.
IBM J. Res. Develop. 16 (1972), 431 - 433.

[2] Dekker, T.J.: Correctness proofs and machine arithmetic.
Proc. of the IFIP TC2 Working Conference on Performance Evaluation of Numerical Software; ed. by L.D. Fosdick, Amsterdam, North Holland, 1979 (to appear).

[3] Hestenes, M.R. and E. Stiefel: Methods of conjugate gradients for solving linear systems.
NBS J. Res. 49 (1952), 409 - 436.

[4] Householder, A.S.: The theory of matrices in numerical analysis.
New York etc., Blaisdell, 1964.

[5] Kammerer, W.J. and M.Z. Nashed: On the convergence of the conjugate gradient method for singular linear operator equations.
SIAM J. Mumer. Anal. 9 (1972), 165 - 181.

[6] Powell, M.J.D.: Some convergence properties of the conjugate gradient method.
Mathematical programming 11 (1976), 42 - 49.

[7] Reid, J.K.: On the method of conjugate gradients for the solution of large sparse systems of linear equations.
Proc. of a Conference on Large Sparse Sets of Linear Equations; ed. by J.K. Reid, New York, Academic Press, 1971, 231 - 254.

[8] Wozniakowski, H.: Round-off error analysis of a new class of conjugate gradient algorithms.
Pittsburgh, Carnegie-Mellon Univ., Dept. Comp. Sc., 1978 (Report CMU-CS-78-153).