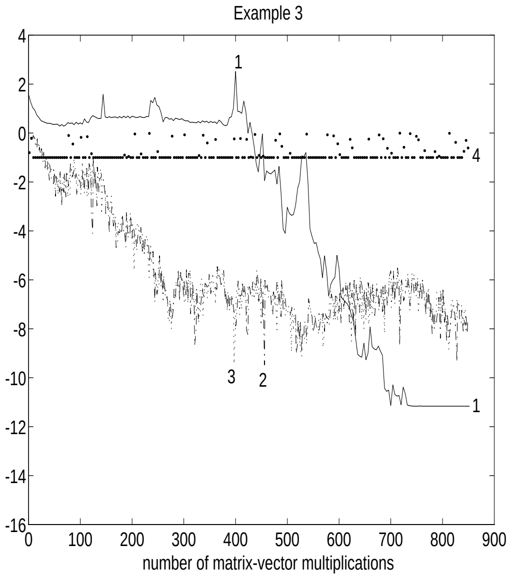


Universiteit Utrecht

Department
of
Mathematics

Maintaining convergence properties of BiCGstab methods in finite precision arithmetic

by

Gerard L.G. Sleijpen and Henk A. Van der Vorst

Preprint
nr. 861
July 4, 1994

revised version
February, 1995

Maintaining convergence properties of BiCGstab methods in finite precision arithmetic

GERARD L.G. SLEIJPEN* AND HENK A. VAN DER VORST*

Abstract. It is well-known that Bi-CG can be adapted so that hybrid methods with computational complexity almost similar to Bi-CG can be constructed, in which it is attempted to further improve the convergence behavior. In this paper we will study the class of BiCGstab methods.

In many applications, the speed of convergence of these methods appears to be determined mainly by the incorporated Bi-CG process, and the problem is that the Bi-CG iteration coefficients have to be determined from the BiCGstab process. We will focus our attention to the accuracy of these Bi-CG coefficients, and how rounding errors may affect the speed of convergence of the BiCGstab methods. We will propose a strategy for a more stable determination of the Bi-CG iteration coefficients and by experiments we will show that this indeed may lead to faster convergence.

Key words. Non-symmetric linear systems, iterative solvers, Bi-CG, Bi-CGSTAB, BiCGstab(ℓ).

AMS subject classification. 65F10.

1. Introduction. The BiCGstab methods can be viewed as Bi-CG combined with repeated low degree GMRES processes, like GMRES(1) in Bi-CGSTAB. Therefore, we start with a brief overview of Bi-CG.

Bi-CG [7, 12] is an iterative solution method for linear systems

$$Ax = b \tag{1.1}$$

in which the $n \times n$ matrix A is nonsingular. In typical applications n will be large and A will be sparse. For ease of presentation, we assume A and b to be real.

Starting with an initial guess x_0 for the solution x and a "shadow" residual $\widetilde{r}_0$ (most often one takes $\widetilde{r}_0 = r_0$), Bi-CG produces sequences of approximations x_k, residuals r_k, and search directions u_k by

$$u_k = r_k - \beta_k u_{k-1}, \quad x_{k+1} = x_k + \alpha_k u_k, \quad r_{k+1} = r_k - \alpha_k A u_k, \tag{1.2}$$

where the Bi-CG coefficients α_k and β_k are such that r_k and Au_k are orthogonal to the shadow Krylov subspace $\mathcal{K}_k(A^T;\widetilde{r}_0)$.

In principle we are free to select any basis for the shadow Krylov subspace that suits our purposes. We will represent the basis vectors of this subspace in polynomial form.

If (ψ_k) is a sequence of polynomials of degree k with a non-trivial leading coefficient θ_k then the vectors $\psi_0(A^T)\widetilde{r}_0, \ldots, \psi_{k-1}(A^T)\widetilde{r}_0$ form a basis of $\mathcal{K}_k(A^T;\widetilde{r}_0)$ and we have (see [24] or [21]):

$$\beta_k = \frac{\theta_{k-1}}{\theta_k}\frac{\rho_k}{\sigma_{k-1}} \quad \text{and} \quad \alpha_k = \frac{\rho_k}{\sigma_k} \quad \text{where} \quad \left\{ \begin{array}{l} \rho_k := (r_k, \psi_k(A^T)\widetilde{r}_0), \\ \sigma_k := (Au_k, \psi_k(A^T)\widetilde{r}_0). \end{array} \right. \tag{1.3}$$

In finite precision arithmetic computation the values of the iteration coefficients depend quite critically on the choice of the basis vectors for the shadow Krylov subspace. For example, if we make the straightforward choice $\psi_k(t) = t^k$, then the basis vectors tend to be more and more in the direction of the dominating eigenvector of A^T. Depending on how well the dominant eigenvalue of A^T is separated from the others, this would imply that eventually the new vectors in $\mathcal{K}_k(A;r_0)$ are effectively made only orthogonal with respect to this dominating eigenvector. This then would lead to a new vector r_k, that is almost orthogonal to the k-th basis vector for the shadow Krylov subspace, and hence we may expect large relative errors in the new iteration coefficients.

Of course, even if all computational steps are done as accurately as possible (in finite precision), eventually the computed Bi-CG coefficients will differ in all digits from the exact ones. As is well known

*Mathematical Institute, University of Utrecht, P.O. Box 80.010, NL-3508 TA Utrecht, The Netherlands, E-mail: sleijpen@math.ruu.nl vorst@math.ruu.nl.

for the Bi-CG process itself (cf. [16, 10, 1]), this can be attributed to a global loss of bi-orthogonality. Since Bi-CG seems to work rather well as long as some local bi-orthogonality is maintained (that means that the local bi-orthogonalization is done accurately enough, see also [9]), we expect to recover the convergence behavior of the incorporated Bi-CG process (in the BiCGstab methods) if we compute the iteration coefficients as accurately as possible. Therefore, we want to avoid all additional perturbations that might be introduced by an unfortunate choice of the polynomial process that is carried out on top of the Bi-CG process.

In Sect. 2 we will study the choice of the set ψ_k, and we will identify polynomials that lead to sufficiently stable computation of the Bi-CG iteration coefficients.

The polynomials ψ_k can also be used for a different purpose in the Bi-CG process. Sonneveld [24] was the first to suggest to rewrite the inner products, not only to avoid the operations with A^T, e.g.,

$$\rho_k = (r_k, \psi_k(A^T)\widetilde{r}_0) = (\psi_k(A)r_k, \widetilde{r}_0) = (\mathbf{r}_k, \widetilde{r}_0), \tag{1.4}$$

but also to allow the construction of recursions for the vectors $\mathbf{r}_k := \psi_k(A)r_k$. In this way the polynomials ψ_k can be used for a further reduction of the residual in some norm. In fact Sonneveld suggested the specific choice $\psi_k = \phi_k$, where $r_k = \phi_k(A)r_0$ (i.e., ϕ_k is the Bi-CG iteration polynomial), and this led to the well-known CGS method [24]. More recently, other hybrid Bi-CG methods have emerged as well. In all these approaches the Bi-CG iteration vectors are not computed explicitly.

In the BiCGstab methods [25, 11, 23] ψ_k is chosen as the product of low degree minimum residual (like GMRES) polynomials. We will study these choices in Sect. 3 in view of our new insights on the choice of the ψ_k. It will turn out that the quest for a stable computation of the iteration coefficients is not always in concordance with optimal residual reducing properties.

Conventions 1. Throughout this paper $\|\cdot\|$ will denote the Euclidean norm.
2. For methods to be discussed, the residual r_k at the kth step is in $\mathcal{K}_{k+1}(A; r_0)$ and can be written as a kth degree polynomial ψ_k in A acting on r_0: $r_k = \psi_k(A)r_0$ and $\psi_k(0) = 1$. In connection with this we will call the polynomial for method M, the *M-polynomial* associated with A and r_0 (see, e.g., [26]), or the *M(k)-polynomial* if we want to specify the degree k. In particular the *OR-polynomial* ψ_k^{OR} corresponds to the situation where the residual is orthogonal with respect to the Krylov subspace $\mathcal{K}_k(A; r_0)$ (FOM [18] and GENCG [4] define implicitly OR-polynomials). The *MR-polynomial* ψ_k^{MR} defines the residual that is minimal in the Krylov subspace $\mathcal{K}_{k+1}(A; r_0)$ (as in GMRES [19]).
3. We will often use phrases like "reduces the residual" or "small residual". This will always mean that these residual vectors are reduced (or are small) with respect to the Euclidean norm.

2. The Bi-CG iteration coefficients. To understand why the Bi-CG coefficients can be inaccurate, we concentrate first on the ρ_k (see (1.4)). Then, as we will show, the effects of inaccurate computation of σ_k can be understood in a similar way. The computed ρ_k will be inaccurate if r_k is nearly orthogonal to $\psi_k(A^T)\widetilde{r}_0$. As is well known, this will happen if the incorporated Lanczos process nearly breaks down (i.e. $(\phi_k(A)r_0, \psi_k(A^T)\widetilde{r}_0) \approx 0$ for any polynomial ψ_k of exact degree k), and this may be attributed to an unlucky choice of $\widetilde{r}_0$. This kind of break down may be circumvented by so-called look-ahead techniques (see, e.g., [17, 9]). However, apart from this a bad choice of ψ_k may lead to a small ρ_k as well. Here, we only consider how this choice of ψ_k causes further instability in the computation of the iteration coefficients. We assume that the Lanczos process itself does not (nearly) break down.

The relative error ε_k, due to rounding errors, in ρ_k can be bounded sharply by (see, e.g., [6])

$$|\varepsilon_k| \leq 1.01\, n\, \overline{\xi}\, \frac{(|r_k|, |\psi_k(A^T)\widetilde{r}_0|)}{|(r_k, \psi_k(A^T)\widetilde{r}_0)|} \leq 1.01\, n\, \overline{\xi}\, \frac{\|r_k\|\, \|\psi_k(A^T)\widetilde{r}_0\|}{|(r_k, \psi_k(A^T)\widetilde{r}_0)|} \leq \frac{1.01\, n\, \overline{\xi}}{\widehat{\rho}_k}, \tag{2.1}$$

where

$$\widehat{\rho}_k := \frac{|(r_k, \psi_k(A^T)\widetilde{r}_0)|}{\|r_k\|\, \|\psi_k(A^T)\widetilde{r}_0\|}, \tag{2.2}$$

$\overline{\xi}$ is the relative machine precision and n the dimension of the problem. For a small relative error we want to have $\widehat{\rho}_k$ (the scaled ρ_k) as large as possible.

Because of the orthogonality of r_k with respect to $\mathcal{K}_k(A^T;\widetilde{r}_0)$ it follows that

$$(r_k, \psi_k(A^T)\widetilde{r}_0) = \gamma_k (r_k, (A^T)^k \widetilde{r}_0), \tag{2.3}$$

where γ_k is the leading coefficient of ψ_k: $\psi_k(t) = \sum_{j\leq k} \gamma_j t^j$. Hence,

$$\widehat{\rho}_k = \frac{|\gamma_k|}{\|\psi_k(A^T)\widetilde{r}_0\|} \frac{|(r_k, (A^T)^k \widetilde{r}_0)|}{\|r_k\|}. \tag{2.4}$$

The second quotient in this expression for $\widehat{\rho}_k$ does not depend on ψ_k, and therefore, the expression for $\widehat{\rho}_k$ is maximal over all polynomials ψ_k with fixed leading coefficient γ_k, when ψ_k is an appropriate multiple of the OR-polynomial ψ_k^{OR} associated with A^T and $\widetilde{r}_0$, that is

$$\psi_k^{\mathrm{OR}}(A^T)\widetilde{r}_0 \perp \mathcal{K}_k(A^T;\widetilde{r}_0), \quad \text{and} \quad \psi_k^{\mathrm{OR}}(0) = 1.$$

The appropriate multiple is to take care that the polynomial has leading coefficient γ_k, but since the expression for $\widehat{\rho}_k$ is invariant under scaling for ψ_k, we conclude that the OR-polynomial is the polynomial that makes $\widehat{\rho}_k$ maximal.

For $\sigma_k = (Au_k, \psi_k(A^T)\widetilde{r}_0)$ we can follow the same line of reasoning. Since by construction in Bi-CG the vectors Au_k are orthogonal to lower dimensional shadow Krylov subspaces, it follows that the relative error in σ_k, due to rounding errors can be bounded by an expression which is also minimal for ψ_k^{OR}. Note that the Bi-CG iteration coefficients are formed from ratios of ρ_k's and σ_k's, so that errors in each of these add to the inaccuracy in the α_k's and β_k's.

Since Bi-CG is designed to avoid all the work for the construction of an orthogonal basis, it would be expensive to construct the ψ_k^{OR} as the basis generating polynomials for the shadow Krylov subspace. In Bi-CG a compromise is made by taking the $\psi_k = \phi_k$ which creates a bi-orthogonal basis. At least for (near) symmetric matrices this is (almost) optimal.

3. Small residuals and accurate Bi-CG coefficients. Now the question arises whether we can select polynomials ψ_k, with $\psi_k(0) = 1$, that satisfy the following requirements:

1. ψ_k leads to sufficiently stable Bi-CG coefficients,
2. ψ_k can be used to further reduce the Bi-CG residual, that is $\mathbf{r}_k = \psi_k(A)r_k$ is (much) smaller in norm than r_k,
3. the ψ_k can be (implicitly) formed by short recurrences.

3.1. Some choices for the polynomial ψ_k. A choice that comes close, in many relevant cases, to fulfilling all these requirements is the one suggested by Sonneveld [24]: $\psi_k = \phi_k$, which leads to CGS. However, there are two disadvantages associated with this choice. The first is that there is no reason why ϕ_k should lead to a further reduction (and often it does not), the second is that all irregularities in the convergence behavior of Bi-CG are magnified in CGS (although the negative effects of this on the accuracy of the approximated solution can be largely reduced [15, 22]).

An obvious alternative is to select ψ_k as the product of k first degree MR (or GMRES) polynomials, which leads to Bi-CGSTAB [25], or as n/ℓ factors of ℓ degree MR-polynomials [21, 23].

An obvious problem, when replacing the Bi-CG polynomial by other polynomials which are chosen as to reduce the residual vector, is that these polynomials do not necessarily lead implicitly to an optimal basis for the shadow Krylov subspace.

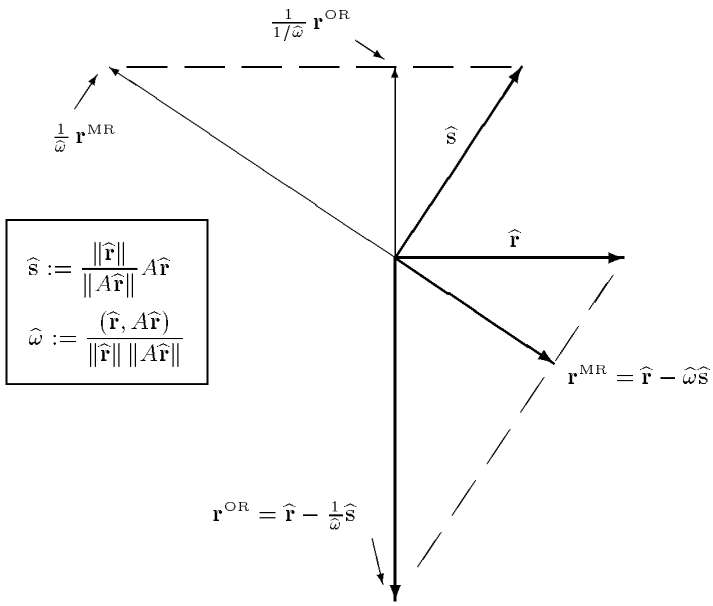
We will first concentrate on suitable (inexpensive) polynomial methods that help to further reduce the Bi-CG residual. For hybrid Bi-CG methods, where $\mathbf{r}_k = \psi_k^{\mathrm{H}}(A)r_k$, we have that ρ_k is computed as $\rho_k = (\mathbf{r}_k, \widetilde{r}_0)$ (cf. (1.4)), and in finite precision arithmetic there will be a relative error $\varepsilon_k^{\mathrm{H}}$ in the evaluation of this inner product that can be bounded as

$$|\varepsilon_k^{\mathrm{H}}| \leq 1.01\, n\, \overline{\xi}\, \frac{(|\mathbf{r}_k|, |\widetilde{r}_0|)}{|(\mathbf{r}_k, \widetilde{r}_0)|} \leq 1.01\, n\, \overline{\xi}\, \frac{\|\mathbf{r}_k\|\, \|\widetilde{r}_0\|}{|(\mathbf{r}_k, \widetilde{r}_0)|} \leq \frac{1.01\, n\, \overline{\xi}}{\widehat{\rho}_k^{\mathrm{H}}}, \tag{3.1}$$

where

$$\widehat{\rho}_k^{\mathrm{H}} := \frac{|(\mathbf{r}_k, \widetilde{r}_0)|}{\|\mathbf{r}_k\|\, \|\widetilde{r}_0\|}. \tag{3.2}$$

FIG. 3.1. *Amplification effects of GMRES(1) and FOM(1).*



For similar reasons as in Sect. 2, we have

$$\widehat{\rho}_k^{\rm H} = \frac{|\gamma_k|}{\|\psi_k^{\rm H}(A)r_k\|}\,\frac{|(A^k r_k, \widetilde{r}_0)|}{\|\widetilde{r}_0\|}, \tag{3.3}$$

where γ_k is the leading coefficient of $\psi_k^{\rm H}$. Again, the OR-polynomial $\psi_k^{\rm OR}$ (associated with A and r_k), which minimizes $\|\psi_k^{\rm H}(A)r_k\|/|\gamma_k|$, would lead to a maximal value for $\widehat{\rho}_k^{\rm H}$. Of course, there is no guarantee or reason why this polynomial should also maximize the expression for $\widehat{\rho}_k$ in the Bi-CG representation for the inner product, but given the fact that we want that the polynomial acts directly on r_k this is the best we can do.

In the BiCGstab methods the $\psi_k^{\rm H}$ is chosen as a product of MR-polynomials, for obvious reasons. The MR-polynomial (associated with A and r_k) minimizes $\|\mathbf{r}_k\| = \|\psi_k^{\rm H}(A)r_k\|$ over all polynomials $\psi_k^{\rm H}$ for which $\psi_k^{\rm H}(0) = 1$, and hence seems to be more appropriate for creating small residuals $\mathbf{r}_k$. But it would be too expensive to perform k steps of FOM or GMRES in order to compute $\mathbf{r}_k = \psi_k^{\rm H}(A)r_k$ (using r_k as initial residual): we not only strive for a large reduction but also for inexpensive steps. The product of MR(1)-polynomials (i.e. MR-polynomials of degree 1) as in Bi-CGSTAB is a compromise between the wish for small residuals and inexpensive steps. In view of our discussion above, however, we might also consider OR(1)-polynomials in order to achieve more stability in some cases, giving up some of the reduction.

With OR(1)-polynomials $(1 - \omega_k\lambda)$ for which $(I - \omega_k A)\widehat{\mathbf{r}} \perp \widehat{\mathbf{r}}$, where $\widehat{\mathbf{r}} := \psi_{k-1}^{\rm H}(A)r_k$, we compromise between (locally) more accurate coefficients and inexpensive steps. Although these polynomials occasionally cure stagnation of Bi-CGSTAB (see also [5]) they also may amplify residuals, which again leads to inaccurate approximations (as explained by (7) in [23]) or even to overflow.

We will now show that, if the angle between $\widehat{\mathbf{r}}$ and $A\widehat{\mathbf{r}}$ is more than 45° (i.e. $|(\widehat{\mathbf{r}}, A\widehat{\mathbf{r}})| \leq \frac{1}{2}\sqrt{2}\,\|\widehat{\mathbf{r}}\|\,\|A\widehat{\mathbf{r}}\|$) then the OR(1)-polynomial locally amplifies the residual, while the MR(1)-polynomial leads to a smaller value for $\widehat{\rho}_k^{\rm H}$ (cf. (3.3)).

This property is easily explained with Fig. 3.1, where $\widehat{\mathbf{s}}$ is a scalar multiple of $A\widehat{\mathbf{r}}$, scaled such that $\|\widehat{\mathbf{s}}\| = \|\widehat{\mathbf{r}}\|$. The residual $\mathbf{r}^{\rm OR}$ is obtained by applying the OR(1)-polynomial to $\widehat{\mathbf{r}}$, $\mathbf{r}^{\rm MR}$ results from the MR(1)-polynomial. Clearly, in the situation as sketched in the figure, $\|\mathbf{r}^{\rm MR}\| < \|\widehat{\mathbf{r}}\| < \|\mathbf{r}^{\rm OR}\|$, while scaling the polynomials such that the leading coefficient is identical (in the figure $\|\widehat{\mathbf{r}}\|/\|A\widehat{\mathbf{r}}\|$) changes the order: $\|\mathbf{r}^{\mathrm{OR}}\|/|1/\widehat{\omega}| < \|\widehat{\mathbf{s}}\| < \|\mathbf{r}^{\mathrm{MR}}\|/|\widehat{\omega}|$.

We have to be careful with such amplifications when they are extremely large or when they occur in a consecutive number of iteration steps (as will be the case in a stagnation phase of the process). In such cases, any of the two choices may slow down the convergence: the OR(1)-polynomials because of amplifying $\|\mathbf{r}_k\|$; the MR(1)-polynomials because of shrinking $\widehat{\rho}_k^{\mathrm{H}}$ and thus affecting the accuracy of the Bi-CG coefficients. Apparently, Bi-CGSTAB may also be expected to converge poorly if in a number of consecutive steps the angle between $\widehat{\mathbf{r}}$ and $A\widehat{\mathbf{r}}$ is more than 45°. Especially when, in Bi-CGSTAB, GMRES(1) stagnates in a consecutive number of steps, it is important to have accurate Bi-CG coefficients, because, in such a case any convergence is to be expected only from the Bi-CG part of Bi-CGSTAB. Unfortunately, this is precisely the situation where the MR(1)-polynomials spoil the accuracy of the coefficients, while the OR(1)-polynomials spoil the accuracy of the approximations or lead to overflow. Therefore, we have to find a cure by other modifications.

We will suggest other choices for ω_k (as in (3.20), see also [5]) which occasionally cure these problems. However, they often cannot completely prevent poor convergence. One reason is that, with these first degree factors, we are (implicitly) building a power basis for the shadow Krylov subspace (if the ω_k are close to each other), and we have seen in Sect. 2 that this is highly undesirable.

We may expect better convergence results by performing a composition of ℓ steps, using minimizing polynomials of degree ℓ, with $\ell > 1$: that is, by taking ψ_k as a product of MR(ℓ)-polynomials as in BiCGstab(ℓ) (see [11, 21, 23]), provided that the MR(ℓ)-polynomials lead to significant reductions. We will consider this approach in much more detail.

In BiCGstab(ℓ) the polynomial ψ_k is constructed as a product of polynomials of degree ℓ: for $k = m\ell$, $\psi_k = p_{m-1}\cdot\ldots\cdot p_0$ where p_j is of degree ℓ. To investigate what properties these polynomial factors p_j should have, we consider $k = m\ell$, concentrate on p_m, and we define

$$\widehat{\mathbf{r}} := \psi_k(A)\phi_{k+\ell}(A)r_0 = \psi_k(A)r_{k+\ell}. \tag{3.4}$$

In BiCGstab(ℓ) the vector $\widehat{\mathbf{r}}$ is computed explicitly in the Bi-CG part of the algorithm. The new residual $\mathbf{r}_{k+\ell}$ will be $\mathbf{r}_{k+\ell} := p_m(A)\widehat{\mathbf{r}}$ (for implementational details, see [21, 23]).

3.2. Minimizing polynomials of low degree. For the derivations and results in this section we assume exact arithmetic. However, we expect that they also have some validity in finite precision arithmetic, and as we will see in Sect. 4, the experiments do largely confirm our expectations.

As before, we wish to maximize $\widehat{\rho}_k^{\mathrm{H}}$ for polynomials p_m of degree ℓ for which $p_m(0) = 1$. As we have seen in Sect. 2, this means that we have to restrict ourselves to polynomials with a fixed leading coefficient γ_ℓ.

In order to make our formulas more readable, we define

$$|\!|\!|\mathbf{r}|\!|\!| := \frac{\|\mathbf{r}\|}{|\gamma_\ell|} \quad \text{where} \quad \mathbf{r} := p(A)\widehat{\mathbf{r}} \quad \text{and} \quad p(t) = \sum_{j=0}^{\ell}\gamma_j t^j. \tag{3.5}$$

In order to have $\widehat{\rho}_{k+\ell}^{\mathrm{H}}$,

$$\widehat{\rho}_{k+\ell}^{\mathrm{H}} = \frac{|(\mathbf{r}, \widetilde{r}_0)|}{\|\mathbf{r}\|\|\widetilde{r}_0\|}, \tag{3.6}$$

(approximately) maximal, we should have that

$$|\!|\!|p(A)\widehat{\mathbf{r}}|\!|\!| \quad \text{is (approximately) minimal.} \tag{3.7}$$

As is well known $\mathbf{r}^{\mathrm{MR}}$ solves

$$\min\{\,\|\mathbf{r}\| \mid \mathbf{r} = p(A)\widehat{\mathbf{r}},\ \mathrm{degr}(p) \le \ell,\ p(0) = 1\}$$

if and only if

$$\mathbf{r}^{\mathrm{MR}} \perp A\widehat{\mathbf{r}}, A^2\widehat{\mathbf{r}}, \ldots, A^\ell\widehat{\mathbf{r}}. \tag{3.8}$$

In a similar way, one can show that $\mathbf{r}^{\mathrm{OR}}$ solves

$$\min\left\{\, |\!|\!|\mathbf{r}|\!|\!| \mid \mathbf{r}=p(A)\widehat{\mathbf{r}},\ \mathrm{degr}(p)\leq \ell,\ p(0)=1\right\}$$

(cf. (3.7)) if and only if

$$\mathbf{r}^{\mathrm{OR}} \perp \widehat{\mathbf{r}}, A\widehat{\mathbf{r}}, \ldots, A^{\ell-1}\widehat{\mathbf{r}}. \tag{3.9}$$

The residual $\mathbf{r}^{\mathrm{MR}}$ is the ℓth residual of a minimal residual method (as GMRES), and $\mathbf{r}^{\mathrm{OR}}$ is the ℓth residual of an orthogonal residual method (as FOM), each for the problem $Ax=b$ with the same initial residual $\widehat{\mathbf{r}}$.

The following theorem compares how good $\mathbf{r}^{\mathrm{MR}}$ is in maximizing $\widehat{\rho}_k^{\mathrm{H}}$ and how well $\mathbf{r}^{\mathrm{OR}}$ helps to reduce $\widehat{\mathbf{r}}$. A method like ORTHODIR produces explicitly an orthonormal basis for which Theorem 3.1 can be applied.

We will call a sequence of vectors $\widehat{\mathbf{r}}_1,\ldots,\widehat{\mathbf{r}}_\ell$ a *Krylov basis* for $\mathcal{K}_\ell(A;\widehat{\mathbf{r}}_1)$ if $\widehat{\mathbf{r}}_1,\ldots,\widehat{\mathbf{r}}_j$ and $\widehat{\mathbf{r}}_1,\ldots,A^{j-1}\widehat{\mathbf{r}}_1$ span the same space for each $j=1,\ldots,\ell$.

THEOREM 3.1. *Let $\widehat{\mathbf{r}}_1,\ldots,\widehat{\mathbf{r}}_{\ell-1}$ be an orthonormal Krylov basis for $\mathcal{K}_{\ell-1}(A;A\widehat{\mathbf{r}})$. Let the vectors $\widetilde{\mathbf{r}}_0$ and $\widetilde{\mathbf{r}}_\ell$ be obtained by orthogonalizing $\widehat{\mathbf{r}}$ and $A^\ell\widehat{\mathbf{r}}$ with respect to $\mathcal{K}_{\ell-1}(A;A\widehat{\mathbf{r}})$:*

$$\widetilde{\mathbf{r}}_0 := \widehat{\mathbf{r}} - \sum_{j=1}^{\ell-1}(\widehat{\mathbf{r}},\widehat{\mathbf{r}}_j)\,\widehat{\mathbf{r}}_j \quad \textit{and} \quad \widetilde{\mathbf{r}}_\ell := A^\ell\widehat{\mathbf{r}} - \sum_{j=1}^{\ell-1}(A^\ell\widehat{\mathbf{r}},\widehat{\mathbf{r}}_j)\,\widehat{\mathbf{r}}_j.$$

Let ϱ be defined as

$$\varrho := \frac{(\widetilde{\mathbf{r}}_\ell,\widetilde{\mathbf{r}}_0)}{\|\widetilde{\mathbf{r}}_\ell\|\,\|\widetilde{\mathbf{r}}_0\|}. \tag{3.10}$$

Then

$$\mathbf{r}^{MR} = \widetilde{\mathbf{r}}_0 - \varrho\,\frac{\|\widetilde{\mathbf{r}}_0\|}{\|\widetilde{\mathbf{r}}_\ell\|}\,\widetilde{\mathbf{r}}_\ell \quad \textit{and} \quad \mathbf{r}^{OR} = \widetilde{\mathbf{r}}_0 - \frac{1}{\varrho}\,\frac{\|\widetilde{\mathbf{r}}_0\|}{\|\widetilde{\mathbf{r}}_\ell\|}\,\widetilde{\mathbf{r}}_\ell, \tag{3.11}$$

$$\|\mathbf{r}^{MR}\| = \sqrt{1-\varrho^2}\,\|\widetilde{\mathbf{r}}_0\| \quad \textit{and} \quad \|\mathbf{r}^{OR}\| = \frac{\sqrt{1-\varrho^2}}{|\varrho|}\,\|\widetilde{\mathbf{r}}_0\|, \tag{3.12}$$

$$|\!|\!|\mathbf{r}^{MR}|\!|\!| = \frac{\sqrt{1-\varrho^2}}{|\varrho|}\,\|\widetilde{\mathbf{r}}_\ell\| \quad \textit{and} \quad |\!|\!|\mathbf{r}^{OR}|\!|\!| = \sqrt{1-\varrho^2}\,\|\widetilde{\mathbf{r}}_\ell\|. \tag{3.13}$$

Proof. One may easily verify that $s := \widetilde{\mathbf{r}}_0 - \gamma\widetilde{\mathbf{r}}_\ell = p(A)\widehat{\mathbf{r}}$ for some polynomial p of degree ℓ with leading coefficient γ and $p(0)=1$. Moreover, since $\widetilde{\mathbf{r}}_0$ and $\widetilde{\mathbf{r}}_\ell$ are orthogonal to $A\widehat{\mathbf{r}},\ldots,A^{\ell-1}\widehat{\mathbf{r}}$, the vector s is orthogonal to these vectors as well.
Define $\nu := \|\widetilde{\mathbf{r}}_0\|/\|\widetilde{\mathbf{r}}_\ell\|$. With $\gamma=\varrho\nu$, s is also orthogonal to $\widetilde{\mathbf{r}}_\ell$, and consequently, orthogonal to $A^\ell\widehat{\mathbf{r}}$. By (3.8), $s=r^{\mathrm{MR}}$ for this $\gamma=\varrho\nu$. Similarly, the choice $\gamma=\nu/\varrho$ makes s orthogonal to $\widetilde{\mathbf{r}}_0$, which implies that $s=\mathbf{r}^{\mathrm{OR}}$ (cf. (3.9)), and this completes the proof of (3.11).

The expressions in (3.12) for the norms of the optimal residuals follow from Pythagoras' theorem: $\widetilde{\mathbf{r}}_\ell \perp \mathbf{r}^{\mathrm{MR}}$ and $\widetilde{\mathbf{r}}_0 \perp \mathbf{r}^{\mathrm{OR}}$. Combining these expressions with the values for the leading coefficient γ ($\gamma=\varrho\nu$, $\gamma=\nu/\varrho$, respectively) gives (3.13). □

The vector $\widetilde{\mathbf{r}}_0$ in the theorem is precisely the residual in the $(\ell-1)$-th step of a minimal residual method. Therefore, if ς is the residual reduction step ℓ by this MR method (i.e. $\varsigma := \|\mathbf{r}_\ell^{\mathrm{MR}}\|/\|\mathbf{r}_{\ell-1}^{\mathrm{MR}}\|$) then $\varrho=\sqrt{1-\varsigma^2}$ and we have the following corollary.

COROLLARY 3.2.

$$\frac{\|\mathbf{r}_\ell^{MR}\|}{\|\mathbf{r}_\ell^{OR}\|} = |\varrho| \quad \textit{and} \quad \frac{|\!|\!|\mathbf{r}_\ell^{MR}|\!|\!|}{|\!|\!|\mathbf{r}_\ell^{OR}|\!|\!|} = \frac{1}{|\varrho|} \quad \textit{where} \quad \sqrt{1-\varrho^2} = \frac{\|\mathbf{r}_\ell^{MR}\|}{\|\mathbf{r}_{\ell-1}^{MR}\|}. \tag{3.14}$$

Property (3.12) can also be found in [3], [26] (there, the authors focussed on GMRES while our formulation follows the ORTHODIR approach).

3.3. Global effects. The concept of Lanczos breakdown is well known. We will see that it can be translated in terms of angles between the Krylov subspaces and their "shadow subspaces". It is less obvious what *near-breakdown of the Lanczos process* could mean. We will say that no near-breakdown takes place if the angle between the Krylov subspace $\mathcal{K}_k := \mathcal{K}_k(A; r_0)$ and the shadow Krylov subspace $\widetilde{\mathcal{K}}_k := \mathcal{K}_k(A^T; \widetilde{r}_0)$ is uniformly (in k) sufficiently smaller than $\frac{1}{2}\pi$:

$$\inf_k \cos\angle(\mathcal{K}_k, \widetilde{\mathcal{K}}_k) = \inf_k \left(\inf_{v \in \mathcal{K}_k} \sup_{\widetilde{v} \in \widetilde{\mathcal{K}}_k} \frac{|(v, \widetilde{v})|}{\|v\| \, \|\widetilde{v}\|} \right) \geq \kappa > 0. \tag{3.15}$$

In particular, we then have that

$$\inf_k \sup \left\{ \frac{|(r_k, \widetilde{v})|}{\|r_k\| \, \|\widetilde{v}\|} \mid \widetilde{v} \in \widetilde{\mathcal{K}}_{k+1} \right\} > 0, \tag{3.16}$$

and, since $r_k \perp \widetilde{\mathcal{K}}_k$, we see that $(r_k, \widetilde{r}_k) \neq 0$ (no-breakdown of the Lanczos process).

For hybrid Bi-CG methods as CGS, Bi-CGSTAB and BiCGstab(ℓ) we translate property (3.16) as

$$\kappa_k := \sup \left\{ \frac{|(\psi(A) r_k, \widetilde{r}_0)|}{\|\psi(A) r_k\| \, \|\widetilde{r}_0\|} \mid \psi \text{ pol. of degree } k \right\}, \quad \kappa_k \geq \kappa > 0, \text{ for all } k. \tag{3.17}$$

We may expect to obtain the most accurate ρ_k by using the kth OR-polynomial $\phi_k^{\rm OR}$ of k steps of an OR method with initial residual r_k. Then, ρ_k may expect to be endowed with a relative error of size $n\overline{\xi}/\kappa_k$. In practise, if we use another polynomial ψ_k of degree k, we may expect an error in ρ_k that is larger than $n\overline{\xi}/\kappa_k$ by a multiplicative factor

$$\frac{\|\psi_k(A) r_k\|}{\|\phi_k^{\rm OR}(A) r_k\|}. \tag{3.18}$$

If this factor is large, say $\geq \kappa/(n\overline{\xi})$, the scalar ρ_k and hence the Bi-CG coefficients can not be expected to have any correct digit. It is difficult to analyze this factor as a function of k, since the initial residual r_k changes for each k. Clearly, since $\|\phi_{k+\ell}^{\rm OR}(A) r_{k+\ell}\| \leq \|p_m^{\rm OR}(A)\psi_k(A) r_{k+\ell}\| = \|p_m^{\rm OR}(A)\widehat{\mathbf{r}}\|$ and $\psi_{k+\ell} = p_m \psi_k = p_m \cdots p_0$,

$$\frac{\|\psi_{k+\ell}(A) r_{k+\ell}\|}{\|\phi_{k+\ell}^{\rm OR}(A) r_{k+\ell}\|} = \frac{1}{\varrho_m} \frac{\|p_m^{\rm OR}(A)\widehat{\mathbf{r}}\|}{\|\phi_{k+\ell}^{\rm OR}(A) r_{k+\ell}\|} \geq \frac{1}{\varrho_m} \quad \text{where} \quad \varrho_m := \frac{\|p_m^{\rm OR}(A)\widehat{\mathbf{r}}\|}{\|p_m(A)\widehat{\mathbf{r}}\|}. \tag{3.19}$$

In our discussion, we assume that the factor in (3.18) has at least the order of magnitude of the product $\prod_{j=0}^m 1/\varrho_j$ (with $k = m\ell$) of the factors $1/\varrho_j$ (in (3.19)) per sweep. The conclusion that we should avoid that $\prod 1/\varrho_j \geq \kappa/(n\overline{\xi})$ seems to be supported by results from numerical experiments (cf. Sect. 4).

3.4. Discussion. The OR(ℓ)-polynomial is the best choice for obtaining more accurate Bi-CG coefficients, but the MR(ℓ)-polynomial will do almost as well if there is a significant error reduction at the ℓth step of GMRES (with initial residual $\widehat{\mathbf{r}}$) (cf. (3.13) and (3.14)). In that case the effect of OR(ℓ) is practically equivalent with the effect of MR(ℓ) (cf. (3.12) and (3.14)), and MR(ℓ)-polynomials may be slightly preferable.

However, if GMRES (with initial residual $\widehat{\mathbf{r}}$) does not reduce the residual well at the ℓth step then $|\varrho| \ll 1$ (cf. (3.14)), and, hence, the MR(ℓ)-polynomial may be expected to lead to an inaccurate $\rho_{k+\ell}$, while the OR(ℓ)-polynomial will enlarge the residual significantly. Likewise, if in a consecutive number of sweeps with BiCGstab(ℓ) the factor $|\varrho|$ is less than $\frac{1}{2}\sqrt{2}$, say, then the choice of MR(ℓ)-polynomials may lead to inaccurate Bi-CG coefficients as well, because of an accumulation of rounding errors in ρ_k (and σ_k). On the other hand, the OR(ℓ)-polynomials may lead to unacceptable large residuals, or even to overflow, after a consecutive number of amplifications of the residual.

We propose to make a compromise by choosing some intermediate between the OR(ℓ)- and MR(ℓ)-polynomial in case of a poor reduction by the latter one. This choice is inspired by equation (3.11):

$$\mathbf{r}_{k+\ell} = \widetilde{\mathbf{r}}_0 - \widehat{\gamma}_k \frac{\|\widetilde{\mathbf{r}}_0\|}{\|\widetilde{\mathbf{r}}_\ell\|}\widetilde{\mathbf{r}}_\ell, \quad \text{where} \quad \widehat{\gamma}_k := \frac{\varrho}{|\varrho|}\max(|\varrho|, \Omega) \tag{3.20}$$

with $\widetilde{\mathbf{r}}_0$, $\widetilde{\mathbf{r}}_\ell$, and ϱ as in Theorem 3.1 and $\Omega \in [0, \infty)$. In Sect. 3.4.1 we will present some strategies for the choice of Ω and we will comment on computational details.
However, although this approach often helps to cure our instability problems, it is not a panacea which can circumvent always an occasionally poor convergence of BiCGstab(ℓ). In such a situation, increasing the value of ℓ may be an alternative.

We may expect better convergence for BiCGstab(ℓ) if we increase the value for ℓ, and we will argue why. It helps for our discussion to compare BiCGstab(ℓ) with Bi-CGSTAB (=BiCGstab(1)), that is, we compare 1 sweep of BiCGstab(ℓ) with ℓ sweeps of Bi-CGSTAB. In ℓ sweeps of Bi-CGSTAB, ℓ times a MR(1)-polynomial is applied (each time for a different starting vector: $\psi_{k+j}(A)r_{k+j+1}$, $j = 0, \ldots, \ell-1$). By selecting higher degree MR(ℓ)-polynomials, as in BiCGstab(ℓ), we hope to profit for two different reasons:

1. One sweep of GMRES(ℓ) may be expected to result in a better residual reduction than ℓ steps of GMRES(1),

   Remark. In the case that GMRES(1) reduces well, we do not have to fear a loss of speed of convergence due to inaccurate Bi-CG coefficients. However, since GMRES may accelerate (superlinear convergence behavior), we may expect to obtain a smaller residual $\mathbf{r}_k$ with BiCGstab(ℓ) than with Bi-CGSTAB.

2. ℓ steps of GMRES(1) contribute ℓ times to a decrease of $\widehat{\rho}_k$ (hence contributing ℓ times to increasingly larger rounding errors in ρ_k), while one sweep of GMRES(ℓ) contributes only once; the decreasing effect in each single step of ℓ steps of GMRES(1) may be expected to be comparable or worse than the effect of only one sweep with GMRES(ℓ).

   Remark. If the ℓth step of GMRES does not reduce well (i.e. we have a small $|\varrho|$) then the MR(ℓ)-polynomial "amplifies" the inaccuracy on $\rho_{k+\ell}$ by $1/|\varrho|$ (in comparison with the OR(ℓ)-polynomial). In such a situation we may not expect to obtain a significant reduction by any of the steps of GMRES(1). It is even worse: in ℓ steps of Bi-CGSTAB, we should expect an "amplification" by the GMRES(1) steps in the inaccuracy of $\rho_{k+\ell}$ by a factor like $(1/|\varrho|)^\ell$ or more. More specifically, in a stagnation phase of the iterative method, it is more likely to have rather accurate Bi-CG coefficients, when we use BiCGstab(ℓ), than with Bi-CGSTAB.

In our discussion, the expected superlinear convergence behavior of GMRES plays a rather important role. However, as is well known, GMRES may as well converge slower or even stagnate for quite a while in any phase of the iteration process. In order to profit from a possible good reduction of the MR($\ell - 1$)-polynomial, in cases where the MR(ℓ)-polynomial gives only a poor additional reduction, we may use the modification as suggested in (3.20).

3.4.1. Computational details. The computation of the leading coefficient and ϱ, as suggested in Theorem 3.1 and formula (3.20), can be done with relatively inexpensive operations involving ℓ-vectors only.

If $\widehat{\mathbf{r}}_0 = \widehat{\mathbf{r}}, \widehat{\mathbf{r}}_1, \ldots, \widehat{\mathbf{r}}_\ell$ is a Krylov basis of $\mathcal{K}_{\ell+1}(A; \widehat{\mathbf{r}})$ and $\mathbf{R} := [\widehat{\mathbf{r}}_0 | \ldots | \widehat{\mathbf{r}}_\ell]$ then $\widetilde{\mathbf{r}}_0 = \mathbf{R}\vec{\eta}_0$ and $\widetilde{\mathbf{r}}_\ell = \mathbf{R}\vec{\eta}_\ell$ for some $\vec{\eta}_0, \vec{\eta}_\ell \in I\!R^{\ell+1}$. Consider the $(\ell+1) \times (\ell+1)$-matrix $V := \mathbf{R}^T\mathbf{R}$, the inner product $< \vec{\eta}, \vec{\mu} > := \vec{\mu}^T V \vec{\eta}$ and norm $|\vec{\eta}\,| := \sqrt{< \vec{\eta}, \vec{\eta} >}$. Then (cf. (3.20))

$$\mathbf{r}_{k+\ell} = \mathbf{R}\left(\vec{\eta}_0 - \widehat{\gamma}_k \frac{|\vec{\eta}_0|}{|\vec{\eta}_\ell|}\vec{\eta}_\ell\right), \quad \text{with} \quad \widehat{\gamma}_k := \frac{\varrho}{|\varrho|}\max(|\varrho|, \Omega), \quad \varrho = \frac{< \vec{\eta}_\ell, \vec{\eta}_0 >}{|\vec{\eta}_\ell|\,|\vec{\eta}_0|} \tag{3.21}$$

and Ω some scalar in $[0, \infty)$.

With $\Omega = 0$ we have the MR(ℓ)-polynomial which gives optimal reduction of $\widehat{\mathbf{r}}$ with respect to $\|\cdot\|$. For very small ϱ or for a consecutive number of non-large ϱ (say, $|\varrho| < \frac{1}{2}$), $\Omega = 0$ may result in inaccurate Bi-CG coefficients, since, as compared with the OR-polynomial, the MR-polynomial amplifies $\widehat{\mathbf{r}}$ with respect to $\|\cdot\|$ by $1/|\varrho|$ (i.e. $\|\mathbf{r}^{\mathrm{MR}}\| = \|\mathbf{r}^{\mathrm{OR}}\|/|\varrho|$): if in m sweeps of BiCGstab(ℓ) each ϱ is $|\varrho| \leq \beta < 1$ we may expect an amplification of the relative rounding error by at most $(1/\beta)^m$ (comparing local effects of OR- and MR-polynomials and assuming that local differences accumulate).

With $\Omega = 1/|\varrho|$ we have the OR(ℓ)-polynomial. Although this polynomial gives optimal reduction of $\widehat{\mathbf{r}}$ with respect to $\|\cdot\|$, it may lead to large residuals.

With $\Omega > 0$ we may avoid large amplifications of the inaccuracy in the Bi-CG coefficients. The choice $\Omega = 1$ distributes negative effects of small ϱ equally amongst residual reduction (amplifying $\|\mathbf{r}^{\mathrm{MR}}\|$ by at most $\sqrt{2}$: $\|\mathbf{r}_{k+\ell}\| \leq \sqrt{2}\,\|\mathbf{r}^{\mathrm{MR}}\|$) and rounding error amplification (amplifying $\|\mathbf{r}^{\mathrm{OR}}\|$ by at most $\sqrt{2}$: $\|\mathbf{r}_{k+\ell}\| \leq \sqrt{2}\,\|\mathbf{r}^{\mathrm{OR}}\|$). For $\ell = 1$, this choice amplifies both $\|\widehat{\mathbf{r}}\|$ and $\|A\widehat{\mathbf{r}}\|$ whenever $|\varrho| < \frac{1}{2}$ (that is, $\|\mathbf{r}_{k+1}\| > \|\widehat{\mathbf{r}}\|$ and $\|\mathbf{r}_{k+1}\| > \|A\widehat{\mathbf{r}}\|$), and reduces both quantities for other values of ϱ. A significant amplification of the rounding error in a few steps need not worry us as long as the coefficients are still rather accurate (in 16 digit arithmetic a cumulative amplification by, say, 10^8 may still be very acceptable). Therefore, the best choice of Ω will depend on the length of the stagnation phase: using (3.20), we expect a cumulative rounding error amplification by at most $(1+\Omega^{-2})^{L/(2\ell)}$, where $2L$ is number of matrix-vector multiplications in the stagnation phase (again assuming that the local differences accumulate). Therefore, for larger ℓ a smaller Ω may be acceptable. Moreover, in the upper bound, we did not take into account the reducing effect of the OR-polynomial itself: the amplification factor $1/|\varrho|$ of the MR-polynomial may be harmless if $\|\mathbf{r}^{\mathrm{OR}}\|$ is small. For instance, for $\ell = 1$ and $|\varrho| < \frac{1}{2}\sqrt{2}$, we have that $\|\mathbf{r}^{\mathrm{MR}}\| < \|A\widehat{\mathbf{r}}\|$.

3.5. Minimizing polynomials and σ_k. Following the arguments for ρ_k, using the concept of near-breakdown of the LU-decomposition, we can see that, for accurate σ_k, the $\|A\psi_k(A)u_k\|$ should be as small as possible. In the mth sweep, where $\psi_k = p_{m-1}\cdots p_0$ is given ($k = m\ell$) and p_m is constructed, $\|Ap_m(A)\widehat{\mathbf{u}}\|$, with $\widehat{\mathbf{u}} := \psi_k(A)u_{k+\ell}$, should be as small as possible. Unfortunately, the construction of ψ_k is linked to the residual r_k. Note that we do not have this problem in the Bi-CG process. For Bi-CG, the polynomial ψ_k that will give the most accurate coefficients ρ_k and σ_k is linked to the initial shadow residual $\widetilde{r}_0$ (see Sect. 2): $\|\psi_k(A^T)\widetilde{r}_0\|$ should be minimal.

The following observation links the scaled σ_k (cf. (3.22)) to the scaled ρ_k and the residual r_k and gives some theoretical support to the strategy of concentrating on non-small scaled ρ_k only.

Since $Au_k = \frac{1}{\alpha_k}(r_k - r_{k+1})$ and $(\psi_k(A)r_{k+1}, \widetilde{r}_0) = 0$ we have,

$$\frac{1}{\widehat{\sigma}_k^{\mathrm{H}}} := \frac{\|A\psi_k(A)u_k\|\,\|\widetilde{r}_0\|}{|(A\psi_k(A)u_k, \widetilde{r}_0)|} = \frac{\|\psi_k(A)r_k - \psi_k(A)r_{k+1}\|\,\|\widetilde{r}_0\|}{|(\psi_k(A)r_k, \widetilde{r}_0)|}. \tag{3.22}$$

Hence, with $\mathbf{r}_k := \psi_k(A)r_k$, we can bound the scaled σ_k by the scaled ρ_k and the growth of the residual in step k:

$$\frac{1}{\widehat{\sigma}_k^{\mathrm{H}}} \leq \frac{\|\mathbf{r}_k\|\,\|\widetilde{r}_0\|}{|(\mathbf{r}_k, \widetilde{r}_0)|}\left(1 + \frac{\|\psi_k(A)r_{k+1}\|}{\|\psi_k(A)r_k\|}\right) = \frac{1}{\widehat{\rho}_k^{\mathrm{H}}}\left(1 + \frac{\|\psi_k(A)r_{k+1}\|}{\|\psi_k(A)r_k\|}\right). \tag{3.23}$$

If our strategy to choose the polynomial ψ_k prevents $\widehat{\rho}_k^{\mathrm{H}}$ te become too small and the residuals cannot grow much in one step then our strategy works for σ_k as well.

One can show that (see [14]),

$$\frac{\|r_{k+1}\|}{\|r_k\|} \leq \frac{\theta_+^{(k)}}{\theta_-^{(k)}}\frac{1}{\lambda_k}, \quad \text{with} \quad \lambda_k := \frac{|(r_{k+1}, \widetilde{r}_{k+1})|}{\|r_{k+1}\|\,\|\widetilde{r}_{k+1}\|}, \tag{3.24}$$

and $\theta_-^{(k)}$ and $\theta_+^{(k)}$ are maximal and minimal, respectively, such that,

$$\theta_-^{(k)} \leq \sup_{\widetilde{v}\in\widetilde{\mathcal{K}}_k} \frac{|(Av, \widetilde{v})|}{\|v\|\,\|\widetilde{v}\|} \leq \theta_+^{(k)} \quad \text{for all} \quad v \in \mathcal{K}_{k+1}. \tag{3.25}$$

The $\theta_j^{(k)}$ are the singular values of A projected on the Krylov subspace $\mathcal{K}_k$ and the quotient space $I\!R^n/\widetilde{\mathcal{K}}_{k+1}$. The LU-decomposition in Bi-CG breaks down in step k if and only if the smallest singular value $\theta_-^{(k)}$ is zero. By scaling by $\theta_+^{(k)}$, we obtain $\theta_-^{(k)}/\theta_+^{(k)}$, that can be viewed as a quantification of the *near-breakdown of the LU-decomposition.* ((3.25) can be related to the Babuška-Brezzi condition, well-known in mixed finite element theory, cf. [2], Sect. 4). Apparently, (3.24) tells us that the growth of the Bi-CG residuals at step k can be bounded in terms of the "distance" λ_k to Lanczos breakdown and the distance $\theta_-^{(k)}/\theta_+^{(k)}$ to LU-decomposition breakdown.

If Bi-CG incorporated in the BiCGstab process does not suffer from near breakdown of the Lanczos process nor from the LU-decomposition, then we expect

$$\|\psi_k(A)r_{k+1}\|/\|\psi_k(A)r_k\|$$

(cf. (3.23)) to be of moderate size.

4. Numerical experiments. In the previous sections, we have focussed on the scaled ρ_k (i.e. $\widehat{\rho}_k^{\mathrm{H}}$ as defined in (3.2)) and the scaled σ_k. Although these scaled values (where we have applied Cauchy-Schwartz) are actually smaller than the values that we display in the figures they do not differ significantly in our numerical examples. The figures also show that the values of $\widehat{\rho}_k^{\mathrm{H}}$ and $\widehat{\sigma}_k^{\mathrm{H}}$ are rather close to each other, as they should in view of our arguments in Sect. 3.5.

All figures show

- the $\log_{10}$ of the norm of the true residual $b - A\mathbf{x}_k$ (curve 1, — in figure) for $k = m\ell$,
- the $\log_{10}$ of $|(\mathbf{r}_k, \widetilde{r}_0)|/(|\mathbf{r}_k|, |\widetilde{r}_0|)$ (curve 2, $- \cdot - \cdot$ in figure), also for intermediate k,
- the $\log_{10}$ of $|(A\mathbf{u}_k, \widetilde{r}_0)|/(|A\mathbf{u}_k|, |\widetilde{r}_0|)$ (curve 3, $\cdots$ in figure),
- and the $\log_{\frac{1}{0.7}}$ of $|\widehat{\gamma}_k|$ (curve 4, the $\bullet$'s in figure), where $\widehat{\gamma}_k$ is the scaled leading coefficient of the polynomial p_m that we actually used (cf. 3.20) and 0.7 is the value for Ω if we modify the methods by (3.20) (i.e., if $\Omega \neq 0$).

We count the iteration phases by numbers of matrix-vector products (MV), since this makes it possible to compare effectively different BiCGstab variants.

The numerical results clearly indicate that limiting the relative size of the leading coefficient of the polynomial may help to cure the effects of stagnation of BiCGstab(ℓ), and also it may help to increasing the value of ℓ (our standard choice was $\ell = 1$ in these examples). In some situations one modification may help, while in other situations the other modification may help or a combination of both.

4.1. Example 1. First we consider an advection dominated 2-nd order partial differential equation, with Dirichlet boundary conditions, on the unit cube (this equation was taken from [13]):

$$-u_{xx} - u_{yy} - u_{zz} + 1000\, u_x = f.$$

The function f is defined by the solution $u(x,y,z) = \exp(xyz)\sin(\pi x)\sin(\pi y)\sin(\pi z)$. This equation was discretized using $10 \times 10 \times 10$ finite volumes and central differences for u_x, resulting in a seven-diagonal linear system of order 1000. No preconditioning has been used in this example (or other examples in order to make the differences between approaches more visible).

As we see, Bi-CGSTAB more or less stagnates (see Fig. 4.1), which creates the kind of situation that we were particularly interested in, and that we want to cure. After about 30 MVs, the scalars ρ_k and σ_k, and consequently the Bi-CG coefficients, do not have any correct digit. None of the scaled leading coefficients ϱ is extremely small (according to Fig. 4.1, $|\varrho| \geq (0.7)^{10} = 0.0282$; the $\bullet$'s in the figure show $\log_{\frac{1}{0.7}}|\varrho|$). Apparently, the accumulation of the amplifications of $\|\widehat{\mathbf{r}}\|$ by the MR(1)-polynomials leads to this situation. The graph of $|(\mathbf{r}_k, \widetilde{r}_0)|/(|\mathbf{r}_k|, |\widetilde{r}_0|)$ shows nicely the predicted exponential decrease for $k \leq 10$ (that is for less than 20 MVs; $\prod_{k\leq 10}|\varrho| \geq (0.0282)^{10} = 3.18\,_{10}{-16}$). For larger k, $k \geq 11$, these values are smaller than the machine precision.

Although the modification with $\Omega = 0.7$ improves the relative size of the ρ_k and σ_k (only after $k = 50$ we lose all significant digits, cf. Fig. 4.2), we now even have divergence: the amplification of $\|\widehat{\mathbf{r}}\|$ per step

FIG. 4.1. *Standard* Bi-CGSTAB

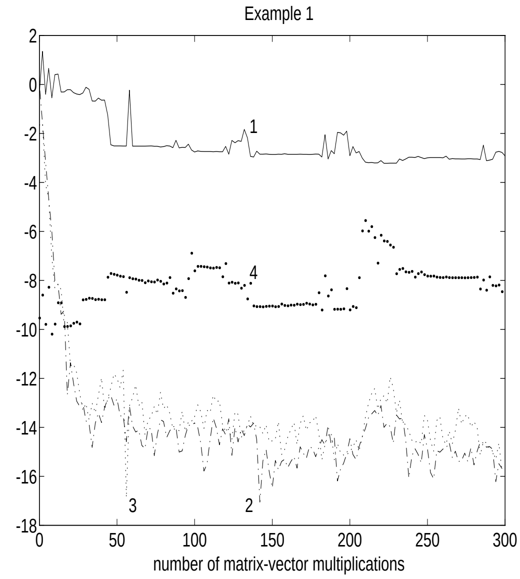


FIG. 4.2. *Modified* $\Omega = 0.7$

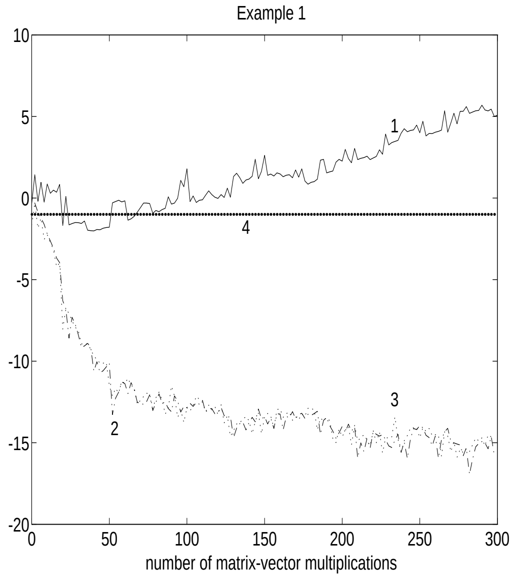


FIG. 4.3. *Standard* BiCGstab(2)

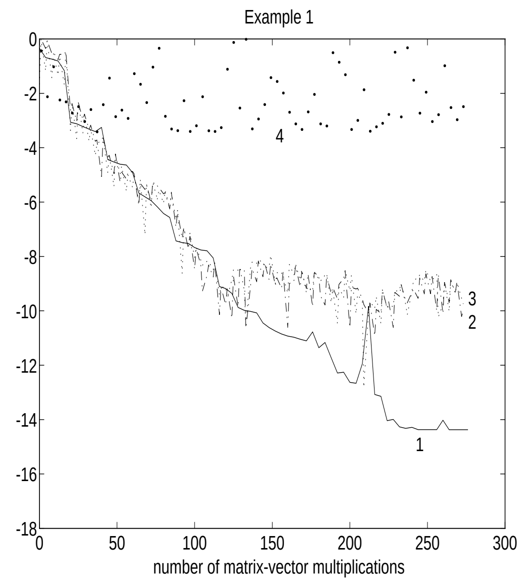
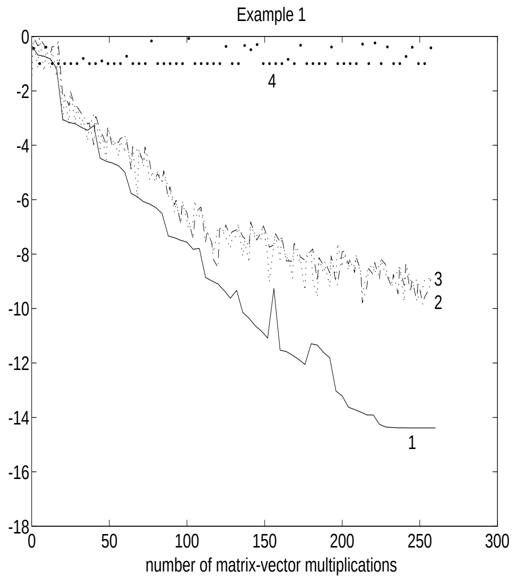


FIG. 4.4. *Modified* $\Omega = 0.7$



is not compensated by better convergence behavior of the incorporated Bi-CG process. Note that, for $k \leq 10$ (that is, $\#\mathrm{MV} \leq 20$), $|(\mathbf{r}_k, \widetilde{r}_0)|/(|\mathbf{r}_k|, |\widetilde{r}_0|)$ decreases proportionally to $(1/\sqrt{1+\Omega^{-2}})^k = 0.573^k$.

Using BiCGstab(2) improves the situation: both versions of BiCGstab(ℓ) converge nicely. According to Fig. 4.3, $|\varrho| \geq (0.7)^3 = 0.343$. During the first 60 MVs, we have to deal with these ϱ's 15 times. This leads us to expect a decrease of $|(\mathbf{r}_k, \widetilde{r}_0)|/(|\mathbf{r}_k|, |\widetilde{r}_0|)$ by at most $\geq (0.343)^{15} = 1.07_{10}{-}7$. This is confirmed by the numerical results. Here, BiCGstab(2) seems to be able to retain locally five correct digits of the Bi-CG coefficients, which is, apparently, enough to keep the incorporated Bi-CG process converging as it should.

The modification with $\Omega = 0.7$ further improves the situation (but only slightly).

FIG. 4.5. *Standard* Bi-CGSTAB

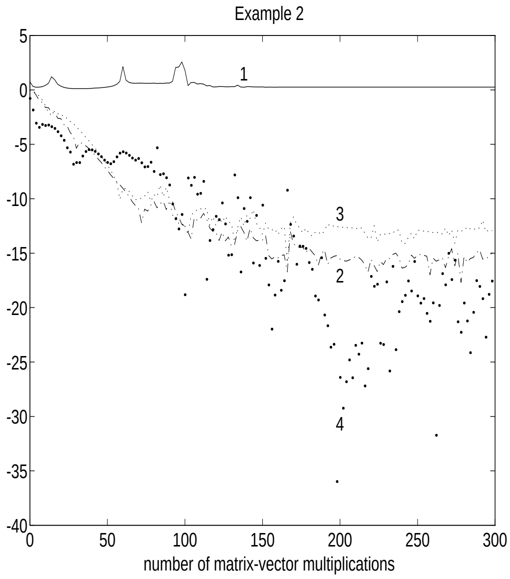
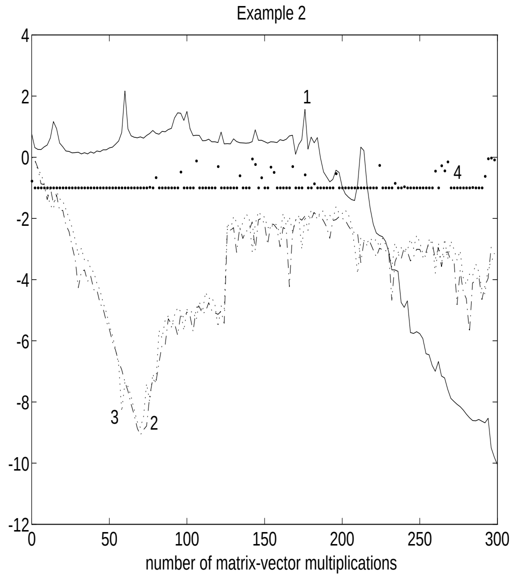


FIG. 4.6. *Modified* $\Omega = 0.7$



4.2. Example 2. For this comparison we have chosen the matrix that arises from a 63×63 finite volume discretization of

$$-u_{xx} - u_{yy} + \gamma(xu_x + yu_y) + \beta u = f \tag{4.1}$$

on the unit cube with Dirichlet boundary conditions, with $\gamma = 100$ and $\beta = -200$. This example has been suggested in [8]. We have chosen the right hand side b such that the solution x of the equation $Ax = b$ is the vector $(1, 1, \ldots, 1)^T$. The zero vector was used as an initial guess.

Here Bi-CGSTAB stagnates also (see Fig. 4.5). After about 150 MVs, the Bi-CG coefficients do not have any correct digits left, due to the cumulative effect of non-small $|\varrho|$'s ($|\varrho| \approx (0.7)^5 = 0.168$). The modification with $\Omega = 0.7$ (cf. Fig. 4.6) improves the relative size of the ρ_k and σ_k significantly, enough to survive the phase where $|\varrho|$'s are ≈ 0.17. As soon as the MR(1)-polynomial is more effective in reducing the norm of the residual (for #MV≥ 80, we find $|\varrho| \geq 0.7$), the relative size of the ρ_k and σ_k grows.

With BiCGstab(2) the situation does not seem to improve (see Fig. 4.7). In the initial phase, the values of $|\varrho|$ for the MR(2)-polynomials are much smaller than the values of $|\varrho|$ for the MR(1)-polynomials. Although, up to #MV≈ 80 the decrease of the scaled ρ_k and σ_k is a little bit better than with Bi-CGSTAB, this is apparently not enough to help survive the phase where the $|\varrho|$ is too small. However, (not shown in the figure) BiCGstab(2) did eventually converge (that is, $\|\mathbf{r}_k\| \leq 10^{-14}\|\mathbf{r}_0\|$) in 420 MVs, needing 120 MV more than the modified Bi-CGSTAB. 120 MVs is the length of the phase where the scaled ρ_k and σ_k are very small ($\approx 10^{-10}$).

The convergence behavior of modified BiCGstab(2) with $\Omega = 0.7$ (see Fig. 4.8) is comparable to the one of modified Bi-CGSTAB. (in Fig. 4.6).

4.3. Example 3. For this example we have selected a problem similar to the one in Example 2. Here, the matrix arises from a 66×66 finite volume discretization of (4.1), now with $\gamma = 1000$ and $\beta = 10$ as in [20].

The example shows that a combination of our strategies for Bi-CGSTAB (increasing ℓ and limiting the size of the scaled leading coefficient $\widehat{\gamma}_k$, cf. (3.21)) can cure the accuracy problems, whereas each of the strategies independently may fail.

As Fig. 4.10 shows, limiting the leading coefficient improves the accuracy of the Bi-CG coefficients in the initial phase of the process but can not prevent a loss of all digits (for #MV> 200). Now the

FIG. 4.7. *Standard* BiCGstab(2)

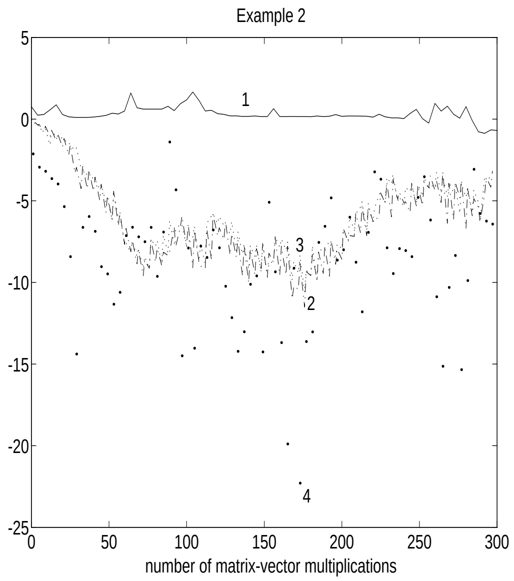


FIG. 4.8. *Modified* $\Omega = 0.7$

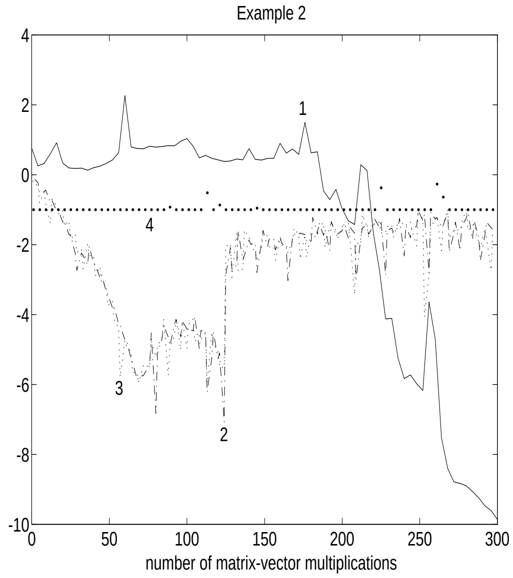


FIG. 4.9. *Standard* Bi-CGSTAB

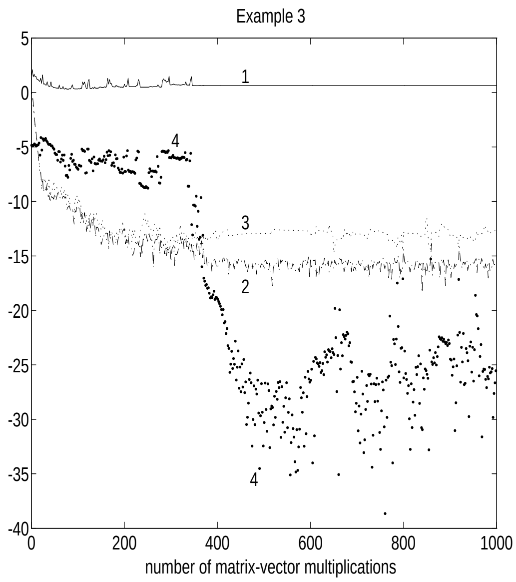
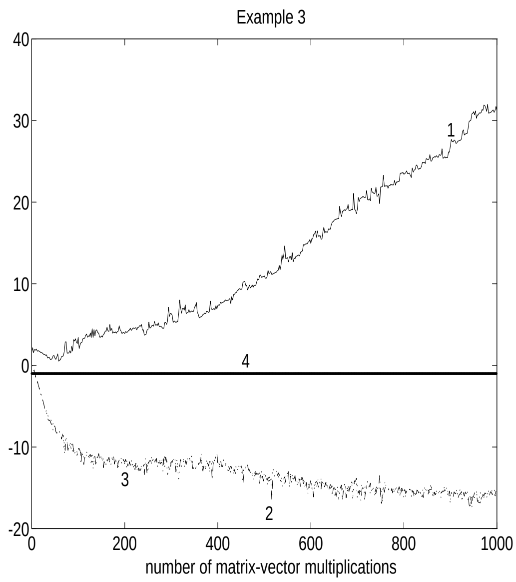


FIG. 4.10. *Modified* $\Omega = 0.7$



amplifying effect on $\|\mathbf{r}_k\|$ of this choice of the polynomial can clearly be seen (from #MV= 200 to #MV= 400 the residual grows with $(\sqrt{1+\Omega^2})^{100} = 4.6\,10^8$).

5. Conclusions. In order to maintain the convergence properties of the Bi-CG component in hybrid Bi-CG methods, it is necessary to select polynomial methods for the hybrid part that permit to compute the Bi-CG coefficients as accurately as possible.

The (dynamical) combination of two strategies for the improvement of the local accuracy seems to be very attractive. These strategies are:

A. Take products of degree ℓ polynomials with $\ell > 1$ as in BiCGstab(ℓ) rather than of degree 1 polynomials as in Bi-CGSTAB.

B. Try to limit the size of the leading coefficient of these polynomials, by switching occasionally

FIG. 4.11. *Standard* BiCGstab(2)

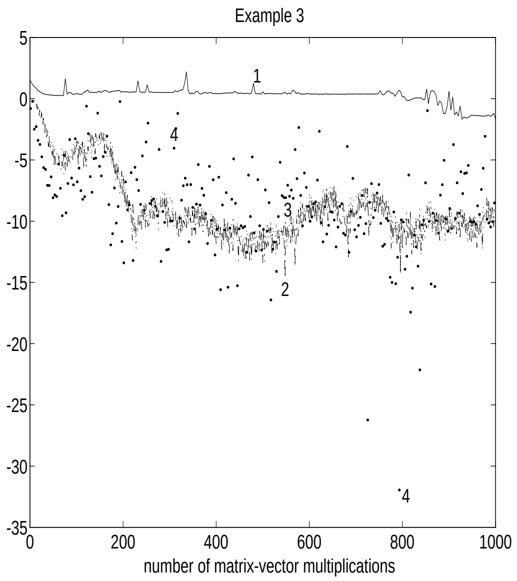
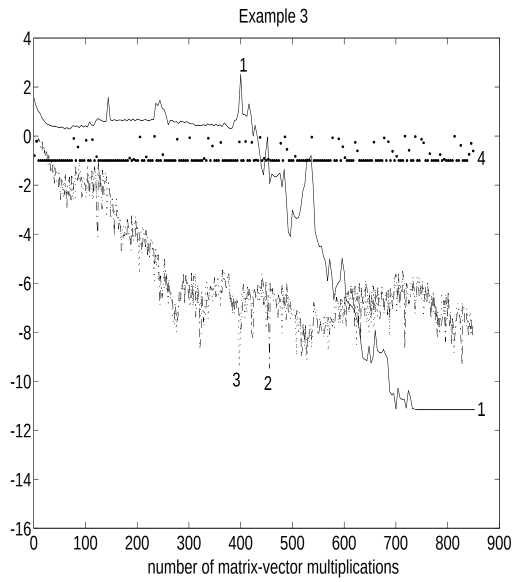


FIG. 4.12. *Modified* $\Omega = 0.7$



between FOM and GMRES processes.

This approach often leads to improve convergence and may help to overcome phases of stagnation. Our strategies are rather inexpensive, relative to the work per matrix-vector product.

Acknowledgement. We thank the referees for their suggestions to improve our presentation.

REFERENCES

[1] Z. BAI, *Error analysis of the Lanczos algorithm for the nonsymmetric eigenvalue problem*, Math. Comp., 62 (1994), pp. 209–226.
[2] R.E. BANK AND T.F. CHAN, *An analysis of the composite step biconjugate gradient method*, Numer. Math., 66 (1993), pp. 295–319.
[3] P.N. BROWN, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 58–78.
[4] S.C. EISENSTAT, H.C. ELMAN AND M.H. SCHULTZ, *Variational iterative methods for nonsymmetric systems of linear equations*, Siam J. Numer. Anal., 20 (1983), pp. 345–357.
[5] T.F. CHAN, E. GALLOPOULOS, V. SIMONCINI, T. SZETO AND C.H. TONG, *A Quasi-Minimal Residual Variant of the Bi-CGSTAB Algorithm for Nonsymmetric Systems*, SIAM J. Sci. Comput., 15 (1994), pp. 338–347.
[6] G.H. GOLUB AND C.F. VAN LOAN, *Matrix Computations*, Second edition, The John Hopkins University Press, Baltimore and London, 1989.
[7] R. FLETCHER, *Conjugate gradient methods for indefinite systems*, In Proc. of the Dundee Biennial Conference on Numerical Analysis, G. Watson, ed., Springer-Verlag, New York, 1975.
[8] R.W. FREUND, *A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems*, SIAM J. Sci. Comput., 14 (1993), pp. 470–482.
[9] R.W. FREUND, M.H. GUTKNECHT AND N. NACHTIGAL, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 137–158.
[10] A. GREENBAUM, *Behavior of slightly perturbed Lanczos and conjugate gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.
[11] M.H. GUTKNECHT, *Variants of BiCGStab for Matrices with Complex Spectrum*, SIAM J. Sci. Comput., 14 (1993), pp. 1020–1033.
[12] C. LANCZOS, *Solution of systems of linear equations by minimized iteration*, J. Res. Nat. Bur. Stand., 49 (1952), pp. 33–53.
[13] U. MEIER YANG, *Preconditioned Conjugate Gradient-Like methods for Nonsymmetric Linear Systems*, Preprint, Center for Research and Development, University of Illinois at Urbana-Champaign, 1992.
[14] J. MODERSITZKI AND G.L.G. SLEIJPEN, *An error analysis of CSBCG*, In preparation
[15] A. NEUMAIER, *Oral presentation at the Oberwolfach meeting "Numerical Linear Algebra"*, Oberwolfach, April 1994.

[16] C. PAIGE, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra Appl., 34 (1980), pp. 235–258.
[17] B.N. PARLETT, D.R. TAYLOR AND Z.A. LIU, *A look-ahead Lanczos algorithm for unsymmetric matrices*, Math. Comp., 44 (1985), pp. 105–124.
[18] Y. SAAD, *Krylov subspace methods for solving large linear unsymmetric linear systems*, Math. Comput., 37 (1981), pp. 105–126.
[19] Y. SAAD AND M.H. SCHULTZ, *GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.
[20] Y. SAAD, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput., 14, (1993), pp. 461–469.
[21] G.L.G. SLEIJPEN AND D.R. FOKKEMA, *BiCGstab(ℓ) for linear equations involving matrices with complex spectrum*, ETNA, 1 (1993), pp. 11–32.
[22] G.L.G. SLEIJPEN AND H.A. VAN DER VORST, *Reliable updated residuals in hybrid Bi-CG methods*, Preprint 886, Dept. Math., University Utrecht (1994).
[23] G.L.G. SLEIJPEN, H.A. VAN DER VORST AND D.R. FOKKEMA, *BiCGstab(ℓ) and other hybrid Bi-CG methods*, Numerical Algorithms, 7 (1994), pp. 75–109.
[24] P. SONNEVELD, *CGS, a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 36–52.
[25] 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 (1992), pp. 631–644.
[26] H.A. VAN DER VORST AND C. VUIK, *The superlinear convergence behaviour of GMRES*, J. Comput. Appl. Math., 48 (1993), pp. 327–341.
[27] D.M. YOUNG AND K.C. JEA, *Generalized conjugate-gradient acceleration of nonsymmetrizable iterative methods*, Linear Algebra Appl., 34 (1980), pp. 159–194.