Title	IDR VERSUS OTHER KRYLOV SUBSPACE SOLVERS (The latest developments in theory and application on scientific computation)
Author(s)	ZEMKE, JENS-PETER M.
Citation	数理解析研究所講究録 (2012), 1791: 37-46
Issue Date	2012-04
URL	http://hdl.handle.net/2433/172841
Right	
Type	Departmental Bulletin Paper
Textversion	publisher

数理解析研究所講究録
第1791巻 2012年 37-46

Proceedings of a Workshop in Kyoto

IDR VERSUS OTHER KRYLOV SUBSPACE SOLVERS*

JENS-PETER M. ZEMKE†

Abstract. We compare members of the IDR family for the solution of linear systems and eigenvalue problems with traditional Krylov subspace solvers. This comparison is based on a description of IDR as a means to construct generalized Hessenberg decompositions, whereas traditional Krylov methods construct Hessenberg decompositions.

Key words. IDR; IDR(s); eigenvalues; Krylov subspace methods.

AMS subject classifications. 65F15 (primary); 65F10; 65F50

1. Introduction. Krylov subspace methods are named after the Russian naval engineer Алексей Николаевич Крылов (Aleksei Nikolaevich Krylov), who in 1931 wrote a paper on a method to compute the coefficients of the characteristic polynomial of a matrix, cf. [7]. In 1940 the first modern Krylov subspace method was developed [5]. The best known Krylov subspace methods are based on Lanczos's [8, 9] and Arnoldi's [1] method. The first IDR method is [28]; the IDR(s) methods [21, 27] are relatively new. The generalization to use larger shadow spaces of dimension $s \in \mathbb{N}$ offers advantages: these appear to be more stable than the original IDR variant, BiCGStab [25, 24], and most of its relatives.

We simply term all methods form the IDR family, e.g., original IDR, BiCGStab, IDR(s) and IDRStab [22, 19, 23] amongst others, as *IDR methods* or *Sonneveld methods*. Sonneveld methods are linked to Lanczos processes termed Lanczos($s, 1$). This process is based on a left block Krylov subspace and a simple right Krylov subspace. Even though this link explains some of the details, it does not account for all the subtleties associated with Sonneveld methods.

1.1. Notation. We use standard notation. The identity matrix of size $n \times n$ is denoted by $\mathbf{I} = \mathbf{I}_n$, its column vectors by $\mathbf{e}_j$ and its elements by the Kronecker delta δ_{ij}. The vector of the sums of all columns, i.e., the vector of all ones, is denoted by $\mathbf{e}$. The matrix $\mathbf{O} = \mathbf{O}_n$ denotes the zero matrix of size $n \times n$, the zero column vector of length n is denoted by $\mathbf{o} = \mathbf{o}_n$. The sizes are omitted if easily deducible from the context. We are interested in the properties of a square matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, e.g., its inverse and/or some of its eigenvalues. Unreduced Hessenberg matrices are denoted by letter $\mathbf{H}_k \in \mathbb{C}^{k\times k}$, upper triangular matrices by letter $\mathbf{U}_k \in \mathbb{C}^{k\times k}$. Extended counterparts of $\mathbf{I}_k$, $\mathbf{H}_k$, and $\mathbf{U}_k$ exist, which are denoted by $\underline{\mathbf{I}}_k$, $\underline{\mathbf{H}}_k$, and $\underline{\mathbf{U}}_k$, respectively. The rectangular matrices $\underline{\mathbf{I}}_k \in \mathbb{C}^{(k+1)\times k}$ and $\underline{\mathbf{U}}_k \in \mathbb{C}^{(k+1)\times k}$ are obtained by appending a row of zeros at the bottom, $\underline{\mathbf{H}}_k$ is an unreduced extended Hessenberg matrix that has $\mathbf{H}_k$ as leading square part. The columns of $\underline{\mathbf{I}}_k$ are denoted by $\underline{\mathbf{e}}_j$. The vector of all ones of length $k+1$ is denoted by $\underline{\mathbf{e}} \in \mathbb{C}^{k+1}$. The inverse, transpose, complex conjugate transpose, and pseudo-inverse (or Moore-Penrose inverse) is denoted by appending $^{-1}$, $^{\mathsf{T}}$, $^{\mathsf{H}}$, and †, respectively.

We remark that like in [4] we use a simplified way to denote Krylov subspace methods, e.g., we write GMRes in place of GMRES as is done in the original publication, since the acronym stands for the phrase Generalized Minimal RESidual.

2. Classical Krylov subspace methods. Essentials of classical Krylov subspace methods can be captured by a so-called *Hessenberg decomposition* [5, 4]

$$\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_{k+1}\underline{\mathbf{H}}_k, \quad k \in \mathbb{N}, \quad k < n, \tag{2.1}$$

*Version of November 30, 2011, 15:21.

†Institut für Numerische Simulation, Technische Universität Hamburg-Harburg, D-21073 Hamburg, Germany (zemke@tu-harburg.de).

where $\mathbf{Q}_{k+1} = (\mathbf{Q}_k, \mathbf{q}_{k+1}) = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{k+1}) \in \mathbb{C}^{n\times(k+1)}$ accounts for the basis vectors $\mathbf{q}_j$, $1 \leqslant j \leqslant k+1$, produced to span the $(k+1)$st Krylov subspace ($\mathbf{q} := \mathbf{q}_1$)

$$\mathcal{K}_{k+1} := \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{q}) := \mathsf{span}\{\mathbf{q}, \mathbf{A}\mathbf{q}, \mathbf{A}^2\mathbf{q}, \ldots, \mathbf{A}^k\mathbf{q}\} = \mathsf{span}\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{k+1}\}, \quad (2.2)$$

and an unreduced extended Hessenberg matrix $\underline{\mathbf{H}}_k \in \mathbb{C}^{(k+1)\times k}$ that in some manner collects information about the action of the operator $\mathbf{A}$ on the Krylov subspace $\mathcal{K}_k(\mathbf{A}, \mathbf{q})$.

These Hessenberg decompositions as well as their perturbed counterparts are related to polynomial approximation theory [30]. The iterates of linear linear system solvers for

$$\mathbf{A}\mathbf{x} = \mathbf{r}_0, \quad \mathbf{A} \in \mathbb{C}^{n\times n}, \mathbf{r}_0 \in \mathbb{C}^n, \quad (2.3)$$

based on the Orthogonal Residual (OR) and the Minimal Residual (MR) approach, respectively, to be defined below, can be thought of as if they are obtained as the evaluation of interpolation polynomials. The OR iterates satisfy

$$\mathbf{x}_k = \mathcal{L}_k[z^{-1}](\mathbf{A})\mathbf{r}_0, \quad (2.4)$$

where $\mathcal{L}_k[z^{-1}]$ is the interpolation of the function $f : z \mapsto z^{-1}$ at the eigenvalues of the trailing Hessenberg matrix $\mathbf{H}_k$, i.e., the Ritz values, counting multiplicity. Similarly, the MR iterates satisfy

$$\underline{\mathbf{x}}_k = \underline{\mathcal{L}}_k[z^{-1}](\mathbf{A})\mathbf{r}_0, \quad (2.5)$$

where $\underline{\mathcal{L}}_k[z^{-1}]$ is the interpolation of the function $f : z \mapsto z^{-1}$ at the the eigenvalues of the matrix $(\underline{\mathbf{H}}_k^\dagger \underline{\mathbf{I}}_k)^{-1}$, i.e., the harmonic Ritz values, counting multiplicity.

The convergence properties of linear system solvers are intimately related to the convergence properties of (harmonic) Ritz values. Optimal methods like Arnoldi [1] and GMRES [15] compute the best polynomial with respect to some approximation problem, i.e., Arnoldi computes a least squares approximation to an annihilating polynomial,

$$\|p_k(\mathbf{A})\mathbf{r}_0\| = \min_{p_k \in \mathbb{P}_{k,k}}, \quad \mathbb{P}_{k,k} := \Big\{p \mid p(z) = z^k + \sum_{i=0}^{k-1} a_i z^i\Big\}, \quad (2.6)$$

and GMRES computes a least squares approximation to an residual polynomial (these are polynomials with trailing coefficient one),

$$\|p_k(\mathbf{A})\mathbf{r}_0\| = \min_{p_k \in \mathbb{P}_{k,0}}, \quad \mathbb{P}_{k,0} := \Big\{p \mid p(z) = 1 + \sum_{i=1}^{k} a_i z^i\Big\}. \quad (2.7)$$

This optimality ensures that the methods terminate with the solution once the degree of the minimal polynomial of the vector $\mathbf{r}_0$ has been reached.

These optimal polynomials can be expressed in terms of the (extended) Hessenberg matrices $\mathbf{H}_k$ and $\underline{\mathbf{H}}_k$. The residual polynomial for the Arnoldi approximation, defined only for regular $\mathbf{H}_k$ (which is used in all OR approximations, e.g., in CG) is given by the polynomial

$$\mathcal{R}_k(z) := \det(\mathbf{I}_k - z\mathbf{H}_k^{-1}), \quad (2.8)$$

the residual polynomial for the GMRES approximation (which is used in all MR approximations) is given by the polynomial

$$\underline{\mathcal{R}}_k(z) := \det(\mathbf{I}_k - z\underline{\mathbf{H}}_k^\dagger \underline{\mathbf{I}}_k). \quad (2.9)$$

The roots of these polynomials, i.e., the eigenvalues of $\mathbf{H}_k$ and the eigenvalues of the inverse of a section of the pseudoinverse are known as the Ritz values and harmonic Ritz values, respectively.

The success of the Lanczos's method and related methods like CG, e.g., [8, 6, 9] shows that methods which are not optimal for general matrices are still of interest. This is due to the fact that these methods are based on coupled *short recurrences*. In exact arithmetic, provided that no breakdown occurs, Lanczos's methods also show the finite termination property. As Lanczos's methods change their behavior in an unexpected manner when executed in finite precision, the convergence analysis has to be adopted to perturbed methods.

2.1. Lanczos's method. Like Hessenberg's method [5], Lanczos's method [8, 9] is based on bi-orthogonality. In contrast to Hessenberg, who uses the standard unit vectors $\mathbf{e}_j$ as left vectors, Lanczos does not use a fixed set of vectors, but instead for this purpose computes a basis of a left Krylov subspace

$$\widehat{\mathcal{K}}_{k+1} := \mathcal{K}_{k+1}(\widehat{\mathbf{A}}, \widehat{\mathbf{q}}) = \mathsf{span}\,\{\widehat{\mathbf{q}}, \widehat{\mathbf{A}}\widehat{\mathbf{q}}, \widehat{\mathbf{A}}^2\widehat{\mathbf{q}}, \ldots, \widehat{\mathbf{A}}^k\widehat{\mathbf{q}}\} = \mathsf{span}\,\{\widehat{\mathbf{q}}_1, \widehat{\mathbf{q}}_2, \ldots, \widehat{\mathbf{q}}_{k+1}\}, \tag{2.10}$$

where $\widehat{\mathbf{A}}$ is the adjoint in some bilinear or sesquilinear form. For ease of presentation, we think of $\widehat{\mathbf{A}}$ being the Hermitean adjoint, i.e., $\widehat{\mathbf{A}} = \mathbf{A}^{\mathsf{H}}$, and the form being the usual inner product in $\mathbb{C}^n$. The method of Lanczos computes (formally) bi-orthogonal bases of $\widehat{\mathcal{K}}_k$ and $\mathcal{K}_k$ via some two-sided Gram-Schmidt process. As all vectors in the Krylov subspace $\mathcal{K}_k$ correspond to polynomials in $\mathbf{A}$,

$$\mathbf{q}_k \in \mathcal{K}_k \quad \Rightarrow \quad \mathbf{q}_k = \sum_{j=1}^{k} \mathbf{A}^{j-1}\mathbf{q}\alpha_j = p_{k-1}(\mathbf{A})\mathbf{q}, \tag{2.11}$$

the products of left vectors $\sum_{j=1}^{k} \widehat{\mathbf{A}}^{j-1}\widehat{\mathbf{q}}\widehat{\alpha}_j$ and right vectors $\sum_{j=1}^{k} \mathbf{A}^{j-1}\mathbf{q}\alpha_j$ can be expressed using solely the so-called moments of $\mathbf{A}$, compare with [8, § V., Eqn. (34) p. 258]:

$$\langle \widehat{\mathbf{A}}^i\widehat{\mathbf{q}}, \mathbf{A}^j\mathbf{q}\rangle = \langle \widehat{\mathbf{q}}, \mathbf{A}^{i+j}\mathbf{q}\rangle = c_{i+j}, \quad 0 \leqslant i, j \leqslant n. \tag{2.12}$$

Lanczos derives a three-term recurrence for the solutions $\boldsymbol{\eta}_k = \left(\eta_{0,k}, \ldots, \eta_{k-1,k}\right)^{\mathsf{T}}$ to the Hankel systems

$$\begin{pmatrix} c_0 & c_1 & \cdots & c_k \\ c_1 & c_2 & \cdots & c_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ c_k & c_{k+1} & \cdots & c_{2k} \end{pmatrix} \begin{pmatrix} \eta_{0,k} \\ \vdots \\ \eta_{k-1,k} \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ h_k \end{pmatrix} \tag{2.13}$$

with a certain $h_k \in \mathbb{C}$ that eventually becomes zero, see [8, § VI., Eqn. (50), p. 262]. This "progressive" form of his algorithm is the basis for his method of "minimized iterations" [8, § VII., pp. 265–268], which is the well-known reduction to tridiagonal form by means of a two-sided Gram-Schmidt process. The quantities constructed in this algorithm can be captured with two Hessenberg decompositions

$$\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_{k+1}\underline{\mathbf{T}}_k, \quad \widehat{\mathbf{A}}\widehat{\mathbf{Q}}_k = \widehat{\mathbf{Q}}_{k+1}\underline{\widehat{\mathbf{T}}}_k, \quad \text{where} \quad \widehat{\mathbf{T}}_k = \mathbf{T}_k^{\mathsf{H}}. \tag{2.14}$$

Here, $\mathbf{T}_k \in \mathbb{C}^{k\times k}$ and $\widehat{\mathbf{T}}_k \in \mathbb{C}^{k\times k}$ denote the leading square parts of the unreduced extended tridiagonal (i.e., Hessenberg) matrices $\underline{\mathbf{T}}_k$ and $\underline{\widehat{\mathbf{T}}}_k$, respectively.

3. Sonneveld methods. Sonneveld methods are based on the IDR Theorem. IDR spaces, a special case of Sonneveld subspaces [19, Definition 2.2, p. 2690], are defined as follows. Define $\mathcal{G}_0$ by

$$\mathcal{G}_0 = \mathcal{K}(\mathbf{A}, \mathbf{q}) = \mathcal{K}_n(\mathbf{A}, \mathbf{q}) = \mathsf{span}\,\{\mathbf{q}, \mathbf{A}\mathbf{q}, \ldots, \mathbf{A}^{n-1}\mathbf{q}\} \subset \mathbb{C}^n. \tag{3.1}$$

In case of non-derogatory $\mathbf{A} \in \mathbb{C}^{n\times n}$ and a generic starting vector $\mathbf{q} \in \mathbb{C}^n$, $\mathcal{G}_0 = \mathbb{C}^n$. IDR Sonneveld spaces $\mathcal{G}_j$ are recursively defined by

$$\mathcal{G}_j = g_j(\mathbf{A})(\mathcal{G}_{j-1} \cap \mathcal{S}), \quad g_j(z) = \eta_j z + \mu_j, \quad \eta_j, \mu_j \in \mathbb{C},\ \eta_j \neq 0, \quad j = 1, 2, \ldots. \tag{3.2}$$

Here, $\mathcal{S}$ is a space of codimension $s \in \mathbb{N}$. The IDR Theorem is given as follows:

THEOREM 3.1 (IDR Theorem [21]). *Under mild conditions on the matrices* $\mathbf{A}$ *and the space* $\mathcal{S}$,

(i) $\mathcal{G}_j \subsetneq \mathcal{G}_{j-1}$ *for all* $\mathcal{G}_{j-1} \neq \{\mathbf{o}_n\}$, $j > 0$.
(ii) $\mathcal{G}_j = \{\mathbf{o}_n\}$ *for some* $j \leqslant n$.

For the proof we refer to [21, 17].

The relation of IDR to Krylov subspaces is given in [3, 16, 4, 17]. The latter includes an alternate description of Sonneveld spaces [17, Theorem 11, p. 1104] based on left block Krylov subspaces. Implementations of the recursion (3.2) are given in [21, 27, 26]:

Initialization: compute $s+1$ basis vectors $\mathbf{g}_i$, $1 \leqslant i \leqslant s+1$, in $\mathcal{K}_{s+1} \subset \mathcal{G}_0$ using your favorite Krylov subspace method. We advocate the use of Arnoldi/GMRES.

Recursion: for $j > 0$ until convergence perform the following:

Intersection: compute a linear combination $\mathbf{v}_i$ of vectors in $\mathcal{G}_{j-1} \cap \mathcal{S}$. Typically, $s+1$ vectors are used in this stage, mostly the newest vectors, as is done in [21, 27, 23, 26], or a fixed set of s vectors for several steps and one that changes, as is done in [19]. Numerically, using *all* vectors available is more robust (but more costly); this was observed in several experiments.

Update: if constructing the first vector in a new space $\mathcal{G}_j$, chose a new linear polynomial g_j of exact degree 1. Here, the remarks on the accuracy of the Lanczos coefficients apply, the techniques from [18] find here applications. Alternatives to minimization include the use of eigenvalue information, either using another Krylov subspace method [16] or the purified Sonneveld pencil of the Sonneveld method [4].

Map: compute the new vector $g_j(\mathbf{A})\mathbf{v}_i$ in $\mathcal{G}_j$ and compute a new basis vector as linear combination of $g_j(\mathbf{A})\mathbf{v}_i$ with other vectors in $\mathcal{G}_j$. The first vector is essentially unique up to scaling; experiments show that computing linear combinations increases the numerical stability significantly.

Sonneveld Krylov subspace methods can be described by a so-called *generalized Hessenberg decomposition* [4]

$$\mathbf{A}\mathbf{V}_k = \mathbf{A}\mathbf{G}_k\mathbf{U}_k = \mathbf{G}_{k+1}\underline{\mathbf{H}}_k, \quad k \in \mathbb{N}, \quad k < n, \tag{3.3}$$

where $\mathbf{V}_k = \mathbf{G}_k\mathbf{U}_k$, with $\mathbf{U}_k \in \mathbb{C}^{k\times k}$ upper triangular, and all other matrices are defined like in Eqn. (2.1).

The small change from the Hessenberg decomposition (2.1) to the *generalized* Hessenberg decomposition (3.3) is the main change in devising new algorithms or applying well-known techniques from the pool of existing Krylov subspace method techniques.

The relation to Lanczos's methods with s left-hand vectors can be used to give a rule of thumb for the convergence of IDR(s) when these left-hand or shadow vectors are chosen at random [20]. We present two sets of plots to highlight some aspects of the statements in [20].

Our first example, Figure 3.1, is based on a random matrix of size 100×100 shifted such that it is positive real. GMRES needs 100 steps to compute the solution and converges in a nice superlinear curve. Such matrices are very nice test examples for numerical experiments involving Krylov subspace methods: the convergence has almost no peaks and the probability of a breakdown is small.

Our second example, Figure 3.2, is based on a Frank matrix of size 100×100. GMRES stagnates more or less in a first phase, and only when almost the full space has been spanned, makes rapid progress. The Frank matrix is an upper Hessenberg matrix with integer entries that has only real eigenvalues that come in reciprocal pairs. The small eigenvalues are ill-conditioned. These ill-conditioned eigenvalues are

not approximated very well by Krylov subspace methods, which typically results in a delay in the convergence. The Frank matrix is a good test example to depict the failure of (short recurrence type) Krylov subspace methods.

In [20], Sonneveld relates GMRES and a random projection Galërkin method, which in turn is then related to a Lanczos's method with s random shadow vectors. We omit the full random Galërkin method, but show how in theory the Lanczos's method with s random shadow vectors relates to GMRES. The two plots in the first rows of Figure 3.1 and Figure 3.2 depict the relation of GMRES to Lanczos$(s,1)$, first with reorthogonalization, then without reorthogonalization. The reorthogonalization procedure does not give reliable results beyond step 100. We implemented the OR approach using the backslash operator in MATLAB for the small resulting block-tridiagonal Hessenberg square matrices, which is by no means the stablest way. Thus, the curves are not reliable once there have been some peaks in the convergence.

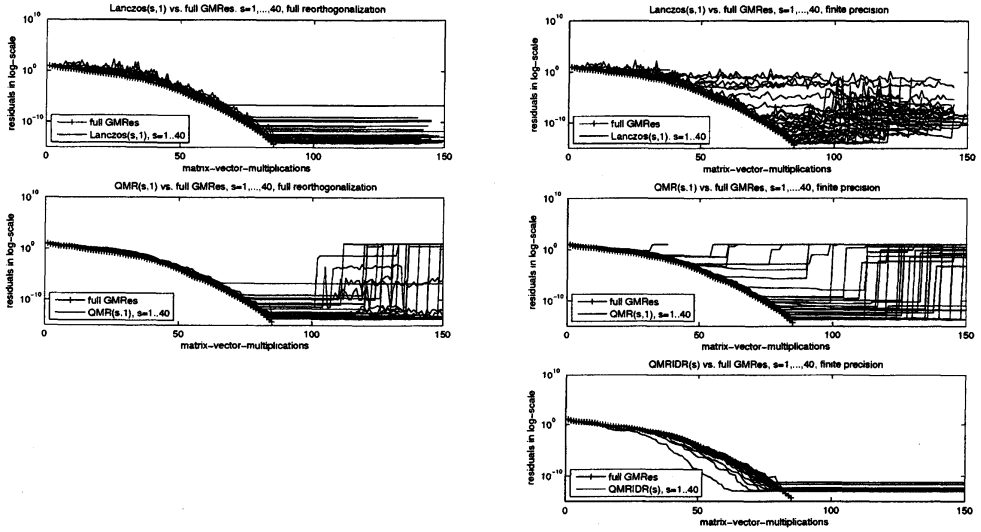

FIG. 3.1. *Lanczos$(s,1)$, QMR$(s,1)$, and QMRIDR(s) versus* GMRES*: Example based on a random matrix. The shadow vectors are the first vectors of a randomly chosen matrix in $\mathbb{R}^{100\times 40}$. The parameter s varies between 1 and 40. The curves of QMRIDR(s) are tuned by omitting every $(s+1)$th step. The* GMRES *convergence curve is indicated by plus signs, the solid lines are the convergence curves of Lanczos$(s,1)$, QMR$(s,1)$, and QMRIDR(s), respectively.*

As GMRES is of type MR and Lanczos$(s,1)$ is of type OR, we switch from Lanczos$(s,1)$ to QMR$(s,1)$, the MR variant of Lanczos$(s,1)$. The resulting plots with and without reorthogonalization are depicted in the second row of Figure 3.1 and Figure 3.2. We see in both examples that the QMR$(s,1)$ curves are even closer to the GMRES curve. It is an interesting future investigation to derive an explicit formula for the residual surplus, comparable to the formula in [20]. We remark that similar to the computation of these curves in the case of Lanczos$(s,1)$, the reorthogonalization procedure does not give reliable results beyond step 100. Also, we implemented the MR approach using the backslash operator in MATLAB for the small resulting block-tridiagonal rectangular Hessenberg matrices, which is by no means the stablest way. Thus, the curves are not reliable once there have been some peaks in the convergence of the corresponding Lanczos$(s,1)$ curves.

The last row of Figure 3.1 and Figure 3.2 only contains a single plot, namely, the convergence of QMRIDR(s) in finite precision versus the convergence curve of GMRES. The reason might be obvious: we do not know of any scheme to mimic the "true" convergence of unperturbed QMRIDR(s), as IDR based methods only implicitly compute left block Krylov subspaces.

There are some observations we state here for future investigation: the "convergence model" [20] by Sonneveld seems to hold true in "nice cases" like the one given

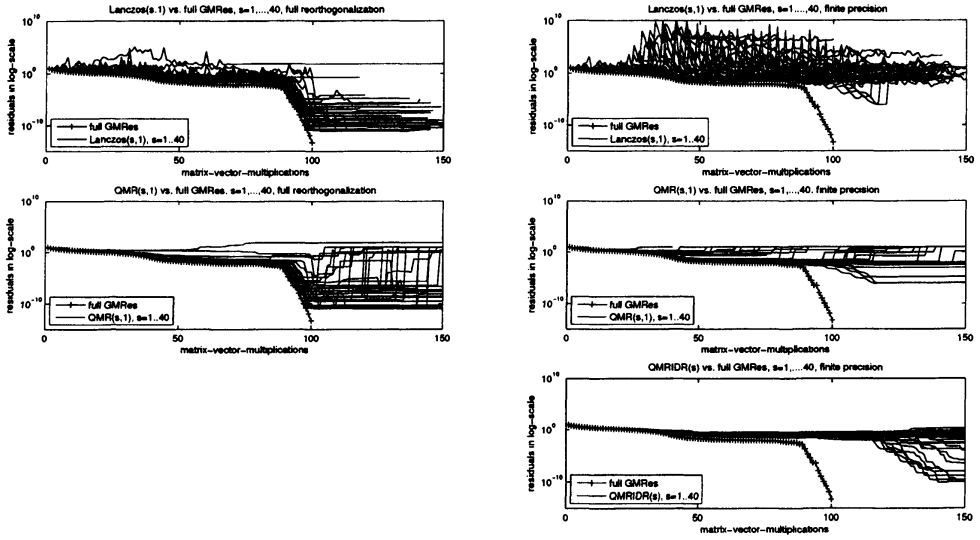

FIG. 3.2. *Lanczos*$(s,1)$, *QMR*$(s,1)$, *and QMRIDR*(s) *versus* GMRES*: Example based on a Frank matrix. The shadow vectors are the first vectors of a randomly chosen matrix in* $\mathbb{R}^{100\times 40}$. *The parameter* s *varies between* 1 *and* 40. *The curves of QMRIDR*(s) *are tuned by omitting every* $(s+1)$*th step. The* GMRES *convergence curve is indicated by plus signs, the solid lines are the convergence curves of Lanczos*$(s,1)$, *QMR*$(s,1)$, *and QMRIDR*(s), *respectively.*

in Figure 3.1 and seems to fail in "bad cases" like the one given in Figure 3.2. The crucial question is, if we can figure out whether we deal with a "nice" or a "bad" case while using the algorithm to compute a solution. The convergence of all methods based on short recurrences is delayed in finite precision with respect to the number of matrix-vector-multiplications. Is it possible to come up with a rule of thumb or even mathematical theory that explains this behavior?

4. Some classical techniques and remarks on results. We very briefly sketch the application of some classical Krylov subspace techniques to Sonneveld methods. The Ritz approach is based on the Sonneveld pencil $(\mathbf{H}_k, \mathbf{U}_k)$ [4]

$$\mathbf{H}_k \mathbf{s}_j = \theta_j \mathbf{U}_k \mathbf{s}_j \tag{4.1}$$

and gives Ritz pairs $(\theta_j, \mathbf{y}_j := \mathbf{V}_k \mathbf{s}_j = \mathbf{G}_k \mathbf{U}_k \mathbf{s}_j)$ [4, 13, 14]. The harmonic Ritz approach [10, 11, 2, 12]

$$\mathbf{I}_k \underline{\mathbf{s}}_j = \underline{\theta}_j \underline{\mathbf{H}}_k^{\dagger} \underline{\mathbf{U}}_k \underline{\mathbf{s}}_j \tag{4.2}$$

gives harmonic Ritz pairs $(\underline{\theta}_j, \underline{\mathbf{y}}_j := \mathbf{V}_k \underline{\mathbf{s}}_j = \mathbf{G}_k \mathbf{U}_k \underline{\mathbf{s}}_j)$. The Orthogonal Residual (OR) approach: The kth OR solution is given by

$$\mathbf{H}_k \mathbf{z}_k := \mathbf{e}_1 \|\mathbf{r}_0\|, \quad \text{e.g., mostly} \quad \mathbf{z}_k := \mathbf{H}_k^{-1} \mathbf{e}_1 \|\mathbf{r}_0\|, \tag{4.3}$$

the kth OR iterate by

$$\mathbf{x}_k := \mathbf{V}_k \mathbf{z}_k = \mathbf{G}_k \mathbf{U}_k \mathbf{z}_k. \tag{4.4}$$

The Minimal Residual (MR) approach: The kth MR solution is given by

$$\rho_k := \left\| \underline{\mathbf{H}}_k \underline{\mathbf{z}}_k - \underline{\mathbf{e}}_1 \|\mathbf{r}_0\| \right\| = \min, \quad \text{i.e.,} \quad \underline{\mathbf{z}}_k := \underline{\mathbf{H}}_k^{\dagger} \underline{\mathbf{e}}_1 \|\mathbf{r}_0\|, \tag{4.5}$$

the kth MR iterate by

$$\underline{\mathbf{x}}_k := \mathbf{V}_k \underline{\mathbf{z}}_k = \mathbf{G}_k \mathbf{U}_k \underline{\mathbf{z}}_k. \tag{4.6}$$

Other flavors like ORTHORES, ORTHOMIN, and ORTHODIR, and techniques like flexible, multi-shift, and inexact variants can now be developed based on (3.3). Methods like (flexible or multi-shift) QMRIDR [26] provide a smooth transition between the methods of Lanczos and Arnoldi and do not rely on the transposed matrix, but are a little less stable.

4.1. Harmonic Ritz using Sonneveld methods. As the eigenvalue computations and the OR and MR approaches applied to Sonneveld methods have been considered by several authors, we give a numerical example of a harmonic Ritz approach using Sonneveld methods. We compare the eigenvalue approximations returned by the harmonic Ritz approach and the Lanczos Ritz approach for IDRSTAB [22, 19, 23].

In the following example we used the matrix `e05r0500` from Matrix Market. The degree of the stabilizing polynomial was $\ell = 3$, the size of the shadow space was $s = 5$. The resulting undeflated Lanczos-IDRSTAB pencil [14] was of size 120×120 (for the Ritz value computations), the resulting extended Lanczos-IDRSTAB pencil was of size 121×120 (for the harmonic Ritz value computations).

In the first plot, Figure 4.1, the extended Lanczos-IDRSTAB pencil is depicted. Observe that the matrix $\mathbf{U}$ of the Lanczos-IDRSTAB pencil is singular, as every 6th $(s+1=6)$ diagonal element is zero.

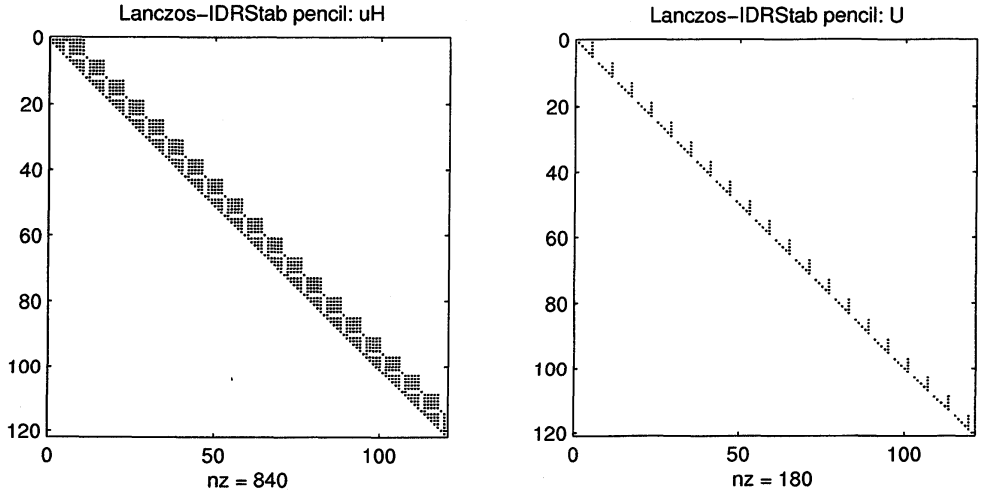

FIG. 4.1. *The resulting Lanczos-*IDRSTAB *pencil of size* 121×120 *for* IDRSTAB *applied with* $\ell = 3$ *and* $s = 5$ *to the matrix* `e05r0500` *from Matrix Market with the starting vector* $\mathbf{r}_0$ *given as the residual of a random initial guess for the linear system with the right-hand side* `e05r0500_rhs1`.

In the second plot, Figure 4.2, the Ritz values and the harmonic Ritz values computed using the Lanczos-IDRSTAB pencil and the extended Lanczos-IDRSTAB pencil are depicted. We observe what is observed frequently for the harmonic Ritz values: they approximate the same eigenvalues as the Ritz values with a slower rate of convergence.

It is well known that it is better to use the Rayleigh quotients of the corresponding harmonic Ritz vectors $\underline{\mathbf{y}}_j$ as approximate eigenvalues, as the harmonic Ritz *vectors* are frequently better approximations to eigenvectors to inner eigenvalues than the classical Ritz vectors $\mathbf{y}_j$. In classical Krylov subspace methods this poses no problem, in Sonneveld methods this technique is not that easily applied.

There are basically two choices, which require future investigation:

- We can use the Sonneveld pencil of IDR(s) or QMRIDR(s) to compute harmonic Ritz pairs, select some that correspond to eigenvalues of $\mathbf{A}$, and compute the Rayleigh quotients of these harmonic Ritz vectors as new eigenvalue approximations. In case of IDRSTAB we have to use the full IDRSTAB pencil [14]. This latter pencil has many zero and infinite eigenvalues which may cause numerical instabilities.
- We can strive for a way to compute harmonic Ritz vectors based on the eigenvectors of a purified and/or deflated pencil. This is sketched for classical Ritz vectors in [4] for IDR(s) and extends to QMRIDR(s). The transition of this technique to IDRSTAB is not straightforward, even in the case of

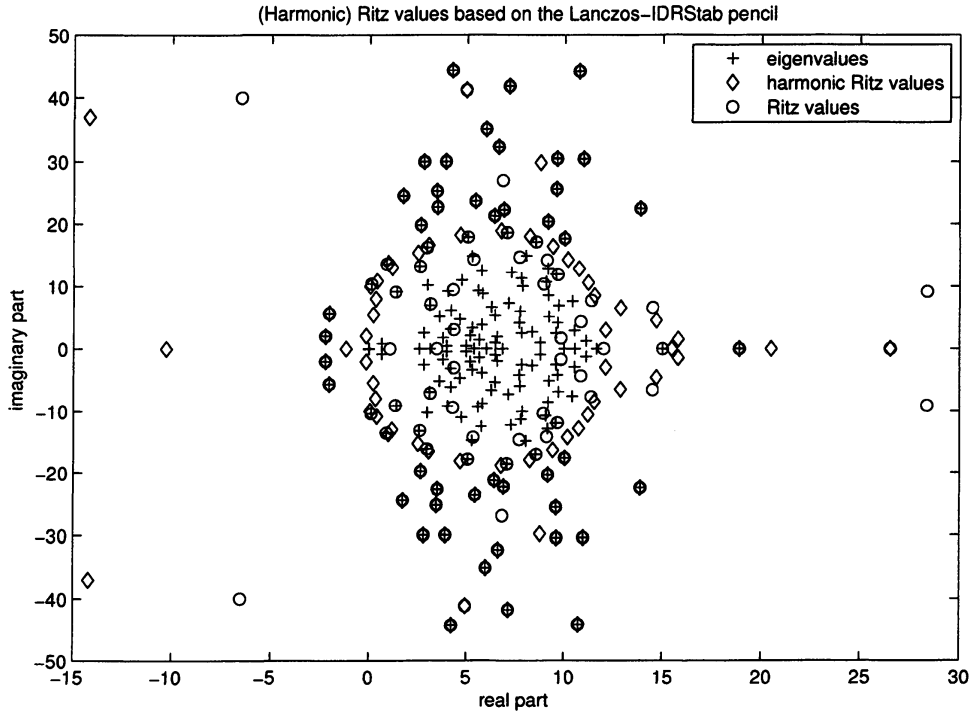

FIG. 4.2. *The Ritz values and the harmonic Ritz values obtained using the (undeflated) Lanczos-*IDRSTAB *pencil of size* 121 × 120 *depicted in Figure 4.1. Cross signs denote eigenvalues of the matrix e05r0500 from Matrix Market, diamonds denote the harmonic Ritz values and circles denote the classical Ritz values. The outer eigenvalues are approximated very well by both approaches, three ghost eigenvalues close to roots of some stabilizing polynomial (which are not depicted) have developed (one real value close to 26, two complex conjugate pairs with real parts close to 5 and imaginary part with absolute value close to 40).*

classical Ritz vectors. The extension to *harmonic* Ritz vectors might even be intractable and requires a new theoretical foundation along the lines of [29].

The former approach has the drawback that it might be impossible to find any good approximations to eigenvalues of $\mathbf{A}$ among the harmonic Ritz values, as these are combinations of a Lanczos$(s, 1)$ process and user-supplied values, e.g., the roots of the stabilizing polynomials. As these informations are "smeared out" in the computation of the harmonic Ritz values, they may suffer great damage and this might render them useless. This depends very strongly on the stabilizing polynomials used, especially when ghost eigenvalues already appeared, see Figure 4.3 obtained using for $\ell = 1$. If we would use the full IDRSTAB pencil or a Sonneveld pencil, the outliers would destroy the good approximation properties of the other harmonic Ritz values close to eigenvalues. We remark that in the generic case the use of a full IDRSTAB pencil or a Sonneveld pencil in eigenvalue computations is harmful.

The latter approach ensures that only the information on the operator $\mathbf{A}$ is used. The harmonic Ritz values are very reliable. As we do need the harmonic Ritz vectors, i.e., the prolonged eigenvectors, to compute the better Rayleigh quotients, we need either a basis for the prolongation, which is usually not at hand (apart from the first steps of IDRSTAB before the first stabilizing polynomial is computed), or we have to compute in a numerically stable manner some "harmonic Ritz vector" of the Sonneveld pencil [4] or the full IDRSTAB pencil [14]. The crucial point is that the classical Ritz values other than the user-supplied polynomial roots survive the process of purification and deflation described in [4, 14] and we do not know what happens to the harmonic Ritz values in these steps.

We could try to recompute the basis vectors that we need for the harmonic Ritz vectors by using the deflated pencils. This seems to work sometimes, and the exact

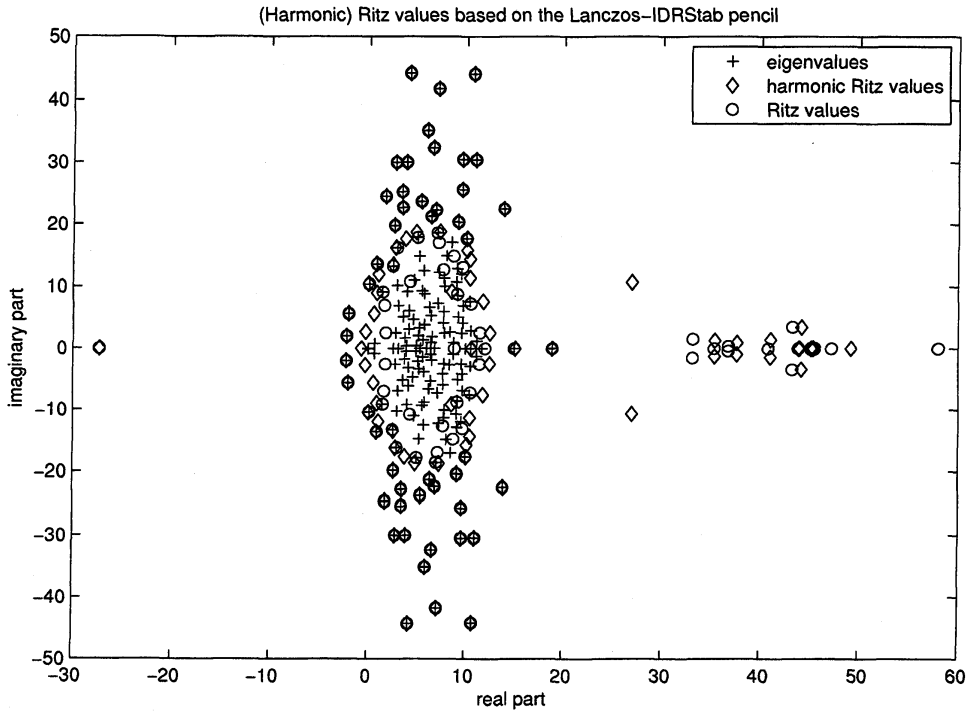

FIG. 4.3. *The Ritz values and the harmonic Ritz values obtained using a (undeflated) Lanczos-*IDRSTAB *pencil of size* 121 × 120 *as depicted in Figure 4.1. Cross signs denote eigenvalues of the matrix, diamonds denote the harmonic Ritz values and circles denote the classical Ritz values.*

circumstances under which this works out fine still have to be investigated.

REFERENCES

[1] W. E. ARNOLDI, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] ROLAND W. FREUND, *Quasi-kernel polynomials and their use in non-Hermitian matrix iterations*, Journal of Computational and Applied Mathematics, 43 (1992), pp. 135–158. Orthogonal polynomials and numerical methods.

[3] MARTIN H. GUTKNECHT, *IDR explained*, Electron. Trans. Numer. Anal., 36 (2009/10), pp. 126–148.

[4] MARTIN H. GUTKNECHT AND JENS-PETER M. ZEMKE, *Eigenvalue computations based on IDR*, Bericht 145, TUHH, Institute of Numerical Simulation, May 2010. Online available at `http://doku.b.tu-harburg.de/volltexte/2010/875/`.

[5] KARL HESSENBERG, *Behandlung linearer Eigenwertaufgaben mit Hilfe der Hamilton-Cayleyschen Gleichung*, Numerische Verfahren, Bericht 1, Institut für Praktische Mathematik (IPM), Technische Hochschule Darmstadt, July 1940. Scanned report and biographical sketch of Karl Hessenberg's life online available at `http://www.hessenberg.de/karl1.html`.

[6] MAGNUS R. HESTENES AND EDUARD STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436 (1953).

[7] Алексей Николаевич Крылов, *О численном решении уравнения, которым в технических вопросах определяются частоты малых колебаний материальных систем*, Известия Академии Наук СССР. Отделение математических и естественных наук. Ser. VII, 4 (1931), pp. 491–539. (Russian).

[8] CORNELIUS LANCZOS, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Research Nat. Bur. Standards, 45 (1950), pp. 255–282.

[9] ———, *Solution of systems of linear equations by minimized iterations*, J. Research Nat. Bur. Standards, 49 (1952), pp. 33–53.

[10] N. J. LEHMANN, *Optimale Eigenwerteinschließungen*, Numerische Mathematik, 5 (1963), pp. 246–272.

[11] RONALD B. MORGAN, *Computing interior eigenvalues of large matrices*, Linear Algebra and Its Applications, 154–156 (1991), pp. 289–309.

[12] C. C. PAIGE, BERESFORD N. PARLETT, AND HENK A. VAN DER VORST, *Approximate solutions and eigenvalue bounds from Krylov subspaces*, Numerical Linear Algebra with Applications,

2 (1995), pp. 115–133.

[13] OLAF RENDEL, *Aspects of eigenvalue computations using Induced Dimension Reduction (IDR)*, bachelor's thesis, Institut für Numerische Simulation, Technische Universität Hamburg-Harburg, 2010.

[14] ANISA RIZVANOLLI, *Eigenwertberechnung mittels IDRStab*, Studienarbeit, Institut für Numerische Simulation, Technische Universität Hamburg-Harburg, 2011.

[15] YOUCEF SAAD AND MARTIN H. SCHULTZ, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[16] VALERIA SIMONCINI AND DANIEL B. SZYLD, *Interpreting IDR as a Petrov-Galerkin method*, SIAM J. Sci. Comput., 32 (2010), pp. 1898–1912.

[17] GERARD L.G. SLEIJPEN, PETER SONNEVELD, AND MARTIN B. VAN GIJZEN, *Bi-CGSTAB as an induced dimension reduction method*, Appl. Numer. Math., 60 (2010), pp. 1100–1114.

[18] GERARD L.G. SLEIJPEN AND HENK A. VAN DER VORST, *Maintaining convergence properties of BiCGstab methods in finite precision arithmetic*, Numer. Algorithms, 10 (1995), pp. 203–223.

[19] GERARD L. G. SLEIJPEN AND MARTIN B. VAN GIJZEN, *Exploiting BiCGstab(ℓ) strategies to induce dimension reduction*, SIAM J. Sci. Comput., 32 (2010), pp. 2687–2709. Received Mar. 11, 2009, electr. publ. Aug. 31, 2010.

[20] PETER SONNEVELD, *On the convergence behaviour of IDR(s)*, Technical Report 10-08, Department of Applied Mathematical Analysis, Delft University of Technology, Delft, 2010.

[21] PETER SONNEVELD AND MARTIN B. VAN GIJZEN, *IDR(s): a family of simple and fast algorithms for solving large nonsymmetric systems of linear equations*, SIAM J. Sci. Comput., 31 (2008/09), pp. 1035–1062.

[22] MASAAKI TANIO AND MASAAKI SUGIHARA, *GIDR(s,L): generalized IDR(s)*, in The 2008 annual conference of the Japan Society for Industrial and Applied Mathematics, Chiba, Japan, September 2008, pp. 411–412. (Japanese).

[23] ———, *GBi-CGSTAB(s, L): IDR(s) with higher-order stabilization polynomials*, J. Comput. Appl. Math., 235 (2010), pp. 765–784. Received May 13, 2009, electr. publ. Jul. 16, 2010.

[24] 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. Statist. Comput., 13 (1992), pp. 631–644.

[25] H. A. VAN DER VORST AND P. SONNEVELD, *CGSTAB, a more smoothly converging variant of CG-S*, Report 90-50, Department of Mathematics and Informatics, Delft University of Technology, 1990.

[26] MARTIN B. VAN GIJZEN, GERARD L.G. SLEIJPEN, AND JENS-PETER M. ZEMKE, *Flexible and multi-shift induced dimension reduction algorithms for solving large sparse linear systems*, Bericht 156, TUHH, Institute of Numerical Simulation, August 2011. Online available at `http://doku.b.tu-harburg.de/volltexte/2011/1114/`.

[27] MARTIN B. VAN GIJZEN AND PETER SONNEVELD, *An elegant IDR(s) variant that efficiently exploits bi-orthogonality properties.*, Technical Report 10-16, Department of Applied Mathematical Analysis, Delft University of Technology, Delft, 2010. (revised version of report 08-21).

[28] P. WESSELING AND P. SONNEVELD, *Numerical experiments with a multiple grid and a preconditioned Lanczos type method*, in Approximation methods for Navier-Stokes problems (Proc. Sympos., Univ. Paderborn, Paderborn, 1979), R. Rautmann, ed., vol. 771 of Lecture Notes in Math., Berlin, Heidelberg, New York, 1980, Springer-Verlag, pp. 543–562.

[29] JENS-PETER M. ZEMKE, *Hessenberg eigenvalue–eigenmatrix relations*, Linear Algebra Appl., 414 (2006), pp. 589–606. Received Nov. 9, 2004, electr. publ. Feb. 17, 2006.

[30] ———, *Abstract perturbed Krylov methods*, Linear Algebra Appl., 424 (2007), pp. 405–434. Received Nov. 3, 2005, electr. publ. Feb. 21, 2007.