CHARLES UNIVERSITY, PRAGUE
Faculty of Mathematics and Physics
Department of Numerical Mathematics

ON SOME OPEN PROBLEMS IN KRYLOV SUBSPACE METHODS

Summary of Ph.D. Thesis
April 2002

Petr Tichý

Branch: M6–Scientific-technical computations

UNIVERZITA KARLOVA, PRAHA
Matematicko-fyzikální fakulta
Katedra numerické matematiky

O NĚKTERÝCH OTEVŘENÝCH PROBLÉMECH V KRYLOVOVSKÝCH METODÁCH

Autoreferát doktorské disertační práce
Duben 2002

Petr Tichý

Branch: M6–Vědecko-technické výpočty

Výsledky tvořící tuto práci byly získány během studia na MFF UK v Praze (1997-2000) a během práce v Ústavu Informatiky AV ČR v Praze (2000-2002).

Disertant: Mgr. Petr Tichý
Katedra numerické matematiky MFF UK
Matematicko-fyzikální fakulta Univerzity Karlovy
Sokolovská 83
186 75 Praha 8
&
Ústav Informatiky AV ČR
Pod vodárenskou věží 2
182 07 Praha 8

Školitel: Doc. RNDr. Jan Zítko, CSc.
Katedra numerické matematiky
Matematicko-fyzikální fakulta Univerzity Karlovy
Sokolovská 83
186 75 Praha 8

Školící pracoviště: Katedra numerické matematiky MFF UK a ÚI AV ČR

Oponenti: Prof. RNDr. Zdeněk Dostál, CSc.
Katedra aplikované matematiky
Fakulta elektrotechniky a informatiky VŠB Ostrava
17. listopadu 15
708 33 Ostrava-Poruba

Dr. Ing. Miroslav Rozložník
Ústav Informatiky AV ČR
Pod vodárenskou věží 2
182 07 Praha 8

Autoreferát byl rozeslán dne

Obhajoba disertace se koná dne v hodin před komisí pro obhajoby doktorských disertačních prací v oboru M6–Vědecko-technické výpočty, na MFF UK, Ke Karlovu 3, Praha 2, v místnosti č. .

S disertací je možno se seznámit na Studijním oddělení doktorandského studia MFF UK, Ke Karlovu 3, Praha 2.

Předseda oborové rady M6: Doc. RNDr. K. Najzar, CSc.
Katedra numerické matematiky MFF UK
Sokolovská 83
186 75 Praha 8

Contents

Introduction

Many real-world problems can be described (using some knowledge from other areas of science) in a mathematical language. For example, a part of reality can be modeled by a system of integral or differential equations. These equations usually describe the real-world problem approximately (the problem is idealized by omitting unimportant details). A solution of this system of integral or differential equations lies in an infinite dimensional space and it is, in general, analytically uncomputable. Therefore we formulate a dicretized problem and look for an approximate solution in a finite dimensional space. After a possible linearization we obtain a system of linear equations

$$\mathbf{A}x = b, \tag{1}$$

one of the basic problems of numerical linear algebra. The whole solution process combines tools from mathematical modeling, applied mathematics, numerical analysis, numerical methods and numerical linear algebra (for more details see [11]). At all stages, the approximation steps are accompanied by errors and the whole solution process should be well-balanced to avoid wasting human and computer resources. When the problem is approximated on one stage with some level of accuracy it does not make a sense to solve the corresponding problem on a next stage with a substantially different accuracy. Therefore, iterative methods are often very suitable for solving the system (1). We can stop iteration process at any iteration step (when the required accuracy level is reached). Moreover, the sparse structure of matrices allows to solve large systems of millions of unknowns without transforming the system matrix or even without forming it.

We mentioned one of the possible processes of formation of systems of linear equations. Of course, systems of linear equations can arise in many ways and in many applications.

This thesis is devoted to Krylov subspace methods for solving the system (1), . It seems that the combination of preconditioning and Krylov subspace methods is suitable and effective for solving the problem described above. The thesis consists of 5 chapters.

The first chapter is introductory. Based on [10], [8] and [5] it gives an overview of the most important Krylov subspace methods. Though there are many Krylov subspace methods and algorithms, they all are based only on a few principles. In the last section we discuss the stopping criteria of algorithms.

The Lanczos method for solving unsymmetric systems of linear equations (LM) is one of the possible generalizations of the CG method to unsymmetric systems. We deal with this method in the second chapter. The goal of this chapter, which contains original results, is to contribute to answering an open question: What is the relationship between the methods with short-term and long-term recurrences? In particular, we wish to contribute to understanding of the role of the shadow vector in the Lanczos process. We formulate a theorem about the relation between general three-term recurrence (the coefficients can be chosen arbitrarily) and the Lanczos three-term recurrence (the coefficients are determined by the orthogonality condition). We explain that it is possible to determine the shadow vector such

that the algorithm of the Lanczos method computes selected residuals of another Krylov subspace method. We discuss the question why the convergence curves of classical Krylov subspace methods are often very close to the convergence curve of the GMRES method.

The third, fourth and fifth chapters consist of original results submitted for publication [12] (joint work with Z. Strakoš) as well as their extensions.

In the third chapter we deal with the estimation of the **A**-norm of the error (**A** is symmetric and positive definite matrix) in the conjugate gradient method (CG). We want to bring more light into the problem of estimating of the **A**-norm of the error. Based on the connection CG to Gauss quadrature, we show how to construct lower estimates of the **A**-norm of the error. We explain that lower estimate based on Gauss quadrature is mathematically equivalent to the estimates derived by algebraical way and also to the original formula of Hestenes and Stiefel [9].

The goal of the fourth chapter is to explain the behaviour of CG in finite precision arithmetic and to prepare basis for the rounding error analysis of estimates of the **A**-norm of the error. We describe the basic idea of mathematical model of CG in finite precision arithmetic based on understanding CG in the sense of Gauss quadrature. The description and the bounds of the rounding errors arising in the CG iterates are presented. A new theorem about the local orthogonality in finite precision arithmetic closes the theoretical part of this chapter. In the numerical experiments, we deal with the actual size of rounding errors in the CG iterates.

Our goal in the fifth chapter is to explain the problem of application of mathematical formulas (derived in exact precision arithmetic) in finite precision arithmetic, and to present rounding error analysis of the formulas that we use for estimating of the **A**-norm of the error. We extend rounding error analysis of the favoured lower bound (given by Hestenes' and Stiefel's formula [9]) presented in our paper [12] and prove that it is numerically stable. We also present detailed rounding error analysis of the new algebraically derived estimate. Our results are illustrated by numerical experiments. We describe also an estimate for the euclidean norm of the error based on [9, Theorem 6:3] and demonstrate numerically the possibility of application of this estimate in finite precision arithmetic.

I. Krylov subspace methods

This chapter gives an overview of Krylov subspace methods. Krylov subspace method is a special case of projective method. It determines approximate solution x_k such that

$$x_k \in x_0 + \mathcal{K}_k(\mathbf{A}, r_0), \quad b - \mathbf{A}x_k \perp \mathcal{L}_k, \tag{2}$$

where x_0 is the initial approximation, $\mathcal{L}_k$ is a space of dimension k and $\mathcal{K}_k(\mathbf{A}, r_0)$ denotes k-th Krylov subspace,

$$\mathcal{K}_k(\mathbf{A}, r_0) \equiv \mathrm{span}\{r_0, \mathbf{A}r_0, \ldots, \mathbf{A}^{k-1}r_0\}.$$

Various Krylov subspace methods are determined by the choice of spaces $\mathcal{L}_k$. In order to work with Krylov subspaces we need to choose their appropriate basis. In

our work we consider two types of basis: *Arnoldi's basis* and *Lanczos' basis*. Arnoldi's basis is an appropriately chosen orthonormal basis and the computation of the basis is, in general, expensive to computer resources (memory and computational costs). Vectors of Lanczos' basis satisfy

$$v_{k+1} \perp \mathcal{K}_k(\mathbf{A}^T, \widetilde{r}_0)$$

where $\widetilde{r}_0$ is an auxiliary nonzero vector (*the shadow vector*). The computation of these basis vectors is not expensive to computer resources.

Denoting by $\mathbf{V}_{k+1} = [v_1, \ldots, v_{k+1}]$ the n by $k+1$ matrix having the basis vectors $v_1, \ldots, v_{k+1}$ as its columns, any vector $r_k \equiv b - \mathbf{A}x_k$ from the linear manifold $r_0 + \mathbf{A}\mathcal{K}_k(\mathbf{A}, r_0) \subset \mathcal{K}_{k+1}(\mathbf{A}, r_0)$ can be written in the form

$$r_k = \mathbf{V}_{k+1}\, q_k.$$

The vector $q_k \in \mathbb{R}^{k+1}$ is the coordinate vector of residual r_k in the basis $v_1, \ldots, v_{k+1}$ and we call it *quasi-residual*. When the quasi-residual q_k is determined such that the vector r_k is a multiple of the last basis vectors v_{k+1}, we speak about *Galerkin's quasi-residual*. The vector q_k with minimal norm from the all possible quasi-residuals is called *minimal quasi-residual*.

It is clear, from the facts given above, that in order to determine residuals and approximate solutions (using appropriately chosen basis and quasi-residual) it is necessary to resolve two problems: the construction of basis vectors and the construction of residuals and approximate solutions from the basis vectors. Algorithmic formulation of Krylov subspace methods is not unique and depends on algorithms that realize individual stages of computations. The choice of appropriate algorithms is important mainly in connection with the usage of algorithms in finite precision arithmetic.

We present four basic Krylov subspace methods GMRES, FOM, QMR and LM, and deal with their implementations. From the algorithm BiCG (this algorithm is an implementation of the Lanczos metod), we derive algorithms of Lanczos-type product methods (CGS, BiCGStab, TFQMR) and explain the basic idea of hybrid BiCG-methods (e.g. BiCGStab(l)).

In the end of the chapter, we discuss the stopping criteria of algorithms. We explain in the context of perturbation theory that if no other (more relevant and more sophisticated) criterion is available then *normwise relative backward error*

$$\beta(x_k) \equiv \frac{\|r_k\|}{\|b\| + \|\mathbf{A}\|\|x_k\|}$$

should be preferred to the (relative) residual norm $\|r_j\|/\|r_0\|$. The number $\beta(x_k)$ denotes the minimal size of perturbations $\Delta\mathbf{A}$ and Δb in the matrix $\mathbf{A}$ and in the vector b ($\|\Delta\mathbf{A}\| \leq \beta(x_k)\|\mathbf{A}\|$, $\|\Delta b\| \leq \beta(x_k)\|b\|$) such that x_k is the exact solution of the perturbed system $(\mathbf{A} + \Delta\mathbf{A})x_k = b + \Delta b$.

II. On the shadow vector in the Lanczos method

In the previous chapter, we explained that it is possible to compute effectively the basis vectors

$$v_{k+1} \perp \mathcal{K}_k(\mathbf{A}^T, \widetilde{r}_0) \tag{3}$$

where $\widetilde{r}_0$ is an auxiliary nonzero vector, often called the shadow vector or also the left starting vector. The recurrence for computing basis vectors v_{k+1} is reduced to three-term recurrence and it is possible to compute the following basis vector only from the basis vectors given by previous two iterations. As a consequence, we can compute the residuals and the approximations given by the conditions

$$x_k \in x_0 + \mathcal{K}_k(\mathbf{A}, r_0), \qquad r_k \perp \mathcal{K}_k(\mathbf{A}^T, \widetilde{r}_0), \tag{4}$$

with short-term recurrences (by the BiCG algorithm). We defined some other methods (QMR, CGS, BiCGStab and so on) that are based on the basic condition (3) and take advantage of the Lanczos' basis. All these methods are not expensive on computer resources and the quality of the computed approximate solution is often comparable with the quality of the approximate solution given by long-term recurrence of the GMRES or FOM methods. The methods that use the Lanczos' basis often give the possibility to gain "good" approximate solution from the linear manifold $x_0 + \mathcal{K}_k(\mathbf{A}, r_0)$ by low expenses on computer resources. However, the convergence of these methods is not well understood.

We wish to contribute to understanding of the role of the shadow vector $\widetilde{r}_0$ in the Lanczos process. One of the possibilities how to look at this problem is following.

Consider the *general three-term recurrence*

$$\mathbf{t} \;=\; \mathbf{A}v_k - \alpha_k v_k - \beta_{k-1} v_{k-1}, \quad \gamma_k = \|\mathbf{t}\|, \quad v_{k+1} = \mathbf{t}/\gamma_k, \tag{5}$$

$v_1 \equiv r_0/\|r_0\|$, $\beta_0 = 0$, $v_0 = \mathbf{o}$ where α_k and β_{k-1} are arbitrary coefficients. The unsymmetric Lanczos algorithm 1 computes Lanczos basis vectors by tree-term recurrence (5) and the coefficients α_k and β_{k-1} are determined by the orthogonality condition (3); then we call this recurrence *the Lanczos recurrence*. Various shadow vectors $\widetilde{r}_0$ determine various coefficients α_k and β_{k-1}. To understand the role of the shadow vector we investigate the connection between the Lanczos recurrence and the general three-term recurrence (5). For the simplicity of notation we assume in this chapter that

$$\dim(\mathcal{K}_n(\mathbf{A}, r_0)) = n. \tag{6}$$

The connection between the Lanczos vectors and the vectors computed by the general three-term recurrence (5) is partially explained in Greenbaum's theorem [6]: *If the three-term recurrence* (5) *(α´s and β´s can be almost anything) is run for no more than $(n+2)/2$ steps, there is a vector $\widetilde{r}_0$ such that recurrence* (5) *is the Lanczos recurrence.*

In the following theorem, we extend results of Greenbaum's paper [6].

ALGORITHM 1: *Unsymmetric Lanczos Algorithm*

input $\mathbf{A},\ r_0,\ \widetilde{r}_0$

initialization

$$
\begin{aligned}
v_0 &= \mathbf{o} \\
w_0 &= \mathbf{o} \\
\beta_0 &= 0 \\
\gamma_0 &= 0 \\
v_1 &= r_0/\|r_0\| \\
w_1 &= \widetilde{r}_0/\widetilde{r}_0^T v_1
\end{aligned}
$$

for $k = 1,\ \ldots$

$$
\begin{aligned}
\alpha_k &= w_k^T \mathbf{A} v_k \\
\mathbf{t} &= \mathbf{A}v_k - \alpha_k v_k - \beta_{k-1} v_{k-1} \\
\gamma_k &= \|\mathbf{t}\| \\
v_{k+1} &= \mathbf{t}/\gamma_k \\
\mathbf{t} &= \mathbf{A}^T w_k - \alpha_k w_k - \gamma_{k-1} w_{k-1} \\
\beta_k &= v_{k+1}^T \mathbf{t} \\
w_{k+1} &= \mathbf{t}/\beta_k
\end{aligned}
$$

end for

Theorem 1 *Given the vectors $v_1, \ldots, v_{k+1}$ computed by the general three-term recurrence* (5)*. Then it is possible to compute these vectors by the unsymmetric Lanczos algorithm* 1 *if and only if $r_0 \notin \mathcal{W}_k$ and*

$$\beta_j \neq 0, \qquad j = 1, \ldots, k-1, \tag{7}$$

where

$$\mathcal{W}_k \equiv \bigcup_{i=1}^{k} \mathcal{K}_i(\mathbf{A}, v_{i+1}). \tag{8}$$

If $k \leq n/2$ and if the coefficients β_j satisfy the condition (7) *then it is possible to compute the vectors $v_1, \ldots, v_{k+1}$ by the unsymmetric Lanczos algorithm* 1*.*

Choosing the coefficient γ_k in the recurrence (5) as

$$\gamma_k = -(\alpha_k + \beta_{k-1}), \tag{9}$$

the vectors v_{k+1} lie in the linear manifold $r_0 + \mathbf{A}\mathcal{K}_k(\mathbf{A}, r_0)$ and we can consider them to be the residual vectors. Denoting $r_k \equiv v_{k+1}$ and using the relation $r_k = b - \mathbf{A}x_k$, we can derive, from recurrence for computing r_k, the recurrence for computing the corresponding approximation x_k. Similarly, if we choose the coefficient γ_k in the unsymmetric Lanczos algorithm 1 according to (9), denote $r_k \equiv v_{k+1}$, and add the recurrence for computing the corresponding approximation x_k then we obtain the LMA algorithm – Lanczos method's algorithm. This algorithm computes the residuals and the approximations given by the conditions (4). In our work we generalize the result of the theorem 1 also for the connection between Krylov subspace method realized by three-term recurrence (for computing residuals and approximation) and the Lanczos method implemented by the LMA algorithm.

Now turn to another issue. Let r_k be arbitrary residual,

$$r_k \in r_0 + \mathbf{A}\mathcal{K}_k(\mathbf{A}, r_0). \tag{10}$$

If the shadow vector satisfies the conditions

$$\widetilde{r}_0 \perp \mathcal{K}_k(\mathbf{A}, r_k), \qquad \widetilde{r}_0^T r_0 \neq 0, \tag{11}$$

then

$$r_k \perp \mathcal{K}_k(\mathbf{A}^T, \widetilde{r}_0). \tag{12}$$

When we run the Lanczos method with the starting parameter $\widetilde{r}_0$ chosen according to (11), the k-th residual of the Lanczos method (denote it by r_k^L) fulfils the conditions (10) and (12). Since the residuals r_k^L and r_k are determined by these conditions uniquely, it holds

$$r_k = r_k^L. \tag{13}$$

Therefore, the short-term implementations (LMA, BiCG) of the Lanczos method are able to compute (if no breakdown occurs) any residual r_k from the linear manifold $r_0 + \mathbf{A}\mathcal{K}_k(\mathbf{A}, r_0)$, e.g. a residual of the GMRES method. In our thesis, we discuss the question whether it is almost-any possible to compute the given residual vector using the LMA algorithm.

We generalize the idea given by (11). Define the space

$$\mathcal{Z}_{2^l} \equiv \bigcup_{i=0}^{l} \mathcal{K}_{2^i}(\mathbf{A}, r_{2^i}), \qquad 2^l \leq n/2, \tag{14}$$

where r_{2^i} are arbitrary residuals (e.g. the residuals of the GMRES or of the FOM method) lying in the linear manifold $r_0 + \mathbf{A}\mathcal{K}_{2^i}(\mathbf{A}, r_0)$. The space $\mathcal{Z}_{2^l}$ does not contain the initial residual r_0 for almost-any residual r_{2^i} and, therefore, we can choose the shadow vector to be orthogonal to the space $\mathcal{Z}_{2^l}$ in association with $\widetilde{r}_0^T r_0 \neq 0$. If no breakdown occurs then the Lanczos method's algorithm (LMA) started with the parameter $\widetilde{r}_0$ computes the residuals $r_1^L, \ldots, r_k^L$ such that

$$r_0^L = r_0,\ r_1^L = r_1,\ r_2^L = r_2,\ r_4^L = r_4,\ r_8^L = r_8 \ldots,\ r_{2^l}^L = r_{2^l}.$$

If the long-term recurrences do not have a special shape (i.e. some coefficients are equal to zero) then the LMA algorithm can compute at most $\log_2(n)$ residuals of the given Krylov subspace method.

In the end of the chapter, we formulate one of the reasons why the convergence curves of classical Krylov subspace methods are often very close to the convergence curve of GMRES whenever the convergence curve of GMRES decreases rapidly. We use a model Krylov subspace method that chooses randomly a vector from $k+1$ dimensional unit sphere of the space $\mathcal{K}_{k+1}(\mathbf{A}, r_0)$ and determines the k-th residual as an intersection of the line given by this vector with the manifold $r_0 + \mathbf{A}\mathcal{K}_k(\mathbf{A}, r_0)$.

In numerical experiments, we demonstrate that the Lanczos metod can compute selected residuals of the FOM and GMRES methods. We discuss the question how to construct a random vector. We determine numerically the optimal shadow vector. The optimality is taken in the sense of the closest convergence curves of the GMRES and QMR methods. We observed very close convergence curves.

III. On error estimation in the CG method

Consider a symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and a right-hand side vector $b \in \mathbb{R}^n$. We investigate numerical estimation of errors in iterative methods for solving linear systems (1). In particular, we focus on the conjugate gradient method (CG) of Hestenes and Stiefel [9] and on the lower estimates of the $\mathbf{A}$-norm (also called the energy norm) of the error, which has important meaning in physics and quantum chemistry, and plays a fundamental role in evaluating convergence.

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

$$\begin{aligned} x_j &\in x_0 + \mathcal{K}_j(\mathbf{A}, r_0) \\ \|x - x_j\|_{\mathbf{A}} &= \min_{u\in x_0+\mathcal{K}_j(\mathbf{A},r_0)} \|x - u\|_{\mathbf{A}}, \end{aligned} \tag{15}$$

i.e. they minimize the $\mathbf{A}$-norm of the error

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

over all methods generating approximations in the manifold $x_0 + \mathcal{K}_j(\mathbf{A}, r_0)$. The standard implementation of the CG method was given in [9, (3:1a)-(3:1f)], see algorithm 2. The residual vectors $\{r_0, r_1, \ldots, r_{j-1}\}$ form an orthogonal basis of the j-th

Algorithm 2: *Conjugate gradient method* (CG)

input $x_0,\ \mathbf{A},\ b$

initialization

$$\begin{aligned} r_0 &= b - \mathbf{A}x_0 \\ p_0 &= r_0 \end{aligned}$$

for $j = 0, 1, \ \ldots$

$$\gamma_j = \frac{(r_j, r_j)}{(p_j, \mathbf{A}p_j)}$$

$$\begin{aligned} x_{j+1} &= x_j + \gamma_j\, p_j \\ r_{j+1} &= r_j - \gamma_j\, \mathbf{A}p_j \end{aligned}$$

$$\delta_j = \frac{(r_{j+1}, r_{j+1})}{(r_j, r_j)}$$

$$p_{j+1} = r_{j+1} + \delta_j\, p_j$$

end for

Krylov subspace $\mathcal{K}_j(\mathbf{A}, r_0)$. The orthogonality relations create the elegance of the method described in [9] and represent the fundamental property which links the CG method to the world of classical orthogonal polynomials. The j-th residual can be written as a polynomial in the matrix $\mathbf{A}$ applied to the initial error, $r_j = \varphi_j(\mathbf{A})$. These polynomials are orthogonal with respect to the discrete inner product

$$(f, g) = \sum_{i=1}^{n} \omega_i f(\lambda_i) g(\lambda_i), \qquad \omega_i = (r_0, u_i)^2 / \|r_0\|^2, \tag{16}$$

where u_i are normalized eigenvectors and λ_i are the eigenvalues of matrix $\mathbf{A}$. The eigenvalues λ_i and weights ω_i determine distribution function $\omega(\lambda)$ and the corresponding Riemann-Stieltjes integral

$$\int_{\zeta}^{\xi} f(\lambda)\, d\omega(\lambda) \equiv \sum_{i=1}^{n} \omega_i f(\lambda_i)\,. \tag{17}$$

Consequently, in j-th step, CG implicitly determines weights $\omega_i^{(j)}$ and nodes $\theta_i^{(j)}$ of j-th Gauss quadrature approximation of the integral (17)

$$\int_{\zeta}^{\xi} f(\lambda)\, d\omega(\lambda) = \sum_{i=1}^{j} \omega_i^{(j)} f(\theta_i^{(j)}) + R_j(f) \tag{18}$$

where $R_j(f)$ stands for the (truncation) error in the Gauss quadrature. In [3] it was proved that for $f(\lambda) = \lambda^{-1}$, the identity (18) can be written in the form

$$\frac{\|x - x_0\|_{\mathbf{A}}^2}{\|r_0\|^2} \;=\; C_j \;+\; \frac{\|x - x_j\|_{\mathbf{A}}^2}{\|r_0\|^2} \tag{19}$$

and the value C_j of j-point Gauss quadrature was approximated from the actual Gauss quadrature calculations (or from the related recurrence relations).

In our work we discuss several mathematically equivalent identities to (19). An interesting form of (19) (multiplied by $\|r_0\|^2$) was noticed by Warnick

$$\|x - x_0\|_{\mathbf{A}}^2 \;=\; r_0^T(x_j - x_0) + \|x - x_j\|_{\mathbf{A}}^2\,. \tag{20}$$

We derived a mathematically equivalent identity by simple algebraic manipulations without using Gauss quadrature,

$$\|x - x_0\|_{\mathbf{A}}^2 \;=\; r_j^T(x_j - x_0) + r_0^T(x_j - x_0) + \|x - x_j\|_{\mathbf{A}}^2. \tag{21}$$

The right-hand side of (21) contains, in comparison with (20), additional term $r_j^T(x_j - x_0)$. This term is in exact arithmetic equal to zero, but it has an important correction effect in finite precision computations. Consequently, we found that the simplest identity mathematically equivalent to (19) (multiplied by $\|r_0\|^2$) was present in the Hestenes and Stiefel paper [9, relation (6:2)],

$$\|x - x_0\|_{\mathbf{A}}^2 \;=\; \sum_{i=0}^{j-1} \gamma_i \|r_i\|^2 + \|x - x_j\|_{\mathbf{A}}^2. \tag{22}$$

The numbers $\gamma_i\|r_i\|^2$ are trivially computable; both γ_i and $\|r_i\|^2$ are available at every iteration step.

Using $\|x - x_0\|_{\mathbf{A}}^2 = \|r_0\|^2 C_n$, (19) is written in the form

$$\|x - x_j\|_{\mathbf{A}}^2 \;=\; \|r_0\|^2\,[C_n - C_j]\;.$$

As suggested in [3, pp. 28–29], the unknown value C_n can be replaced, at a price of $m - j$ extra steps, by a computable value C_m for some $m > j$. The paper [3],

however, did not properly use this idea and did not give a proper formula for computing the difference $C_m - C_j$ without cancellation, which limited the applicability of the proposed result. Golub and Meurant cleverly resolved this trouble in [2] and proposed an algorithm for estimating the **A**-norm of the error in the CG method called CGQL.

Consider, in general, (18) for j and $j+d$, where d is some positive integer. The idea is simply to eliminate the unknown integral by subtracting the identities for j and $j+d$. In particular, using (19)–(22) we obtain the mathematically equivalent identities

$$\|x - x_j\|_{\mathbf{A}}^2 = \eta_{j,d} + \|x - x_{j+d}\|_{\mathbf{A}}^2 , \tag{23}$$
$$\eta_{j,d} \equiv \|r_0\|^2 \, [(C_{j+d} - C_j],$$
$$\|x - x_j\|_{\mathbf{A}}^2 = \mu_{j,d} + \|x - x_{j+d}\|_{\mathbf{A}}^2 , \tag{24}$$
$$\mu_{j,d} \equiv r_0^T (x_{j+d} - x_j),$$
$$\|x - x_j\|_{\mathbf{A}}^2 = \vartheta_{j,d} + \|x - x_{j+d}\|_{\mathbf{A}}^2, \tag{25}$$
$$\vartheta_{j,d} \equiv r_0^T (x_{j+d} - x_j) - r_j^T (x_j - x_0) + r_{j+d}^T (x_{j+d} - x_0),$$
$$\|x - x_j\|_{\mathbf{A}}^2 = \nu_{j,d} + \|x - x_{j+d}\|_{\mathbf{A}}^2, \tag{26}$$
$$\nu_{j,d} \equiv \sum_{i=j}^{j+d-1} \gamma_i \|r_i\|^2.$$

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

$$\|x - x_j\|_{\mathbf{A}}^2 \gg \|x - x_{j+d}\|_{\mathbf{A}}^2 , \tag{27}$$

then neglecting $\|x - x_{j+d}\|_{\mathbf{A}}^2$ on the right-hand sides of (23)–(26) gives lower bounds (mathematically equal) for the squared **A**-norm of the error in the j-th step. Under the assumption (27) these bounds are reasonably tight (their inaccuracy is given by $\|x - x_{j+d}\|_{\mathbf{A}}^2$).

Mathematically (in exact arithmetic)

$$\eta_{j,d} = \mu_{j,d} = \vartheta_{j,d} = \nu_{j,d} . \tag{28}$$

In finite precision computations (28) does not hold in general, and the different bounds may give substantially different results.

IV. CG in finite precision arithmetic

For a long time the effects of rounding errors to the Lanczos and CG methods seemed devastating. Orthogonality among the computed vectors $v_1, v_2, \ldots$ was usually lost very quickly, with a subsequent loss of linear independence. Consequently, the finite termination property was lost. Still, despite a total loss of orthogonality among the vectors in the Lanczos sequence $v_1, v_2, \ldots$, and despite a possible regular appearance of Lanczos vectors which were linearly dependent, the Lanczos and the CG methods produced reasonable results. In the fourth chapter we explain why.

We discuss a model of finite precision CG based on [7] and [4]. Ideally (in exact precision arithmetic) convergence of CG is determined by Gauss quadrature for the Riemann-Stieltjes integral

$$\int_{\zeta}^{\xi} \lambda^{-1}\, d\omega(\lambda).$$

Finite precision CG can be viewed as ideal (exact precision) CG applied to a modified problem, for which the convergence is determined by the Riemann-Stieltjes integral

$$\int_{\zeta}^{\xi} \lambda^{-1}\, d\hat{\omega}(\lambda)$$

with a distribution function $\hat{\omega}(\lambda)$ obtained from $\omega(\lambda)$ by *blurring* the individual points λ_i into (infinitely) many points of increase close to each λ_i, and the total size of increase in the neighbourhood of λ_i equal to ω_i. We explain that delay of convergence (due to rounding errors) in finite precision CG computations is determined by the difference between the iteration number and the numerical rank of the matrix of computed Lanczos vectors.

In the rest of the fourth chapter we describe and bound rounding errors arising in the finite precision CG computations. We prove a new theorem about the local orthogonality between the direction vector p_j and the iteratively computed residual r_{j+1}. The result is used later in rounding error analysis of estimates.

Theorem 2 *The local orthogonality between the direction vectors and the iteratively computed residuals is in the finite precision* CG *bounded by*

$$|p_j^T r_{j+1}| \leq \varepsilon \, \|r_j\|^2 \kappa(\mathbf{A})^{1/2}\, O(j\,n + j^2/2) + O(\varepsilon^2) \tag{29}$$

where ε denotes machine epsilon.

In numerical experiments, we use multiple arithmetic [1] to demonstrate that the bounds of rounding errors are in most cases overestimated. We also depict the actual size of the local orthogonality term $|p_j^T r_{j+1}|$.

V. Estimates in finite precision arithmetic

The bounds $\eta_{j,d}$, $\mu_{j,d}$, $\vartheta_{j,d}$ and $\nu_{j,d}$ are mathematically equivalent. We prove that the ideal (exact precision) identities (23)–(26) change numerically to

$$\|x - x_j\|_{\mathbf{A}}^2 = \Delta_{j,d} + \|x - x_{j+d}\|_{\mathbf{A}}^2 + E_{j,d} \tag{30}$$

where $\Delta_{j,d}$ stands for the bounds $\eta_{j,d}$, $\mu_{j,d}$, $\vartheta_{j,d}$ and $\nu_{j,d}$, and $E_{j,d}$ stands for the rounding error due to finite precision arithmetic.

Please notice that the difference between (23)–(26) and (30) *is not trivial*. The ideal and numerical counterparts of each individual term in these identities may be orders of magnitude different (see Fig. 1)! In finite precision arithmetic, rounding errors in the whole computation, not only in the computation of the convergence bounds, must be taken into account.

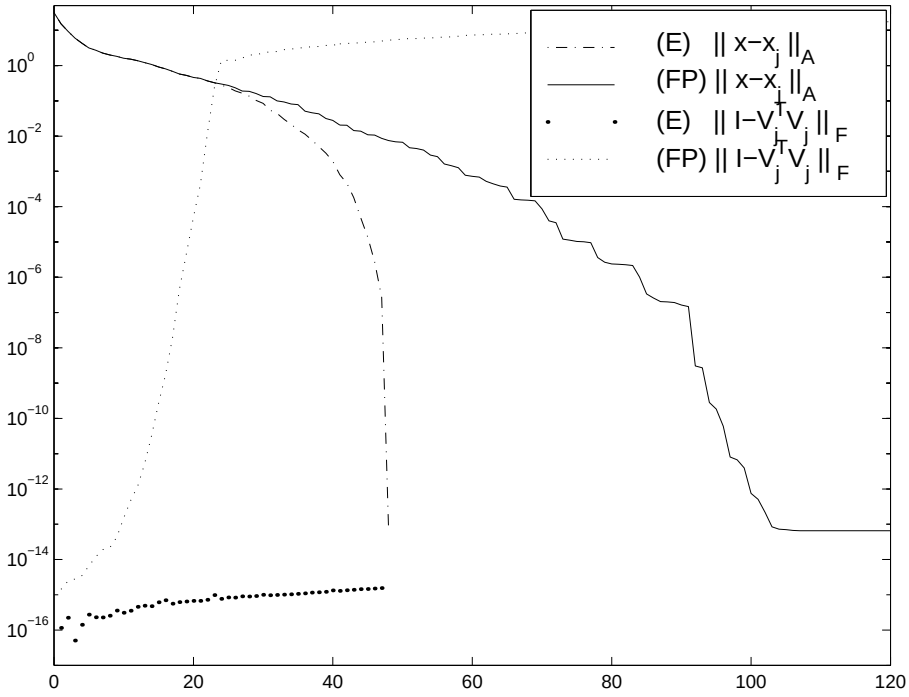

Figure 1: The **A**-norm of the error for the CG implementation with the double reorthogonalized residuals (dashed-dotted line) is compared to the **A**-norm of the error of the ordinary finite precision CG implementation (solid line). The corresponding loss of orthogonality among the normalized residuals is plotted by the dots resp. the dotted line.

Justification of the bound $\eta_{j,d}$ in finite precision arithmetic was given in [3]. This justification is applicable only until $\|x - x_j\|_{\mathbf{A}}$ reaches the square root of the machine precision. In our work we prove much stronger results than the analysis of the finite precision counterpart of (23) given in [3].

We concentrate on lower estimates $\nu_{j,d}$ and $\vartheta_{j,d}$. In our rounding error analysis we show that $E_{j,d}$ is related to $\varepsilon \|x - x_j\|_{\mathbf{A}}$. In more detail, considering the estimate $\nu_{j,d}$, the significant part of perturbation term $E_{j,d}$ can be written in the form

$$2\,\varepsilon\,\{(x - x_{j+d})^T f_{j+d} - (x - x_j)^T f_j\}$$

where $\varepsilon\, f_j \equiv r_j - (b - \mathbf{A}x_j)$ and this nontrivial fact (in [12] we use weaker result) is also demonstrated numerically by using multiple arithmetic [1].

Due to the fact that rounding errors in computing $\nu_{j,d}$ and $\vartheta_{j,d}$ numerically are negligible, the numerically computed values $\nu_{j,d}$ and $\vartheta_{j,d}$ (the difference $x_{j+d} - x_j$ must be computed in a proper way) give a good estimate for the **A**-norm of the error $\|x - x_j\|^2_{\mathbf{A}}$ until the perturbation term $E_{j,d}$ in (30) is reasonably smaller than the square of the computed **A**-norm of the error, i.e. until

$$|E_{j,d}| \ll \|x - x_j\|^2_{\mathbf{A}}. \tag{31}$$

Since $E_{j,d}$ is related to $\varepsilon \|x - x_j\|_{\mathbf{A}}$, we can define $F_{j,d}$ by

$$E_{j,d} \equiv \varepsilon \|x - x_j\|_{\mathbf{A}} F_{j,d}.$$

The relation (31) is then equivalent to

$$\|x - x_j\|_{\mathbf{A}} \gg \varepsilon |F_{j,d}|. \tag{32}$$

The value $F_{j,d}$ represents various terms. Its upper bound is, apart from $\kappa(\mathbf{A})^{1/2}$, which comes into play as an effect of the worst-case rounding error analysis, linearly dependent on an upper bound for $\|x - x_0\|_{\mathbf{A}}$. The value of $F_{j,d}$ is (as similar terms or constants in any other rounding error analysis) not important. What is important is the following possible interpretation of (32): until $\|x - x_j\|_{\mathbf{A}}$ reaches a level close to $\varepsilon \|x - x_0\|_{\mathbf{A}}$, the computed estimates $\nu_{j,d}$ and $\vartheta_{j,d}$ must work. When significant loss of orthogonality occurs, the estimate $\mu_{j,d}$ does not work.

The lower estimate of euclidean norm of the error we get from the identity presented in [9, Theorem 6:3]

$$\|x - x_j\|^2 = \sum_{i=j}^{j+d-1} \frac{\|p_i\|^2}{(p_i, Ap_i)} \left(\|x - x_i\|_A^2 + \|x - x_{i+1}\|_A^2\right) + \|x - x_{j+d}\|^2, \tag{33}$$

and replacing the unknown squares of the **A**-norms of the errors by their lower estimates. Our results are illustrated by numerical experiments.

Conclusions

- We extended results of Anne Greenbaum (Theorem 1) on the role of the shadow vector in the Lanczos process. We showed that there is a shadow vector $\widetilde{r}_0$ such that the Lanczos method computes selected residuals of another Krylov subspace method.
- In the CG method, the lower bound for the **A**-norm of the error based on Gauss quadrature is mathematically equivalent to the original formula of Hestenes and Stiefel [9].
- The local orthogonality between the direction vector and the iteratively computed residual is in the finite precision CG bounded by (29) (Theorem 2).
- The estimate for the **A**-norm of the error $\nu_{j,d}^{1/2}$ is simple and numerically stable. Until $\|x - x_j\|_{\mathbf{A}}$ reaches its ultimate attainable accuracy level, the computed estimate $\nu_{j,d}^{1/2}$ must work. Based on the results presented in our work we believe that this estimate should be incorporated into any software realization of the CG method. There is a small reason for using the other bounds $\eta_{j,d}^{1/2}$ or $\vartheta_{j,d}^{1/2}$ in practical computations.

References

[1] D. H. Bailey, *MPFUN: A multiple precision floating point computation package (Fotran–77), http://www.netlib.org/mpfun/*, NASA Ames Research Center, USA, March 1995.

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

[3] G. H. Golub and Z. Strakoš, *Estimates in quadratic formulas*, Numerical Algorithms, 8 (1994), pp. 241–268.

[4] A. Greenbaum, *Behavior of slightly perturbed Lanczos and Conjugate Gradient recurrences*, Lin. Alg. Appl., 113 (1989), pp. 7–63.

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

[6] ——, *On the role of the left starting vector in the two-sided Lanczos algorithm and nonsymetric linear system solvers*, in Proceedings of the Dundee meeting in Numerical Analysis, D. Griffiths, D. Highham, and G. Watson, eds., Pitman Research Notes in Mathematics Series 380, Longman, 1997.

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

[8] M. H. Gutknecht, *Lanczos-type solvers for nonsymmetric linear system of equations*, Technical Report TR–97–04, Swiss Center for Scientific Computing ETH-Zentrum, Switzerland, 1997.

[9] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bureau Standarts, 49 (1952), pp. 409–435.

[10] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Company, 1996.

[11] Z. Strakoš, *Theory of Convergence and Effects of Finite Precision Arithmetic in Krylov Subspace Methods*, Thesis for the degree Doctor of Science, Institute of Computer Science, February 2001.

[12] Z. Strakoš and P. Tichý, *On error estimation in the Conjugate Gradient method and why it works in finite precision computations*, Submitted to ETNA, (June 2001, released March 2002), p. 26.

List of Publications

Master thesis:

[1] P. Tichý, *Chování BiCG a CGS algoritmů*, diplomová práce, Katedra numerické matematiky, MFF UK, Praha, 1997.

Original published papers:

[2] P. Tichý and J. Zítko, *Derivation of BiCG from the conditions defining Lanczos' method for solving a system of linear equations*, Application of Mathematics 5, 43 (1998), pp. 381–388.

[3] P. Tichý, *BiCGStab and other hybrid BiCG methods*, in WDS'98 Proceedings of Contributed Papers, J. Šafránková, ed., matfyzpress, 1998, pp. 52–58.

[4] P. Tichý, *The shadow vector in the Lanczos method*, in Proceedings of the XUIIth summer school software and algorithms of numerical mathematics Nečtiny, 1999, pp. 309–320.

[5] P. Tichý, *Vztah Lanczosovy metody k ostatním krylovovským metodám*, Doktorandský den 00. Sborník příspěvků., F. Hakl, ed., Ústav informatiky AV ČR, Praha, 2000, pp. 58–63.

[6] P. Tichý, *O odhadu A-normy chyby v metodě sdružených gradientů*, Doktorandský den 01. Sborník příspěvků., F. Hakl, ed., Ústav informatiky AV ČR, Praha, 2001, pp. 4–9.

Original paper submitted for publication:

[7] Z. Strakoš and P. Tichý, *On error estimation in the conjugate gradient method and why it works in finite precision computations*, Submitted to ETNA, (June 2001, released March 2002), p. 26.