COMPUTATIONAL MATHEMATICS
BANACH CENTER PUBLICATIONS, VOLUME 13
PWN–POLISH SCIENTIFIC PUBLISHERS
WARSAW 1984

ROUND-OFF ERROR ANALYSIS OF THE GRADIENT METHOD

J. A. M. BOLLEN

Department of Mathematics, Technological University Twente, The Netherlands

1. Introduction

We study the gradient method (GM) which was first described by Cauchy in 1847. This method belongs to the class of descent methods (DM's) for the determination of the solution $\hat{x} := A^{-1}b$ of a linear system $Ax = b$, where A is symmetric and positive definite. 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 of movement is determined and the process is repeated. The objective function F must have the following three important properties: $F(\hat{x}) = 0$, $F(x) > 0$ if $x \neq \hat{x}$ and F is convex. In the case of the GM the objective function is $F(x) := \big(\hat{x}-x, A(\hat{x}-x)\big)$, expressed in terms of the Euclidean inner product. Obviously, a good choice for moving towards $\hat{x}$ is to move in the (opposite) direction of the gradient vector of the objective function since this is the direction of steepest descent of the objective function. The gradient vector $\nabla F(x)$ of the objective function $F(x) = \big(\hat{x}-x, A(\hat{x}-x)\big)$ at point x_i satisfies

$$(1) \qquad \nabla F(x_i) = -2A(\hat{x}-x_i) = -2(b-Ax_i).$$

The GM is based on this idea; as search direction in the step from i to $i+1$ one chooses the direction of the *residual vector* $r_i := b-Ax_i$.

The GM is of special importance from a theoretical point of view, since it is one of the simplest nonlinear methods for which a satisfactory analysis of the convergence behavior exists in the case of exact computations. Many more advanced methods, like, e.g., 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 [5] is until now the only author who gave a complete round-off error analysis for the GM in order to obtain assertions on the numerical behavior and the attainable accuracy of the GM. Our results derived in this paper are superior to those of Woźniakowski in two aspects. Firstly, we prove step-wise linear convergence of the objective function whereas Woźniakowski gives a result in terms of the limes superior. Secondly, we prove good-behavior, whereas Woźniakowski's result does not even imply numerical stability (for definitions see Subsection 3.1).

Given a definite system $Ax = b$, then 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
(2)        $a_i := (r_i, r_i)/(r_i, Ar_i)$;
(3)        $x_{i+1} := x_i + a_i r_i$;
(4)        $r_{i+1} := b - Ax_{i+1}$;
           $i := i+1$
end.
```

The inner products in the statement for a_i are Euclidean inner products.

We summarize the contents of the paper. In Section 2 we deduce some well-known elementary algebraic properties of the GM concerning speed of convergence of the *natural error* $\|A^{1/2}(\hat{x} - x_i)\|$ and of the *error* $\|\hat{x} - x_i\|$ if no round-off occurs. The main reason for deducing these properties here is that they are basic for studying the method in the presence of round-off. In Subsection 3.1 we present some preliminaries, basic tools and notions needed for our round-off error analysis in the next two subsections. In Subsection 3.2 we derive our main theorems. They contain the numerical analogues of the algebraic properties in Section 2 for the perturbed GM. In Subsection 3.3 we show how this leads to assertions on speed of convergence and attainable accuracy. Finally, in Section 4 we make some final remarks.

In this paper $(\cdot, \cdot)$ stands for the Euclidean inner product, $\|\cdot\|$ in connection with a vector stands for the Euclidean norm and $\|\cdot\|$ in connection with a matrix stands for the spectral norm. $\varkappa$ denotes the condition number $\|A\|\|A^{-1}\|$ of the matrix A. The matrices $A^{1/2}$, $A^{-1/2}$ are the uniquely determined symmetric positive definite matrices satisfying $A^{1/2}A^{1/2} = A$ and $A^{-1/2} = (A^{1/2})^{-1}$. Note that the following relations hold: $(x, Ax) = (A^{1/2}x, A^{1/2}x) = \|A^{1/2}x\|^2$ and $(x, A^{-1}x) = (A^{-1/2}x, A^{-1/2}x) = \|A^{-1/2}x\|^2$.

2. The exact gradient method

In this section we consider the GM iterations if no round-off occurs. A well-known elementary result is given in the following theorem.

THEOREM 1. *At each step the exact GM minimizes the objective function*

$$F(x) = \big((\hat{x}-x), A(\hat{x}-x)\big) = \|A^{1/2}(\hat{x}-x)\|^2 \tag{5}$$

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

$$\frac{\|A^{1/2}(\hat{x}-x_{i+1})\|^2}{\|A^{1/2}(\hat{x}-x_i)\|^2} = 1-\gamma_i^2, \tag{6}$$

where

$$\gamma_i := \frac{\|r_i\|^2}{\|A^{-1/2}r_i\|\|A^{1/2}r_i\|}. \tag{7}$$

Proof. Along the line $x = x_i + ar_i$ we have

$$\begin{aligned} F(x) &= \|A^{1/2}(\hat{x}-x_i-ar_i)\|^2 \\ &= F(x_i) - 2a\big(A(\hat{x}-x_i), r_i\big) + a^2\|A^{1/2}r_i\|^2 \\ &= F(x_i) + \|A^{1/2}r_i\|^2\left(a - \frac{\big(A(\hat{x}-x_i), r_i\big)}{\|A^{1/2}r_i\|^2}\right)^2 - \frac{\big(A(\hat{x}-x_i), r_i\big)^2}{\|A^{1/2}r_i\|^2} \end{aligned} \tag{8}$$

which is minimal for

$$a = \frac{\big(A(\hat{x}-x_i), r_i\big)}{\|A^{1/2}r_i\|^2} = \frac{(r_i, r_i)}{(r_i, Ar_i)} = a_i \tag{9}$$

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

$$\|A^{1/2}(\hat{x}-x_{i+1})\|^2 = \|A^{1/2}(\hat{x}-x_i)\|^2 - (r_i, r_i)^2/\|A^{1/2}r_i\|^2, \tag{10}$$

which proves (6) and (7). ■

Since the GM minimizes $\|A^{1/2}(\hat{x}-x_i)\|$ at each step it seems natural to measure the error this way instead of measuring it by $\|\hat{x}-x_i\|$ or $\|A(\hat{x}-x_i)\|$. Therefore, $\|\hat{x}-x_i\|$ is called the *error* and $\|A^{1/2}(\hat{x}-x_i)\|$ is called the *natural error*. In order to obtain, from (6), an upper bound for the decrement of the natural error at each step we use the *Kantorovich inequality* which states that if A is symmetric and positive definite, then for any vector x one has

$$\frac{(x,x)^2}{(x,Ax)(x,A^{-1}x)} \geqslant \frac{4\varkappa}{(\varkappa+1)^2}. \tag{11}$$

The left-hand side of (11) can be written in terms of norms as $\|x\|^4/(\|A^{1/2}x\| \times \times \|A^{-1/2}x\|)^2$ and consequently it follows from (6) and (7) that

$$\frac{\|A^{1/2}(\hat{x}-x_{i+1})\|^2}{\|A^{1/2}(\hat{x}-x_i)\|^2} \leqslant 1-\frac{4\varkappa}{(\varkappa+1)^2} = \left(\frac{\varkappa-1}{\varkappa+1}\right)^2. \tag{12}$$

Hence the natural error converges step-wise linearly to zero with a ratio no greater than $(\varkappa-1)/(\varkappa+1)$.

Another well-known convergence property of the exact GM reads as follows.

THEOREM 2. *At each step of the exact GM the error* $\|\hat{x}-x_i\|$ *decreases and*

$$\frac{\|\hat{x}-x_{i+1}\|^2}{\|\hat{x}-x_i\|^2} = 1-\sigma_i(2-\varrho_i), \tag{13}$$

where

$$\sigma_i := \frac{\|r_i\|^2\,\|A^{-1/2}r_i\|^2}{\|A^{1/2}r_i\|^2\,\|A^{-1}r_i\|^2} \tag{14}$$

and $\varrho_i := \gamma_i^2$, *with* γ_i *defined by* (7).

Proof. If in the equality $\hat{x}-x_{i+1} = \hat{x}-x_i-a_ir_i$ we take squared norms at both sides we obtain

$$\|\hat{x}-x_{i+1}\|^2 = \|\hat{x}-x_i\|^2-2a_i(\hat{x}-x_i, r_i)+a_i^2(r_i, r_i). \tag{15}$$

Since $a_i = \|r_i\|^2/\|A^{1/2}r_i\|^2$ and $\hat{x}-x_i = A^{-1}r_i$ we obtain (13) after some rearrangements. ■

Since the Euclidean norm and the spectral norm are compatible we have $\sigma_i \geqslant \varkappa^{-1}$ and, using the Schwarz inequality, $\varrho_i \leqslant 1$. Hence from (13) and (14) it follows that

$$\frac{\|\hat{x}-x_{i+1}\|^2}{\|\hat{x}-x_i\|^2} \leqslant 1-\frac{1}{\varkappa}, \tag{16}$$

which implies the step-wise linear convergence to zero of the error with a ratio no greater than $(1-1/\varkappa)^{1/2}$.

3. The perturbed gradient method

In this section we consider the GM if round-off occurs, due to the use of floating point arithmetic.

3.1. Notations, definitions and conventions

Rounding errors. We assume that the GM-algorithm given in Section 1 is performed on a floating point machine with relative precision ε and that adding or subtracting two machine vectors x and y and multiplying a machine vector x and a machine number a yield computed vectors $\mathrm{fl}(x \pm y)$ and $\mathrm{fl}(ax)$ satisfying

$$(17)\qquad \begin{cases} \mathrm{fl}(x \pm y) = (I+F_1)(x \pm y), & \|F_1\| \leqslant \varepsilon, \\ \mathrm{fl}(ax) = (I+F_2)(ax), & \|F_2\| \leqslant \varepsilon. \end{cases}$$

Note that (17) is fulfilled in all practical implementations where F_1 and F_2 are diagonal matrices.

Furthermore, we assume that the matrix by vector product calculation of a machine matrix A and a machine vector x and the inner product calculation of two machine vectors x and y satisfy the relations

$$(18)\qquad \begin{cases} \mathrm{fl}(Ax) = (A+E)x, & \|E\| \leqslant \varepsilon C_1 \|A\|, \\ \mathrm{fl}\big((x, y)\big) = \big((I+D)x, y\big), & \|D\| \leqslant \varepsilon C_2, \end{cases}$$

where C_1 and C_2 are constants depending only on n and ε. Throughout this paper C_1 and C_2 stand for the upper bounds of the round-off matrices E and D according to (18). For the standard algorithms E is a full matrix and D is a diagonal matrix whereas C_1 is of order $n^{3/2}$ and C_2 is of order n. In our round-off error analysis we neglect the possibility of underflow and overflow.

o-notation. In order to simplify the expressions arising from the application of the basic relations (17) and (18) we use a kind of Bachmann–Landau o-notation 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 instance, we write

$$(19)\qquad \varepsilon(1+C_1\varkappa+2\varepsilon C_2\varkappa^{1/2}) = \varepsilon(1+C_1\varkappa+o), \quad [\varepsilon C_2\varkappa^{1/2} \to 0],$$

where the expression between square brackets indicates that o stands for a quantity which is smaller than a constant times $\varepsilon C_2\varkappa^{1/2}$. The formal definition of the o-symbol reads as follows.

DEFINITION 1. Let f and g be two scalar functions defined on a set $R \subset \boldsymbol{R}^l$ $(l \in N)$. Then

$$(20)\qquad f(x) = o, \quad [g(x) \to 0]$$

means

$$(21)\qquad \exists\, K > 0\ \exists\, \delta > 0\colon\ |g(x)| < \delta \Rightarrow |f(x)| \leqslant K|g(x)|. \ \blacksquare$$

The constants K and δ are supposed to be numerical constants not depending on ε, $\varkappa$, C_1 and C_2. The expression between square brackets is referred to as the *restriction* under which (20) holds. The relation (20) only supplies information to those x for which $g(x)$ is small. We remark that we do not define the meaning of o itself; as it is often done in asymptotic analysis (cf. De Bruijn [2]), we only give the interpretation of some complete formulas. For instance, if we write for two scalar functions f_1 and f_2, with $f_2(x) > 0$ $(x \in R)$,

$$(22) \qquad f_1(x) \leqslant f_2(x)(1+o), \quad [g(x) \to 0],$$

then we mean that there exists a scalar function f_3 defined on D such that

$$(23) \qquad \frac{f_1(x) - f_2(x)}{f_2(x)} \leqslant f_3(x) \quad (x \in D) \quad \text{and} \quad f_3(x) = o, \ [g(x) \to 0].$$

If the o-symbol appears in some compound formula or at both sides of an equality or inequality relation, then the o-symbol has to be interpreted as a class of functions. For instance, if we write $f_1(x)o = o$, $[g(x) \to 0]$, then this has to be interpreted as follows. For any function f_2 for which $f_2(x) = o$, $[g(x) \to 0]$ one also has $f_1(x)f_2(x) = o$, $[g(x) \to 0]$. Some rather trivial but often used properties are

$$(24) \qquad \begin{cases} g(x) = o, & [g(x) \to 0], \\ o + o = o, & [g(x) \to 0], \\ oo = o, & [g(x) \to 0], \\ (1+o)^{-1} = 1 + o, & [g(x) \to 0]. \end{cases}$$

These properties indicate that the o-symbol is easy to handle and that is our main reason for using it. Woźniakowski [5] uses the relation $\doteq$ in order to simplify error estimates. Instead of (19) he would write

$$\varepsilon(1 + C_1 \varkappa + 2\varepsilon C_2 \varkappa^{1/2}) \doteq \varepsilon(1 + C_1 \varkappa).$$

However, using this notation without mentioning the underlying restriction one looses significant information concerning uniformity with respect to the relevant parameters as we will see from the analysis in Subsection 3.2. A disadvantage of the use of o-symbols (and also of the use of the relation $\doteq$) is that we do not obtain explicit numerical constants in error estimates. However, in all cases where we derive formulas with o-symbols it is possible to retrace the proof, replacing all o-symbols 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 (cf. Subsection 3.3). However, these definite numbers are rather arbitrary whereas the coefficients in the relations involving o- and $\doteq$-symbols are more or less uniquely determined (compare $1+\varepsilon \leqslant (1-\varepsilon)^{-1} \leqslant 1+(1+1/3)\varepsilon$, $[0<\varepsilon<1/4]$ and $(1-\varepsilon)^{-1} = 1+(1+o)\varepsilon$, $[\varepsilon \to 0]$).

Good-behavior. To denote the quality of the approximate solution computed by an iterative method with floating point arithmetic we use the concept of good-behavior (cf. e.g. Woźniakowski [5]).

DEFINITION 2. An iterative method for solving a linear system $Ax = b$ is said to be *well-behaved* (or, equivalently, *has good-behavior*) if for all initial points x_0 the computed sequence $\{x_i\}$ contains at least one approximation x_i such that

$$(25) \qquad (A+\delta A)x_i = b, \qquad \|\delta A\| \leqslant g\varepsilon\|A\|,$$

for some matrix δA, where g depends only on the dimension of the system. ■

In view of relation (25), good-behavior means that the computed approximate solution x_i is the exact solution of a slightly perturbed system. From a practical point of view, this solution x_i is satisfactory since the elements of the machine matrix A itself in general cannot be a better representation of the elements of the original matrix than with relative precision ε. Therefore, the corresponding error in x_i is *inherent* for the system $Ax = b$. It is easy to verify that (25) is equivalent to the assertion that the *residual* $\|A(\hat{x}-x_i)\|$ satisfies

$$(26) \qquad \|A(\hat{x}-x_i)\| \leqslant g\varepsilon\|A\|\,\|x_i\|.$$

It can also easily be seen that (26) implies

$$(27) \qquad \|A^{1/2}(\hat{x}-x_i)\| \leqslant g\varepsilon\varkappa^{1/2}\|A^{1/2}\|\,\|x_i\|,$$

which, in turn, implies

$$(28) \qquad \|\hat{x}-x_i\| \leqslant g\varepsilon\varkappa\|x_i\|,$$

but the implications do not hold in general vice versa. So, if x_i is the exact solution of a slightly perturbed system in the sense of (25), then the error $\|\hat{x}-x_i\|$ can be of order $\varepsilon\varkappa\|x_i\|$ and therefore this is called the *inherent error*. For similar reasons $\varepsilon\varkappa^{1/2}\|A^{1/2}\|\,\|x_i\|$ is called the *inherent natural error*. An iterative method that computes for all initial points x_0 an approximation x_i whose error $\|\hat{x}-x_i\|$ is at most of the order of the inherent error is called *numerically stable* (cf. Woźniakowski [5]). Thus a well-behaved method certainly is numerically stable but the reverse is not necessarily true.

3.2. The two basic theorems. Before starting off the round-off error analysis of the GM we first deduce some auxiliary results concerning the computation of a residual vector $b-Ax$. These results will be used in the subsequent considerations.

LEMMA 1. *Let b, x be two machine vectors and let*

$$\left\{\begin{array}{ll}\hat{r} := b - Ax \neq 0, & r := \mathrm{fl}\big(b - \mathrm{fl}(Ax)\big),\\ \varphi := \|A\|\,\|x\|/\|\hat{r}\|, & \psi := \|A^{1/2}\|\,\|x\|/\|A^{-1/2}\hat{r}\|.\end{array}\right. \tag{29}$$

Then we have

$$(\hat{r}, r) = \|\hat{r}\|^2(1+o) = \|r\|^2(1+o), \quad [\varepsilon(1+C_1\varphi)\to 0], \tag{30}$$

$$(\hat{r}, A^{-1}r) = \|A^{-1/2}\hat{r}\|^2(1+o) = \|A^{-1/2}r\|^2(1+o), \qquad [\varepsilon\varkappa^{1/2}(1+C_1\psi)\to 0]. \tag{31}$$

Proof. According to (17) and (18) we have

$$\left\{\begin{array}{l}r = \mathrm{fl}\big(b - \mathrm{fl}(Ax)\big) = (I+F)\big(b-(A+E)x\big) = \hat{r} + \delta r,\\ \delta r := F(b-Ax) - (I+F)Ex.\end{array}\right. \tag{32}$$

Consequently,

$$\|\delta r\| \leqslant \varepsilon\|\hat{r}\| + \varepsilon(1+\varepsilon)C_1\|A\|\,\|x\| \tag{33}$$

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

$$\left\{\begin{array}{l}\|\delta r\|/\|\hat{r}\| \leqslant \varepsilon\big(1 + C_1\varphi(1+o)\big),\\ \|A^{-1/2}\delta r\|/\|A^{-1/2}\hat{r}\| \leqslant \varepsilon\varkappa^{1/2}\big(1 + C_1\psi(1+o)\big),\end{array}\right. \tag{34}$$

which can be weakened to

$$\left\{\begin{array}{ll}\|\delta r\|/\|\hat{r}\| = o, & [\varepsilon(1+C_1\varphi)\to 0],\\ \|A^{-1/2}\delta r\|/\|A^{-1/2}\hat{r}\| = o, & [\varepsilon\varkappa^{1/2}(1+C_1\psi)\to 0].\end{array}\right. \tag{35}$$

The first equalities in (30) and (31) follow immediately from (32) and the appropriate inequality of (35). The second equalities follow from the fact that for $l = 0, -\frac{1}{2}$ we have

$$\big|\|A^l r\| - \|A^l\hat{r}\|\big| \leqslant \|A^l\delta r\| \leqslant \|A^l\hat{r}\|o, \tag{36}$$

under the appropriate restriction. ∎

We are now ready to deduce the first basic theorem where the influence of round-off on relation (6) is expressed in terms of a relative error.

THEOREM 3. *Let x_i be an arbitrary machine vector for which $\hat{r}_i := b - Ax_i \neq 0$ and let x_{i+1} be computed from one step GM. Let*

$$r_i := \mathrm{fl}\big(b - \mathrm{fl}(Ax_i)\big), \tag{37}$$

$$\varphi_i := \|A\|\,\|x_i\|/\|\hat{r}_i\|. \tag{38}$$

Then we have

$$\frac{\|A^{1/2}(\hat{x}-x_{i+1})\|^2}{\|A^{1/2}(\hat{x}-x_i)\|^2} = 1-\hat{\gamma}_i^2(1+\nu_{i+1}), \tag{39}$$

where

$$\hat{\gamma}_i := \frac{|(\hat{r}_i, r_i)|}{\|A^{-1/2}\hat{r}_i\|\,\|A^{1/2}r_i\|} \tag{40}$$

and

$$|\nu_{i+1}| \leqslant 4\varepsilon\{1+\varkappa^{1/2}+\varphi_i\}(1+o)+\varepsilon\{\varkappa^{1/2}(C_2+C_1\varkappa^{1/2})+C_1\varphi_i\}o, \tag{41}$$

under the restriction

$$\varepsilon\{\varkappa^{1/2}(1+C_2+C_1\varkappa^{1/2})+(1+C_1)\varphi_i\}\to 0. \tag{42}$$

Proof. From (32) and (34) we know that the computed vector r_i satisfies

$$\begin{cases} r_i = \hat{r}_i+\delta r_i, \\ \|\delta r_i\|/\|\hat{r}_i\| \leqslant \varepsilon(1+C_1\varphi_i)(1+o), \qquad [\varepsilon\to 0]. \end{cases} \tag{43}$$

We consider the computation of a_i from (2).

$$\begin{cases} \mathrm{fl}\big((r_i, r_i)\big) = \big((I+D_i')r_i, r_i\big) = (r_i, r_i)(1+\lambda_i), \\ |\lambda_i| = |(D_i'r_i, r_i)|/(r_i, r_i) \leqslant \varepsilon C_2. \end{cases} \tag{44}$$

Further we have

$$\begin{cases} \mathrm{fl}\big((r_i, Ar_i)\big) = \big((I+D_i'')r_i, (A+E_i)r_i\big) = (r_i, Ar_i)(1+\mu_i), \\ |\mu_i| = \big|(D_i''r_i, Ar_i)+\big((I+D_i'')r_i, E_ir_i\big)\big|/(r_i, Ar_i) \\ \qquad \leqslant \{\varepsilon C_2\|r_i\|\,\|Ar_i\|+\varepsilon C_1\|A\|\,\|r_i\|^2(1+o)\}/\|A^{1/2}r_i\|^2 \\ \qquad \leqslant \varepsilon C_2\varkappa^{1/2}+\varepsilon C_1\varkappa(1+o), \qquad [\varepsilon C_2\to 0]. \end{cases} \tag{45}$$

This yields

$$a_i = \mathrm{fl}\left(\frac{\mathrm{fl}\big((r_i, r_i)\big)}{\mathrm{fl}\big((r_i, Ar_i)\big)}\right) = \frac{(r_i, r_i)(1+\lambda_i)}{(r_i, Ar_i)(1+\mu_i)}(1+\varepsilon_i), \tag{46}$$

with $|\varepsilon_i| \leqslant \varepsilon$.

Hence,

$$\begin{cases} a_i = \dfrac{(r_i, r_i)}{(r_i, Ar_i)}(1+\delta a_i'), \\ |\delta a_i'| = |\lambda_i-\mu_i+\varepsilon_i+\lambda_i\varepsilon_i|/|1+\mu_i| \\ \qquad \leqslant \varepsilon\big(1+C_2(1+\varkappa^{1/2})+C_1\varkappa\big)(1+o), \end{cases} \tag{47}$$

under the restriction

$$\varepsilon\varkappa^{1/2}(C_2+C_1\varkappa^{1/2})\to 0. \tag{48}$$

Since, from (30) and (43),

$$\begin{cases}(r_i, r_i) = (\hat{r}_i, r_i)+(\delta r_i, r_i) = (\hat{r}_i, r_i)(1+\tau_i), \\ |\tau_i| = \dfrac{|(\delta r_i, r_i)|}{|(\hat{r}_i, r_i)|} \leqslant \dfrac{\|\delta r_i\|}{\|\hat{r}_i\|}\,\dfrac{\|\hat{r}_i\|\,\|r_i\|}{|(\hat{r}_i, r_i)|} = \dfrac{\|\delta r_i\|}{\|\hat{r}_i\|}(1+o) \\ \qquad \leqslant \varepsilon(1+C_1\varphi_i)(1+o), \qquad [\varepsilon(1+C_1\varphi_i)\to 0],\end{cases} \tag{49}$$

we obtain from (47)

$$\begin{cases}a_i = \hat{a}_i(1+\delta a_i''), \\ \hat{a}_i := (\hat{r}_i, r_i)/(r_i, Ar_i), \\ |\delta a_i''| = |\tau_i(1+\delta a_i')+\delta a_i'| \leqslant |\tau_i|(1+o)+|\delta a_i'| \\ \qquad \leqslant \varepsilon\big(2+C_1\varphi_i+C_2(1+\varkappa^{1/2})+C_1\varkappa\big)(1+o),\end{cases} \tag{50}$$

under the restriction

$$\varepsilon(1+C_2\varkappa^{1/2}+C_1\varkappa+C_1\varphi_i)\to 0. \tag{51}$$

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

$$\begin{cases}x_{i+1} = \mathrm{fl}\big(x_i+\mathrm{fl}(a_ir_i)\big) = (I+F_i'')\big(x_i+(I+F_i')a_ir_i\big) \\ \qquad = x_i+\hat{a}_ir_i+\delta x_{i+1}, \\ \delta x_{i+1} = F_i''x_i+\hat{a}_i(\delta a_i''r_i+M_ir_i), \\ \|M_i\| = \big\|(1+\delta a_i'')\big(F_i'+F_i''(I+F_i')\big)\big\| \leqslant 2\varepsilon(1+o),\end{cases} \tag{52}$$

under the restriction (51).

From the first equality in (52) it follows that

$$A^{1/2}(\hat{x}-x_{i+1}) = A^{1/2}(\hat{x}-x_i)-\hat{a}_iA^{1/2}r_i-A^{1/2}\,\delta x_{i+1}, \tag{53}$$

and hence, by taking squared norms of both sides,

$$\|A^{1/2}(\hat{x}-x_{i+1})\|^2 = \|A^{1/2}(\hat{x}-x_i)-\hat{a}_iA^{1/2}r_i\|^2 - \\ -2(\hat{r}_i-\hat{a}_iAr_i,\ \delta x_{i+1})+\|A^{1/2}\,\delta x_{i+1}\|^2. \tag{54}$$

From the definition of $\hat{a}_i$ we obtain

$$\|A^{1/2}(\hat{x}-x_{i+1})\|^2 = \|A^{1/2}(\hat{x}-x_i)\|^2-(\hat{r}_i, r_i)^2/\|A^{1/2}r_i\|^2 - \\ -2(\hat{r}_i-\hat{a}_iAr_i,\ \delta x_{i+1})+\|A^{1/2}\,\delta x_{i+1}\|^2, \tag{55}$$

which leads to the basic formula

$$\frac{\|A^{1/2}(\hat{x}-x_{i+1})\|^2}{\|A^{1/2}(\hat{x}-x_i)\|^2} = 1-\hat{\gamma}_i^2(1+\nu_{i+1}), \tag{56}$$

where $\hat{\gamma}_i$ is defined by (40) and

$$\nu_{i+1} := \{2(\hat{r}_i - \hat{a}_i A r_i,\ \delta x_{i+1}) - \|A^{1/2}\,\delta x_{i+1}\|^2\}/\{\hat{a}_i(\hat{r}_i, r_i)\}. \tag{57}$$

It remains to be proved that ν_{i+1} satisfies (41) under the restriction (42). Note that $(\hat{r}_i - \hat{a}_i A r_i, r_i) = 0$ and therefore the term $\hat{a}_i \delta a_i'' r_i$ in the relation for δx_{i+1} in (52) cancels when evaluating the inner product in the numerator of (57). Consequently, from (38), (50) and (52) we obtain, evaluating term by term,

$$\begin{aligned} &|(\hat{r}_i - \hat{a}_i A r_i,\ \delta x_{i+1})|/|\hat{a}_i(\hat{r}_i, r_i)| \\ &\leqslant \|\hat{r}_i\|\cdot\|F_i''\|\,\|x_i\|\,\|A^{1/2} r_i\|^2/(\hat{r}_i, r_i)^2 + \|A r_i\|\,\|F_i''\|\,\|x_i\|/|(\hat{r}_i, r_i)| + \\ &\qquad + \|\hat{r}_i\|\,\|M_i\|\,\|r_i\|/|(\hat{r}_i, r_i)| + \|A r_i\|\,\|M_i\|\,\|r_i\|/\|A^{1/2} r_i\|^2 \\ &\leqslant (\|F_i''\|\varphi_i + \|F_i''\|\varphi_i + \|M_i\| + \|M_i\|\varkappa^{1/2})(1+o) \\ &\leqslant 2\varepsilon(1+\varkappa^{1/2}+\varphi_i)(1+o), \end{aligned} \tag{58}$$

under the restriction (51).

We still have to estimate the second order term in (57). 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 $(a+b)^2 \leqslant 2(a^2+b^2)$ and $(a+b+o+d)^2 \leqslant 4(a^2+b^2+c^2+d^2)$ we find from (38), (50) and (52)

$$\begin{aligned} &\|A^{1/2}\,\delta x_{i+1}\|^2/|\hat{a}_i(\hat{r}_i, r_i)| \\ &\leqslant 4\{\|F_i''\|^2\,\|A\|\,\|x_i\|^2\,\|A^{1/2} r_i\|^2/(\hat{r}_i, r_i)^2 + (\delta a_i'')^2 + \\ &\qquad + \|M_i\|^2\,\|A\|\,\|r_i\|^2/\|A^{1/2} r_i\|^2\} \\ &\leqslant 4\bigl\{\varepsilon^2\varphi_i^2 + 4\varepsilon^2\bigl(4 + C_i^2\varphi_i^2 + 2C_2^2(1+\varkappa) + C_1^2\varkappa^2\bigr) + 4\varepsilon^2\varkappa\bigr\}(1+o), \end{aligned} \tag{59}$$

under the restriction (51).

So, finally we obtain from (57), (58) and (59)

$$\begin{aligned} |\nu_{i+1}| &\leqslant 4\varepsilon\{1+\varkappa^{1/2}+\varphi_i\}(1+o) + \\ &\qquad + 4\varepsilon^2\{4(4+\varkappa) + 8C_2^2(1+\varkappa) + 4C_1^2\varkappa^2 + (1+4C_1^2)\varphi_i^2\}(1+o), \end{aligned} \tag{60}$$

under the restriction (51). As this inequality can be written in the more compact form (41) under the restriction (42) we have proved Theorem 3. ■

Without giving the proof (which is very similar to the proof of Theorem 3) we state the second basic theorem which is the analogue of Theorem 2 in the presence of round-off.

THEOREM 4. *Let x_i be an arbitrary machine vector for which $\hat{r}_i := b - Ax_i \neq 0$ and let x_{i+1} be computed from one step GM. Let*

$$r_i := \mathrm{fl}\bigl(b - \mathrm{fl}(Ax_i)\bigr), \tag{61}$$

and

$$\psi_i := \|A^{1/2}\|\,\|x_i\|/\|A^{-1/2}\hat{r}_i\|. \tag{62}$$

Then we have

$$\frac{\|\hat{x}-x_{i+1}\|^2}{\|\hat{x}-x_i\|^2} = 1-\hat{\sigma}_i(2-\hat{\varrho}_i+\eta_{i+1}), \tag{63}$$

where

$$\hat{\sigma}_i := \frac{(\hat{r}_i, r_i)\,\|A^{-1/2}r_i\|^2}{\|A^{1/2}r_i\|^2\,\|A^{-1}\hat{r}_i\|^2}, \tag{64}$$

$$\hat{\varrho}_i := \frac{(\hat{r}_i, r_i)\,\|r_i\|^2}{\|A^{1/2}r_i\|^2(\hat{r}_i, A^{-1}r_i)}, \tag{65}$$

and

$$|\eta_{i+1}| \leqslant 2\varepsilon\{2\varkappa^{1/2}(1+C_2+C_1\varkappa^{1/2})+2(3+C_2)+ \\ +\big(1+\varkappa^{1/2}(1+2C_1)\big)\psi_i\}(1+o) \tag{66}$$

under the restriction

$$\varepsilon\{\varkappa^{1/2}(1+C_2+C_1\varkappa^{1/2})+(1+C_1\varkappa^{1/2})\psi_i\}\to 0. \;\blacksquare \tag{67}$$

3.3. Good-behavior of the gradient method. If we compare Theorems 1 and 3 and Theorems 2 and 4, then we observe that the perturbed GM behaves at each step more or less like the exact GM for small values of $\varepsilon > 0$. Note that $\hat{\gamma}_i\to\gamma_i$, $\nu_i\to 0$, $\hat{\sigma}_i\to\sigma_i$, $\hat{\varrho}_i\to\varrho_i$ and $\eta_i\to 0$ if $\varepsilon\to 0$. We ask what conclusion can be drawn from the results of Subsection 3.2 as far as the step-wise linear convergence of the natural error and the error are concerned for the perturbed GM. We first concentrate on the natural error and thus on Theorem 3. From relation (39) we conclude that the natural error decreases at the step from i to $i+1$ if and only if $\nu_{i+1} > -1$ and $\hat{\gamma}_i^2 > 0$. Apparently, from (41) and (42), $|\nu_{i+1}| < 1$ if the left-hand part of (42) is small. From the definition of $\hat{\gamma}_i$ and the relations (30) we obtain

$$\hat{\gamma}_i^{-2} \leqslant \varkappa\frac{\|\hat{r}_i\|^2\,\|r_i\|^2}{(\hat{r}_i, r_i)^2} = \varkappa\frac{\|\hat{r}_i\|^4}{(\hat{r}_i, \hat{r}_i)^2}(1+o) = \varkappa(1+o) \tag{68}$$

under the restriction $\varepsilon(1+C_1\varphi_i)\to 0$ and thus under the restriction (42). Therefore, not only $\nu_{i+1} > -1$ but also $\hat{\gamma}_i$ is bounded away from zero if the left-hand part of (42) is small. Hence the natural error decreases by a factor close to $1-\gamma_i^2$ if (42) is small. Unfortunately, the relations (30) and the restriction (42) do not supply explicit bounds for ν_{i+1} and $\hat{\gamma}_i$. Here we encounter a situation where this disadvantage of the

o-notation emerges. On the other hand, as we said already in subsection 3.1, one can easily retrace the proofs and replace all o-symbols by definite estimates involving explicit numerical constants. For instance, one can prove (cf. Bollen [1]) that if

$$\varepsilon\{\varkappa^{1/2}(1+C_2+C_1\varkappa^{1/2})+2(1+C_2)\} \leqslant 1/40 \tag{69}$$

and

$$\varepsilon\{6+C_1\}\varphi_i \leqslant 1/8, \tag{70}$$

then we have

$$|\nu_{i+1}| \leqslant 4/5, \quad \hat{\gamma}_i^{-2} \leqslant 4\varkappa. \tag{71}$$

The restrictions (69) and (70) were chosen quite arbitrarily and the bounds in (71) were obtained by a rather rough estimate; they can easily be improved. Hence as an (as far as the factor 1/20 concerns rather arbitrary) explicit version of Theorem 3 we have the following statement for the GM.

PROPOSITION 1. *If x_{i+1} is computed from one step GM based on an arbitrary machine vector x_i and if furthermore* (69) *and* (70) *are satisfied, then we have*

$$\frac{\|A^{1/2}(\hat{x}-x_{i+1})\|^2}{\|A^{1/2}(\hat{x}-x_i)\|^2} \leqslant 1-\frac{1}{20\varkappa}. \quad \blacksquare \tag{72}$$

We note that inequality (69) only depends on the machine, the implementation and the matrix involved whereas inequality (70) also depends on x_i since this inequality is equivalent to the inequality

$$\|b-Ax_i\| \geqslant 8\varepsilon(6+C_1)\|A\|\,\|x_i\|. \tag{73}$$

Now assume that (69) is satisfied. Then Proposition 1 leads to the following three important conclusions on the GM.

(i) As long as $\|b-Ax_i\| \geqslant 8\varepsilon(6+C_1)\|A\|\,\|x_i\|$, the natural error $\|A^{1/2}(\hat{x}-x_i)\|$ converges step-wise linearly with a ratio no greater than $\left(1-(20\varkappa)^{-1}\right)^{1/2}$.

(ii) If $\|A^{1/2}(\hat{x}-x_{i+1})\| \geqslant \|A^{1/2}(\hat{x}-x_i)\|$ holds for some $i \geqslant 0$, then

$$\|b-Ax_i\| \leqslant 8\varepsilon(6+C_1)\|A\|\,\|x_i\|.$$

(iii) There exists an $i \geqslant 0$ such that $\|b-Ax_i\| \leqslant 8\varepsilon(6+C_1)\|A\|\,\|x_i\|$. (Since otherwise (72) would hold for all $i \geqslant 0$ which would lead to the contradiction $\|b-Ax_i\| \to 0$ $(i \to 0)$.)

Combining these three conclusions into one statement we obtain the following result.

PROPOSITION 2. *If* (69) *is satisfied and if* $\{x_i\}$ *is generated by the GM with an arbitrary initial machine vector* x_0, *then the natural error* $\|A^{1/2}(\hat{x}-x_i)\|$ *converges step-wise linearly with a convergence ratio no greater than* $(1-(20\varkappa)^{-1})^{1/2}$, *at least until an iteration step where the residual satisfies*

$$\|b-Ax_i\| \leqslant 8\varepsilon(6+C_1)\|A\|\|x_i\|. \;\blacksquare \tag{74}$$

This implies that the GM is well-behaved and, consequently, numerically stable. Note that (74) can be used as a stopping criterion for the GM.

As far as the monotonicity of the error $\|\hat{x}-x_i\|$ for the perturbed GM is concerned we first concentrate on an explicit version of Theorem 4. One can prove that in (63) there holds $|\eta_{i+1}| \leqslant 7/10$ if

$$\varepsilon\{\varkappa^{1/2}(1+C_2+C_1\varkappa^{1/2})+(3+C_2)\} \leqslant 1/40 \tag{75}$$

and

$$\varepsilon\{1+\varkappa^{1/2}(1+2C_1)\}\psi_i \leqslant 1/8. \tag{76}$$

From Lemma 1 we obtain under the restriction $\varepsilon\varkappa^{1/2}(1+C_1\psi_i)\to 0$

$$\hat{\sigma}_i^{-1} = \frac{\|A^{1/2}r_i\|^2\|A^{-1}\hat{r}_i\|^2}{\|r_i\|^2\|A^{-1/2}\hat{r}_i\|^2}(1+o) \leqslant \varkappa(1+o) \tag{77}$$

and

$$\hat{\varrho}_i \leqslant \frac{\|A^{-1/2}r_i\|\|A^{-1/2}\hat{r}_i\|}{|(\hat{r}_i, A^{-1}r_i)|} = 1+o, \tag{78}$$

whereas, if (75) and (76) are satisfied, one can prove that the following explicit inequalities hold

$$0 < \hat{\sigma}_i^{-1} < \tfrac{14}{10}\varkappa, \qquad |\hat{\varrho}_i| < \tfrac{12}{10}. \tag{79}$$

Once more we note that inequality (75) only depends on the machine, the implementation and on the matrix involved, whereas inequality (76) is equivalent to the inequality

$$\|A^{1/2}(\hat{x}-x_i)\| \geqslant 8\varepsilon(1+\varkappa^{1/2}(2+C_1))\|A^{1/2}\|\|x_i\|. \tag{80}$$

As an explicit version of Theorem 4 we thus have the following property for the GM.

PROPOSITION 3. *If* x_{i+1} *is computed from one step GM based on an arbitrary machine vector* x_i *and if furthermore* (75) *and* (80) *are satisfied, then we have*

$$\frac{\|\hat{x}-x_{i+1}\|^2}{\|\hat{x}-x_i\|^2} \leqslant 1-\frac{1}{14\varkappa}. \;\blacksquare \tag{81}$$

From this proposition we can draw three conclusions in terms of step-wise linear convergence of the error $\|\hat{x}-x_i\|$ similar to those we derived for the natural error $\|A^{1/2}(\hat{x}-x_i)\|$ from Proposition 1. These conclusions are combined into the following statement.

PROPOSITION 4. *If* (75) *is satisfied and if* $\{x_i\}$ *is generated by the GM with an arbitrary initial vector* x_0, *then the error* $\|\hat{x}-x_i\|$ *converges step-wise linearly with a convergence ratio no greater than* $\left(1-(14\varkappa)^{-1}\right)^{1/2}$, *at least until an iteration step where the natural error satisfies*

$$\|A^{1/2}(\hat{x}-x_i)\| \leqslant 8\varepsilon\left(1+\varkappa^{1/2}(1+2C_1)\right)\|A^{1/2}\|\,\|x_i\|. \quad ■ \tag{82}$$

This implies that the monotonicity of the error cannot break down before the natural error reaches the level of the inherent natural error. Note that (82) does not imply that x_i is the solution of a slightly perturbed linear system.

4. Final remarks

1. The constants C_1 and C_2 do not show up in the first order part of estimate (41). In the error analysis they only appear in the relative error $\delta a_i''$ occurring at the computation of a_i. The objective function $F(x_i+ar_i)$ is quadratic in a and hence, if we are at a distance δ from the point at which this function attains its minimum, 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 (58), which explains the absence of C_1 and C_2. Formulas (41) and (42), however, show that $\varepsilon C_2\varkappa^{1/2}$ and $\varepsilon C_1(\varkappa+\varphi_i)$ have to be small in order to have ν_{i+1} small. A first order round-off error analysis would not have given this information.

2. From Proposition 2 it follows that C_2 has no influence on the reachable level for the residual. This can be explained as follows. Assume that only round-off occurs at the computation of inner products ($C_2 \neq 0$) and not at the basic dyadic arithmetical operations, i.e., vector addition, vector subtraction, scalar by vector product and scalar division nor at the matrix by vector product computations ($C_1 = 0$). Then, retracing the proof of Theorem 3, we obtain successively

$$\begin{cases} \delta r_i = 0, \quad r_i = \hat{r}_i, \quad |\lambda_i| \leqslant \varepsilon C_2, \quad |\mu_i| \leqslant \varepsilon C_2\varkappa^{1/2}, \\ |\delta a_i''| = |\delta a_i'| \leqslant \varepsilon C_2(1+\varkappa^{1/2})(1+o), \qquad [\varepsilon C_2\varkappa^{1/2}\to 0], \end{cases} \tag{83}$$

and

$$x_{i+1} = x_i+\hat{a}_i(1+\delta a_i'')\hat{r}_i. \tag{84}$$

We have (cf. (8)) for the exact GM

$$f(a) := F(x_i + a\hat{r}_i) = F(x_i) - 2a(\hat{r}_i, \hat{r}_i) + a^2 \|A^{1/2}\hat{r}_i\|^2, \tag{85}$$

which is a quadratic function in a and $f(0) = F(x_i)$. Since f is minimal at $\hat{a}_i = (\hat{r}_i, \hat{r}_i)/(\hat{r}_i, A\hat{r}_i)$, f is symmetric around $\hat{a}_i$ and consequently $f(a) < f(0)$ for all a satisfying $|\hat{a}_i - a| < |\hat{a}_i - 0|$. Stated differently, if $a = \omega\hat{a}_i$, for some $0 < \omega < 2$, then $F(x_i + a\hat{r}_i) < F(x_i)$. Hence, if instead of $a = \hat{a}_i$ one takes $a = \omega_i\hat{a}_i$, where $0 < \omega_i < 2$, then still the natural error decreases at the step from i to $i+1$. The factor ω_i could be called a *relaxation factor*.

If all relaxation factors satisfy the condition $|\omega_i - 1| \leqslant \delta$ for some $\delta \in (0, 1)$, then for this process a convergence result similar to (6) holds. Consequently, the factor $1 + \delta a_i''$ occurring in (84), which is due to perturbations at inner product computations, can be regarded upon as a relaxation factor for the exact GM. Hence, if for all i there holds $|\delta a_i''| \leqslant \delta$, then the natural error converges step-wise linearly to zero and therefore C_2 has no influence on the reachable level.

3. We performed (cf. Bollen [1]) several tests with the GM in order to verify our analytical results of Section 3 for the perturbed GM. In all experiments we observed the step-wise linear convergence of the natural error at least until an iteration step for which the residual $\|b - Ax_i\|$ was of order $\varepsilon\|A\|\,\|x_i\|$ and the step-wise linear convergence of the error at least until an iteration step for which the natural error was of order $\varepsilon\varkappa^{1/2}\|A^{1/2}\|\,\|x_i\|$. In most cases where we had to perform many iterations before the reachable level was attained the ultimate convergence ratios were close to $1 - \varkappa^{-1}$.

4. From statement (3) it follows that for the exact GM one has

$$b - Ax_{i+1} = (b - Ax_i) - a_i A r_i, \tag{86}$$

which gives the following recurrence relation for r_i

$$r_{i+1} = r_i - a_i A r_i. \tag{87}$$

Hence, in the algorithm for the exact GM the computation of r_{i+1} might as well be based on this recurrence relation. If, instead of (4), we use relation (87) for the computation of r_{i+1} $(i \geqslant 0)$, then this method is called the *recursive residual gradient method* (RRGM). Of course, if exact arithmetic is used, the approximations $\{x_i\}$ generated by the RRGM and the GM are exactly the same. However, this certainly is not the case when both methods are performed using floating point arithmetic. For the GM 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. For the RRGM the sequence $\{r_i\}$ can be computed without even computing the sequence $\{x_i\}$ and the difference between the recursively computed residual vector r_i and the exact residual vector $b-Ax_i$ s caused by computational round-off at all previous steps.

One can prove (cf. Bollen [1]) that for the perturbed RRGM the natural error $\|A^{-1/2}r_i\|$, expressed in terms of the recursively computed residual vectors r_i based on (87), converges step-wise linearly to zero with a convergence ratio no greater than $(1-(16\varkappa)^{-1})^{1/2}$ if the possibility of underflow is neglected and if

$$\varepsilon\{(1+\varkappa^{1/2})(1+C_2)+\varkappa(3+C_2)\} \leqslant 1/8. \tag{88}$$

We realize that from a practical point of view this is not a very interesting result since convergence of the rescursively computed residual vectors r_i has no direct practical implications. It does not imply that the computed approximations x_i tend to $\hat{x}$. 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, also in the presence of round-off. For results concerning the limiting behavior of the approximations $\{x_i\}$ generated by the perturbed RRGM we refer to Bollen [1].

5. Our definition of good-behavior guarantees that a well-behaved method computes at least one approximation x_i of a slightly perturbed iinear system in the sense of (25). The inequality $\|\delta A\| \leqslant g\varepsilon\|A\|$ does not imply that the approximation x_i is the exact solution of a perturbed linear system where the problem data (the elements of A) are elementwise relatively disturbed by a factor ε, i.e.,

$$(A+\delta A')x_i = b, \qquad |\delta A'| \leqslant g\varepsilon|A|, \tag{89}$$

where the inequality and the absolute values are to be understood in an elementwise sense. A method which computes at least one approximation x_i satisfying (89) is called a *strongly well-behaved* method. From the results of Skeel [4] it follows that a well-behaved method followed by one step iteration refinement in single precision is a strongly well-behaved method and hence the GM followed by one step iterative refinement in single precision computes at least one approximation x_i that satisfies (89).

6. In Bollen [1] we deduce results similar to those of Section 3 for general DM's by following the same strategy, viz., determining numerical analogues of fundamental convergence properties of the exact DM for the perturbed DM. As a special case we consider the conjugate gradient method of Hestenes and Stiefel [3] and we prove that this method computes at least one approximation x_i for which the residual $\|b-Ax_i\|$ is of order $\varepsilon\varkappa^{1/2}\|A\|\,\|x_i\|$.

References

[1] J. A. M. Bollen, *Round-off error analysis of descent methods for solving linear equations*, Doctoral dissertation, Eindhoven University of Technology, 1980.

[2] N. G. de Bruijn, *Asymptotic Methods in Analysis*, North-Holland, Amsterdam 1961.

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

[4] R. D. Skeel, *Iterative refinement implies numerical stability for Gaussian elimination*, Mathematics of Computation 35 (1980), 817–832.

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

Presented to the Semester
Computational Mathematics
February 20 — May 30, 1980