Iterative Solvers: Where Have We Been and Where Do We Go to

Rüdiger Weiss *

Abstract

When the conjugate gradient method has been generalized for the solution of nonsymmetric and indefinite linear systems a confusing variety of different methods has been created, each of them suited for a different problem. We overview these techniques and classify them. From many practical applications we come to the conclusion that only two methods are sufficient to solve most problems: BiCGSTAB(2) and CGNE. The appropriate method of these can be selected by adaptive algorithms. Although there are many theoretical results for most of the methods, a deeper analysis of BiCGSTAB(2) is still missing. We stress that convergence can be accelerated essentially only by preconditioning.

Keywords: linear system, convergence, iterative solution

1 Introduction

Carl Friedrich Gauss states in a letter dated December 26, 1823 to the factory owner Gerling: "Ich empfehle Ihnen diesen Modus zur Nachahmung. Schwerlich werden Sie je wieder direct eliminieren, wenigstens nicht, wenn Sie mehr als 2 Unbekannte haben. Das indirecte Verfahren lässt sich halb im Schlafe ausführen, oder man kann während desselben an andere Dinge denken" ("I recommend to you to follow this mode. You will hardly ever eliminate directly again, at least not if you have more than 2 unknowns. The indirect method can be performed while being half asleep, or you can have something else in mind at the same time.") Gauss is concerned with the solution of linear systems arising from land survey for which he developed an iterative method.

Today with the use of computers linear systems with a dimension in the millions or billions can be solved. If you want to analyze and solve technical problems, you can conduct experiments in reality or in a computer. In the automotive industry, computer simulations have replaced crash tests for a long time already, as it is cheaper to wreck a car simulated in a computer than a real one. The deformations of the car body striking a wall are calculated by the solution of partial differential equations.

A human birth is a quite different thing. However, the head of the child is compressed during the birth by the pelvis of the mother, and thus a birth can be simulated by the same computer program as the car crash. In order to avoid irreversible brain damages during the birth this technique is applied at the Augsburg university hospital in Germany. The shape of the pelvis and the childs head have been recorded by computer tomography before.

*Computing Center, Universität Karlsruhe, 76128 Karlsruhe, Germany (weiss@rz.uni-karlsruhe.de)

Numerical simulation plays an ever increasing part in industry and science. Many technical and scientifical problems are described by systems of differential, integral and eigenvalue equations. After discretization and linearization linear systems of equations have to be solved. These systems are usually sparse and large. But the extensive use of numerical methods is not restricted to technical problems. The entertainment industry has become one of the biggest employers for mathematicians and computer scientists. Movies like "Titanic" or "Star Wars" use extensively computer animations. In sports simulations are used to optimize the equipment, for example the yacht for the America's cup. We also do not keep secret that many linear systems come from military applications.

In order to achieve high accuracy or a realistic animation the size of the systems is large. Therefore, fast solvers are needed. However today, applying iterative methods is not as easy that it can be performed while being half asleep. There are many different techniques, each of them suited for particular problems. The user has to care about the correct methods before he can start the calculation.

2 Preliminaries

Our problem is to solve the linear system

$$Ax = b. \tag{1}$$

The matrix $A \in \mathbb{R}^{n\times n}$ is real and square of dimension n and $x,\ b \in \mathbb{R}^n$. Though many of the presented methods work also for singular matrices we assume that A is not singular in order to get a unique solution and unique error estimates.

Starting from an initial guess x_0 an iterative solution method constructs a sequence of approximations x_k. The residuals are defined by

$$r_k = Ax_k - b\ , \tag{2}$$

and the errors by

$$e_k = x_k - x\ . \tag{3}$$

The relative residuals and errors are connected by

$$\frac{1}{\kappa}\frac{||r_k||}{||r_0||} \le \frac{||e_k||}{||e_0||} \le \kappa\frac{||r_k||}{||r_0||} \le \kappa^2\frac{||e_k||}{||e_0||}\ , \tag{4}$$

where $\kappa = ||A||\ ||A^{-1}||$ is the condition number of A. Iterative methods are usually controlled by the observation of the norm of the residual. In reality the error should be controlled, but the error is not available. If κ is near to unity the norm of the relative residuals is strongly connected to the norm of the relative errors, i. e. if the residuals decrease sufficiently the errors will be reduced as well. For a large condition number the residuals can decrease while the norm of the errors does not decrease or even increases and vice-versa. It can be dangerous to apply a method that minimizes the residuals.

The norm of the residuals of a generalized conjugate gradient method may heavily oscillate. This fact is known from practice. Therefore, the smoothing algorithm was introduced by Schönauer [13] in order to get a norm non-increasing function of the index.

Let x_k be the iterate and r_k the residual resulting from an iterative method. The sequence

$$s_0 = r_0 \ , z_0 = x_0 \ , \tag{5}$$
$$s_k = s_{k-1} + \gamma_k (r_k - s_{k-1}) \ , \tag{6}$$
$$z_k = z_{k-1} + \gamma_k (x_k - z_{k-1}) \tag{7}$$

is a corresponding smoothed method.

The smoothed sequence is a mean of the actual element of the original sequence and the previous element of the smoothed sequence. The smoothed sequence does not react on the original sequence. The different smoothing techniques vary by the determination of the coefficient γ_k.

The original approach of Schönauer is to minimize the residual in the Euclidean norm

$$||s_k|| = min_{\gamma_k} ||s_{k-1} + \gamma_k (r_k - s_{k-1})||. \tag{8}$$

This technique is called minimal residual (MR) smoothing. A direct result of (8) is that the smoothed residual is a function of the iteration index, with a non-increasing norm:

$$\|s_k\| \leq \|s_{k-1}\|, \tag{9}$$

$$\|s_k\| \leq \|r_k\|. \tag{10}$$

The concept of quasi minimal residual (QMR) smoothing was introduced by Freund and Nachtigal [5] for the QMR method. Zhou and Walker [24] generalized the technique and created the name of this kind of smoothing. The idea is to choose γ_k so that

$$s_k = \frac{1}{\sum_{i=0}^{k} \frac{1}{||r_i||^2}} \sum_{i=0}^{k} \frac{1}{||r_i||^2} r_i \tag{11}$$

is satisfied. This choice establishes a weighted mean of the original residuals. If $\|r_i\|$ is large, then the contribution to $\|s_k\|$ is neglectable because $\frac{1}{||r_i||^2}$ is small. Quasi-residual smoothing and residual-minimizing smoothing coincide if the r_k are orthogonal.

Smoothing has some remarkable properties. MR smoothing tranforms classical CG into ORTHOMIN and ORTHORES into GMRES and QMR smoothing transforms BCG into QMR and an auxiliary sequence resulting from CGS into TFQMR, see [20].

Smoothing can be considered from a theoretical and a practical viewpoint. The results of the theory show that methods can be transformed by a very simple process. From a practical viewpoint smoothing should be applied to basic methods with oscillating residual norms in order to obtain in each iteration step a reliable approximation, all the more because the additional work can be neglected. In Figure 1 the basic method BCG is compared with the MR smoothed BICO and the QMR smoothed QMR. Peaks of BCG are eliminated by both techniques and BICO and QMR behave very similar. The number of matrix-vector multiplications is a measure for the number of iterations. Note that the scale is logarithmic for the relative residuals and errors so that the non-smoothed BCG method oscillates in orders of magnitude. There are recent investigations on the stability of different implementations by Gutknecht and Rozložník [8].

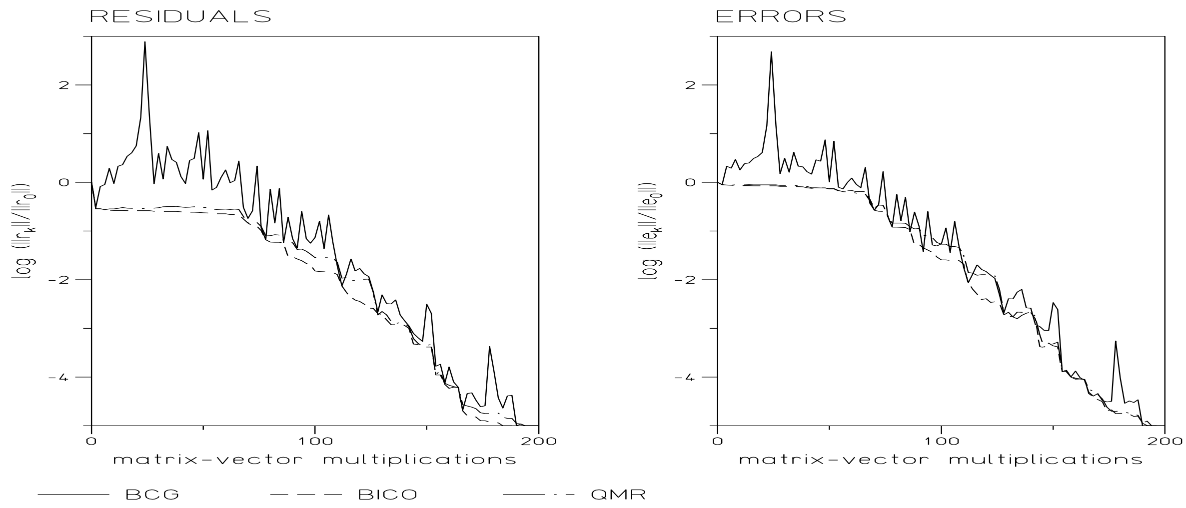

Figure 1: Comparison of different smoothers for a rough model of the Navier-Stokes equations with Reynolds number 500 and BCG as basic method

3 The Symmetric Case

The conjugate gradient (CG) method was developed by Hestenes and Stiefel [9] in the early 1950s. The method can be considered as a direct method because the exact solution is obtained at least in the step n in the absence of roundoff errors. Until the early 1970s the method was not widely used. With the increasing use of computers and the consideration as an iterative method the method recovered importance.

The idea of the classical conjugate gradient method is to chose

$$x_k = x_{k-1} + \sum_{i=1}^{k} \delta_{i,k} r_{i-1} \tag{12}$$

and to minimize the quadratic form

$$\|e_k\|_A^2 = \min_{\delta_{1,k},\ldots,\delta_{k,k}} ||e_{k-1} + \sum_{i=1}^{k} \delta_{i,k} r_{i-1}||_A^2, \tag{13}$$

if A is symmetric and positive definite. In this case the method converges as well.

Equation (12) suggests that the length of the recurrence increases with the iteration but the new iterate of classical CG can be calculated by a short recurrence.

The idea of ORTHOMIN by Vinsome [18] is to choose x_k as in (12), but to determine the $\delta_{i,k}$ from

$$\|r_k\| \quad = \quad \min_{\delta_{1,k},\ldots,\delta_{k,k}} \|r_{k-1} + \sum_{i=1}^{k} \delta_{i,k} A r_{i-1}\|. \tag{14}$$

The method can be obtained from classical CG by MR smoothing.

The idea of Fridman's algorithm [6] is to choose x_k a little bit different than in (12)

$$x_k = x_{k-1} + \sum_{i=1}^{k} \delta_{i,k} A r_{i-1}, \tag{15}$$

property	method
energy norm-minimizing	CG
residual-minimizing	ORTHOMIN
error-minimizing	Fridman's method

Table 1: Comparison of minimization properties in the symmetric case

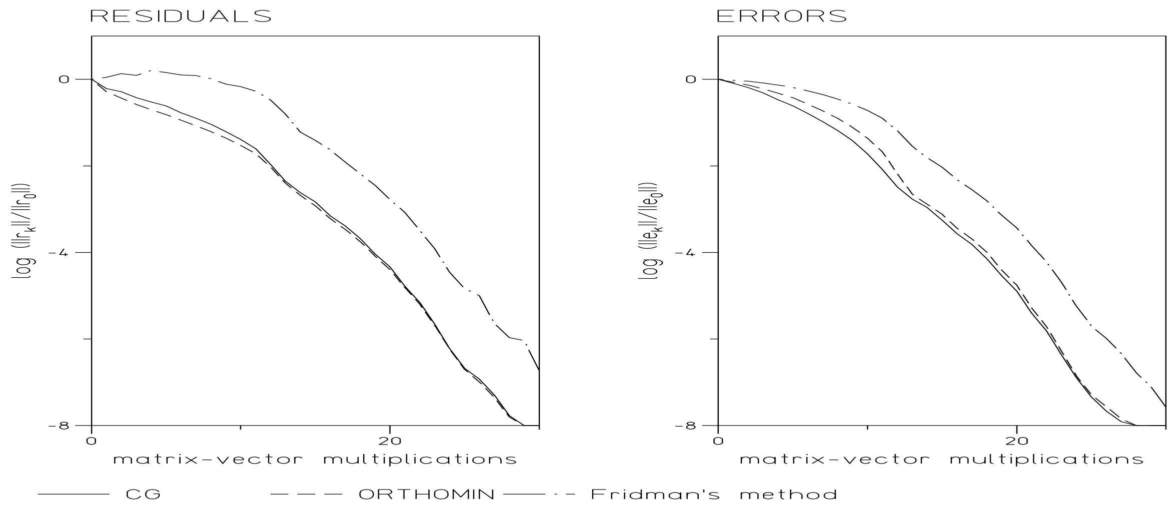

Figure 2: CG, ORTHOMIN and Fridman's method for the sample problem

and to minimize the error in the Euclidean norm

$$||e_k|| \quad = \quad \min_{\delta_{1,k},\ldots,\delta_{k,k}} \|e_{k-1} + \sum_{i=1}^{k} \delta_{i,k} A r_{i-1}\|. \tag{16}$$

The three methods for symmetric matrices are closely related. It can be shown that the iterates of CG and ORTHOMIN can be represented by

$$x_k = x_0 + \sum_{i=0}^{k-1} \beta_{i,k} A^i r_0, \tag{17}$$

while those of Fridman's method fulfil

$$x_k = x_0 + \sum_{i=0}^{k-1} \beta_{i,k} A^{i+1} r_0. \tag{18}$$

The space that is spanned by the vectors $r_0, \ldots, A^i r_0$ is called Krylov space generated by r_0 and A. Therefore these methods are also referred as Krylov subspace methods. The minimization properties are listed in Table 1. All three methods can be implemented with a short recurrence and they deliver similar results.

In Figure 2 we consider the reduction of the residuals and errors for the sample problem

$$\Delta v = h \tag{19}$$

with Dirichlet boundary conditions on the unit cube, where $v = (v_1, v_2, v_3)^T$, $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$. The linear system arises from a finite difference discretization with consistency order 2 generated by the FIDISOL program package [14] on a $20 \times 20 \times 20$ grid. The dimension of the system is $n = 24000$. The convergence of CG, ORTHOMIN and Fridman's method is comparable, see Figure 2. Note that the number of matrix-vector multiplications corresponds to the number of iterations. It should be stressed that a relative reduction of the residuals and errors by 8 orders of magnitude is achieved with 30 i. e. in $1.25 \cdot 10^{-3} \cdot n$ matrix-vector multiplications. Thus, the iterative methods are by a factor of 10^3 faster than direct methods.

4 The Nonsymmetric Case

Starting from the 1970s the classical CG method has been generalized in order to obtain convergent techniques for non-symmetric and non-positive definite systems. A whole family of methods results from this generalization; see e. g. [20].

A direct generalization of classical CG for nonsymmetric, but definite matrices is to minimize the energy norm of the error in the following sense:

$$\|e_k\|_A \leq \sqrt{1 + \frac{\rho^2}{\mu^2}} \min_{\beta_1 \ ,\ldots,\beta_k} \|\sum_{i=1}^{k} \beta_i A^i e_0 + e_0\|_A, \tag{20}$$

where ρ is the spectral radius of the skewsymmetric part of A and μ is the smallest eigenvalue of the symmetric part of A. If A is symmetric and positive definite, then we obtain the classical CG method.

GMRES (Generalized Minimal RESidual method) was proposed by Saad and Schultz [12] and generalizes ORTHOMIN. The residuals are minimized in the Euclidean norm

$$||r_k|| \quad = \min_{\beta_1 \ ,\ldots,\beta_k} ||\sum_{i=1}^{k} \beta_i A^i r_0 + r_0||. \tag{21}$$

GMERR [19] generalizes Fridman's method and minimizes the error

$$||e_k|| \quad = \min_{\beta_1,\ldots,\beta_k} ||\sum_{i=1}^{k} \beta_i \left(A^T\right)^i r_0 + e_0||. \tag{22}$$

Ehrig and Deuflhard [3] show that GMERR converges for all right-hand sides only for normal matrices. The convergence is analyzed in this case in [11].

For all these three methods the storage and computing requirements increase with the iteration step. In general truncated and restarted versions have to be applied. Then the methods become parameter dependent. Restarted and truncated GMRES and ORTHORES methods give similar results, while restarted or truncated GMERR is different. There are examples where GMRES works fine while GMERR converges very slowly. There are other examples where GMERR is very fast but GMRES is very bad.

5 Biconjugate Gradient-Based Methods

The biconjugate gradients (BCG) were introduced by Lanczos [10]. The method can be considered as method generated by the Lanczos algorithm. BCG can also be considered as method applied to the double system

$$\begin{pmatrix} A & 0 \\ 0 & A^T \end{pmatrix} \begin{pmatrix} x \\ x^* \end{pmatrix} = \begin{pmatrix} b \\ b^* \end{pmatrix}, \tag{23}$$

where b^* is arbitrary. The iterates are calculated by a double CG ansatz (12)

$$x_k = x_{k-1} + \sum_{i=1}^{k} \delta_{i,k} r_{i-1}, \tag{24}$$

$$x_k^* = x_{k-1}^* + \sum_{i=1}^{k} \delta_{i,k} r_{i-1}^*, \tag{25}$$

where the residuals have the form

$$r_k = Ax_k - b \tag{26}$$

$$r_k^* = A^T x_k^* - b^*. \tag{27}$$

r_k^* is called shadow residual. The coefficients are obtained from the biorthogonalities

$$r_k^T r_{k-i}^* = 0, \tag{28}$$

$$[r_k^*]^T r_{k-i} = 0, \tag{29}$$

for $i = 1, \ldots, k$. These orthogonalities can be considered as weak formulation for the condition that the residual and the shadow residual vanish for the true solution. The residuals can be represented with a matrix polynomial

$$\Pi_k(A) = \sum_{i=1}^{k} \gamma_i A^i + I \tag{30}$$

so that

$$r_k = \Pi_k(A) r_0. \tag{31}$$

The method may break down if a division by zero occurs. In this case look-ahead strategies, see [2, 4, 22] can avoid breakdowns.

It is known from practice that the convergence behavior of BCG in the Euclidean norm can be very erratic. In order to obtain a better behavior Schönauer applied minimal residual smoothing to BCG. The quasi-minimal residual method (QMR) was developed by Freund and Nachtigal [5]. Zhou and Walker [24] show that quasi-residual smoothing transforms BCG into QMR.

BiCGSTAB (BiConjugate Gradients STABilized) was introduced by van der Vorst [17]. The idea is to substitute the squared matrix polynomial Π_k of BCG, see (30), by the product of a matrix polynomial Ψ_k by Π_k in order to get a faster and smoother convergencei:

$$r_k = \Psi_k(A)\Pi_k(A)r_0. \tag{32}$$

Ψ_k is recursively defined by

$$\Psi_k(A) = (\gamma_k A + I)\Psi_{k-1}(A), \tag{33}$$

where γ_k is determined from a 1-dimensional minimization

$$\|r_k\| = \min_{\gamma_k} \|(\gamma_k A + I)\Psi_{k-1}(A)\Pi_k(A)r_0\|. \tag{34}$$

Gutknecht [7], Sleijpen and Fokkema [15] and Sleijpen, van der Vorst and Fokkema [16] generalized BiCGSTAB so that Ψ_k is determined by an ℓ-dimensional minimization. These methods are called BiCGSTAB(ℓ). Other generalizations called product methods were given by Brezinski and Redivo-Zaglia [1] and generalized product-type methods by Zhang [23]. While for all other methods many theoretical results have been achieved there are still many open questions for this type of methods.

The big advantage of all these methods is that the iterates can be calculated by a short recurrence. BiCGSTAB(2) is among the best biconjugate gradient-based methods. It also beats in many cases the truncated and restarted techniques of the last section.

6 Normal Equations

Let us consider the normal equations

$$AA^T y = b, \tag{35}$$

$$A^T A x = A^T b. \tag{36}$$

The solution of (36) is the solution of (1) and we get the solution of the original system x from (35) by $x = A^T y$. The matrices AA^T and $A^T A$ are symmetric and positive definite and, therefore, classical CG can be applied to the normal equations.

The speed of convergence may be very slow if the eigenvalues of the iteration matrix AA^T are more scattered than those of the original system matrix A. The condition number is squared, i. e. the convergence is slow for matrices close to symmetric matrices, but in the general case this is not true. Theoretically the methods converge always. Thus, they are very robust but not very efficient. They can be implemented by a short recurrence.

If we apply classical CG to the normal equations (35), then the errors of the original system

$$||e_k|| = \min_{\beta_1, \ldots, \beta_k} ||\sum_{i=1}^{k} \beta_i \left(A^T A\right)^i e_0 + e_0|| \tag{37}$$

are minimized. Therefore the method is called CGNE (CG Normal equations Error-minimizing). If we apply classical CG to the normal equations (36), then the residuals of the original system

$$||r_k|| = \min_{\beta_1, \ldots, \beta_k} ||\sum_{i=1}^{k} \beta_i \left(AA^T\right)^i r_0 + r_0||. \tag{38}$$

are minimized and the method is called CGNR (CG Normal equations Residual minimizing). We prefer CGNE because the errors are minimized.

7 Overview

The essential operations for the here discussed methods are matrix-vector multiplications, dot products and triadic operations. For realistic problems the work for the matrix-vector multiplication dominates the work for other operations. The matrix-vector multiplication and triadic operations are optimally vectorizable and parallelizable, thus the methods are optimally suited for today's supercomputers.

For symmetric systems the three variants CG, ORTHOMIN and Fridman's method seem to be comparable. For nonsymmetric systems with definite matrices BiCGSTAB(2) is among the best methods. The theoretical investigations for this method are not yet satisfying. For indefinite systems CGNE is the emergency exit.

The polyalgorithms proposed in [21] monitor the norm of the residuals. One variant for nonsymmetric systems switches from the efficient but less robust method BiCGSTAB to the more robust but less efficient method CGNE if convergence slows down. Coming back to the introduction, the polyalgorithms can be performed by the user while being half asleep. However, if there is no or slow convergence, then preconditioning has to be applied and that is a more difficult task.

References

[1] C. Brezinski and M. Redivo-Zaglia. Look-ahead in Bi-CGSTAB and other product methods for linear systems. *BIT*, 35:169–201, 1995.

[2] C. Brezinski, M. Redivo-Zaglia, and H. Sadok. A breakdown-free Lanczos type algorithm for solving linear systems. *Numer. Math.*, 63:29–38, 1992.

[3] R. Ehrig and P. Deuflhard. GMERR – an error minimizing variant of GMRES. Preprint SC 97-63, Konrad-Zuse-Zentrum für Informationstechnik Berlin, Takustr. 7, 14195 Berlin, Germany, 1997.

[4] R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal. An implementation of the Look-Ahead Lanczos algorithm for non-Hermitian matrices. *SIAM J. Sci. Comput.*, 14(1):137–158, 1993.

[5] R. W. Freund and N. M. Nachtigal. QMR: a quasi-minimal residual method for non-Hermitian linear systems. *Numer. Math.*, 60:315–339, 1991.

[6] V. M. Fridman. The method of minimum iterations with minimum errors for a system of linear algebraic equations with a symmetrical matrix. *USSR Comput. Math. and Math. Phys.*, 2:362–363, 1963.

[7] M. H. Gutknecht. Variants of BiCGStab for matrices with complex spectrum. *SIAM J. Sci. Comput.*, 14:1020–1033, 1993.

[8] M. H. Gutknecht and M. Rozložník. Residual smoothing techniques: Do they improve the limiting accuracy of iterative solvers? Technical report, Swiss Center for Scientific Computing, Zürich, 1999.

[9] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards*, 49:409–435, 1952.

[10] C. Lanczos. Solution of systems of linear equations by minimized iterations. *J. Res. Nat. Bur. Standards*, 49:33–53, 1952.

[11] M. Rozložník and R. Weiss. On stable implementations of the generalized minimal error method. *J. Comp. and Appl. Math.*, 98(1):49–62, 1998.

[12] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.*, 7(3):856–869, 1986.

[13] W. Schönauer. *Scientific Computing on Vector Computers*. North-Holland, Amsterdam, New York, Oxford, Tokyo, 1987.

[14] W. Schönauer, E. Schnepf, and H. Müller. The FIDISOL program package. Internal report 27/85, University of Karlsruhe, Computing Center, 76128 Karlsruhe, Germany, 1985.

[15] G. L. G. Sleijpen and D. R. Fokkema. BiCGStab(l) for linear equations involving unsymmetric matrices with complex spectrum. *Electronic Transactions on Numer. Anal. (ETNA)*, 1:11–32, 1993.

[16] G. L. G. Sleijpen, H. A. van der Vorst, and D. R. Fokkema. BiCGSTAB(l) and other hybrid Bi-CG methods. *Numerical Algorithms*, 7:75–109, 1994.

[17] H. A. van der Vorst. BI-CGSTAB: A fast and smoothly converging variant of BI-CG for the solution of nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 13(2):631–644, 1992.

[18] P. K. W. Vinsome. ORTHOMIN, an iterative method for solving sparse sets of simultaneous linear equations, paper SPE 5739. In *Symp. Numer. Simulation of Reservoir Performance of SPE of AIME, 4th, Los Angeles, California*, 1976.

[19] R. Weiss. Error-minimizing Krylov subspace methods. *SIAM J. Sci. Comput.*, 15(3):511–527, 1994.

[20] R. Weiss. *Parameter-Free Iterative Linear Solvers*. Mathematical Research, vol. 97. Akademie Verlag, Berlin, 1996.

[21] R. Weiss, H. Häfner, and W. Schönauer. LINSOL (LINear SOLver)—description and user's guide for the parallelized version. Internal report 61/95, University of Karlsruhe, Computing Center, 76128 Karlsruhe, Germany, 1995.

[22] Q. Ye. A breakdown-free variation of the nonsymmetric Lanczos algorithms. *Math. of Comput.*, 62(205):179–207, 1994.

[23] S.-L. Zhang. GPBi-CG: Generalized product-type methods based on Bi-CG for solving nonsymmetric linear systems. *SIAM J. Sci. Comput.*, 18(2):537–551, 1997.

[24] L. Zhou and H. F. Walker. Residual smoothing techniques for iterative methods. *SIAM J. Sci. Comput.*, 15(2):297–312, 1994.