Universiteit Utrecht

Department of Mathematics

Iterative linear system solvers with approximate matrix-vector products

by

Jasper van den Eshof, Gerard L.G. Sleijpen, and Martin B. van Gijzen

Preprint

nr. 1293

December, 2003

ITERATIVE LINEAR SYSTEM SOLVERS WITH APPROXIMATE MATRIX-VECTOR PRODUCTS

JASPER VAN DEN ESHOF*, GERARD L.G. SLEIJPEN†, AND MARTIN B. VAN GIJZEN‡

Abstract

There are classes of linear problems for which a matrix-vector product is a time consuming operation because an expensive approximation method is required to compute it to a given accuracy. One important example is simulations in lattice QCD with Neuberger fermions where a matrix multiply requires the product of the matrix sign function of a large sparse matrix times a vector. The recent interest in these type of applications has resulted in research efforts to study the effect of errors in the matrix-vector products on iterative linear system solvers. In this paper we give a very general and abstract discussion of this issue and try to provide insight into why some iterative system solvers are more sensitive than others.

1 Introduction

The central problem in this report is to find an approximate solution to the equation

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

For some linear problems the matrix-vector product can be an expensive operation since a time consuming approximation must be constructed for the product, as for example in simulations in lattice QCD with Neuberger fermions. The recent interest in this, and other applications, has resulted in research efforts to study the impact of errors in the matrix-vector products on iterative linear system solvers, e.g., [3, 4, 6, 8]. The purpose of this paper is to give general and abstract novel insight into why some iterative system solvers are more sensitive than others. This understanding is, for example, important to devise efficient strategies for controlling the errors and, moreover, to choose a suitable iterative solver for a problem. Therefore, we conclude this paper by discussing briefly how this insight can be used to derive strategies for controlling the error. Experiments with these strategies and additional ideas that can be exploited in simulations in lattice QCD with Neuberger fermions are discussed in the forthcoming paper [1].

2 Krylov subspace methods

An important class of iterative solvers for linear systems is the class of *Krylov subspace solvers*. A Krylov subspace method is characterized by the fact that it is an iterative method that constructs its approximate iterate in step j, $\mathbf{x}_j$, from the j dimensional *Krylov subspace*, $\mathcal{K}_j$, defined as the span of $\{\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{j-1}\mathbf{b}\}$. There are various ways of constructing these iterates. Of particular importance are Krylov subspace method

*Department of Mathematics, University of Düsseldorf, Universitätsstr. 1, D-40224, Düsseldorf, Germany. eshof@am.uni-duesseldorf.de

†Department of Mathematics, Utrecht University, P.O. Box 80.010, NL-3508 TA Utrecht, The Netherlands. sleijpen@math.uu.nl

‡ CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse Cedex 01, France. vangijzen@cerfacs.fr

that construct their iterates in an optimal way. Two strategies that are considered in this paper are

1. *Galerkin extraction*: where $\mathbf{x}_j = \mathbf{x}_j^{\mathrm{GAL}} \in \mathcal{K}_j$ such that $\mathbf{r}_j^{\mathrm{GAL}} = \mathbf{r}_j := \mathbf{b} - \mathbf{A}\mathbf{x}_j^{\mathrm{GAL}} \perp \mathcal{K}_j$

2. *Minimal residual extraction*: where $\mathbf{x}_j = \mathbf{x}_j^{\mathrm{MR}} \in \mathcal{K}_j$ and $\mathbf{r}_j^{\mathrm{MR}} = \mathbf{r}_j := \mathbf{b} - \mathbf{A}\mathbf{x}_j$ and $\|\mathbf{b} - \mathbf{A}\mathbf{x}_j\|_2$ is minimal.

The observations we make are not difficult to extend to other extraction methods. Methods like Bi-CGstab [9], however, do not straightforwardly fit into the framework of this paper.

In Krylov subspace methods the Krylov subspace is (implicitly) expanded by applying the matrix to some vector $\mathbf{z}_j$ in step $j+1$. The vectors $\mathbf{z}_j$ for $j = 0, \ldots, k-1$ necessarily form a basis for the Krylov subspace. (Notice that in this paper we assume that the starting vector of the iterative methods is the zero vector.) We try to provide insight into the influence of (allowed) errors in the matrix-vector multiplications on the iterative method. It will therefore come as no surprise that the particular choice of the $\mathbf{z}_j$ will play an important role in the remainder of this paper. Other quantities of interests are the *iterates*, $\mathbf{x}_j$, and the *residuals*, $\mathbf{r}_j := \mathbf{b} - \mathbf{A}\mathbf{x}_j$. We will assume that the following relation links together the quantities of interest after k iteration steps:

$$\mathbf{A}\mathbf{Z}_k = \mathbf{R}_{k+1}\underline{S}_k \quad \text{and} \quad \mathbf{x}_k = \mathbf{Z}_k S_k^{-1} e_1, \tag{2}$$

with $\underline{S}_k$ being a $(k+1) \times k$ upper Hessenberg matrix and S_k the $k \times k$ upper block of $\underline{S}_k$. The definition of S_k depends on the method used but we stress that the recursions described by (2) do not have to be *explicitly* used by the particular method. More specifically, the basis used in the extraction of $\mathbf{x}_j$ from the Krylov subspace may differ from the basis $\mathbf{z}_j, \ldots, \mathbf{z}_{k-1}$ used for the expansion. Throughout this paper capitals are used to group together vectors which are denoted with lower case characters with a subscript that refers to the index of the column (starting with zero for the first column). Hence, $\mathbf{R}_k e_{j+1} = \mathbf{r}_j$.

3 Approximate matrix-vector products

To model perturbations on the exact matrix-vector products we assume that the matrix-vector products are computed by the function $\mathcal{A}_\eta(\mathbf{v})$ that represents approximations to the matrix-vector product $\mathbf{A}\mathbf{v}$ with a relative precision η as

$$\mathcal{A}_\eta(\mathbf{v}) = \mathbf{A}\mathbf{v} + \mathbf{f} \quad \text{with} \quad \|\mathbf{f}\|_2 \leq \eta \|\mathbf{A}\|_2 \|\mathbf{v}\|_2.$$

The precise source of these perturbations can be various and are at this point not of interest. We neglect other errors.

In case the matrix-vector product is computed to some relative precision η_j in step $j+1$, we assume that (2) becomes

$$\mathbf{A}\mathbf{Z}_k + \mathbf{F}_k = \mathbf{R}_{k+1}\underline{S}_k \quad \text{and} \quad \mathbf{x}_k = \mathbf{Z}_k S_k^{-1} e_1. \tag{3}$$

The vector $\mathbf{f}_j$ is the $j+1$-th column of $\mathbf{F}_k$ and it contains the error in the matrix-vector product in step $j+1$ and we, therefore, have that $\|\mathbf{f}_j\|_2 \leq \eta_j \|\mathbf{A}\|_2 \|\mathbf{z}_j\|_2$. It can be easily checked that this assumption is appropriate for all inexact Krylov methods that we consider in this paper. (Notice that we assume that there are no roundoff errors.) The perturbation $\mathbf{F}_k$ in (3) causes that $\mathbf{r}_k$ is not a residual for the vector $\mathbf{x}_k$ defined by the second relation. As a consequence one should be careful when assessing the accuracy of the iterate $\mathbf{x}_k$. Instead we have the following inequality involving the norm of the *residual gap*, that is the distance between the *true residual*, $\mathbf{b} - \mathbf{A}\mathbf{x}_k$, and the *computed residual*, $\mathbf{r}_k$:

$$\|\underbrace{\mathbf{b} - \mathbf{A}\mathbf{x}_k}_{\text{true residual}}\|_2 \quad \leq \quad \|\underbrace{\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)}_{\text{residual gap}}\|_2 \quad + \quad \|\underbrace{\mathbf{r}_k}_{\text{computed residual}}\|_2.$$

We notice that the size of the true residual is unknown in contrast to the size of the computed residual and it follows from (3) that the gap is bounded by

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 = \|\mathbf{F}_k S_k^{-1} e_1\|_2 \leq \sum_{j=0}^{k-1} \eta_j \|\mathbf{A}\|_2 \|\mathbf{z}_j\|_2 |e_{j+1}^* S_k^{-1} e_1|. \tag{4}$$

Focusing on the residual gap is not uncommon in theoretical analyses of the attainable accuracy of iterative methods in the finite precision context, see e.g., [7]. It is based on the frequent observation that the computed residuals eventually become many orders of magnitude smaller than machine precision and, therefore, the attainable precision is determined by the size of the residual gap. A similar technique can be used for getting insight into the effect of approximate matrix-vector products on Krylov methods: if we terminate as soon as $\|\mathbf{r}_k\|_2$ is of order ϵ, then the size of the gap determines the precision of the inexact process. If we would plot the convergence curve of the true residuals then from some point on the residual reduction stagnates and this stagnation level is determined by the size of the residual gap. Extensive numerical experiments confirm this observation. With this assumption we see that the sensitivity of the stagnation level of a particular Krylov subspace method is determined by the quantities $\|\mathbf{z}_j\|_2 |e_{j+1}^* S_k^{-1} e_1|$. The purpose is to investigate the size of these quantities to give some understanding of the sources that influence the sensitivity of Krylov subspace methods.

Some preliminary insight can be given. We have that

$$\underbrace{\mathbf{x}_k}_{\text{extraction}} \quad = \quad \underbrace{\mathbf{Z}_k}_{\text{choice basis}} \quad S_k^{-1} e_1.$$

This shows that the size of the elements of this vector do not only depend on the optimality properties of the iterates (i.e., how $\mathbf{x}_k$ is chosen from $\mathcal{K}_k$) but also on the choice of the basis given by the $\mathbf{z}_j$. On termination, when $\mathbf{x}_k \approx \mathbf{x}$, we expect no essential difference between a Galerkin extraction and minimal residual extraction. On the other hand, linear dependence in the matrix $\mathbf{Z}_k$ implies that elements of the vectors $|S_k^{-1} e_1|$ can be relatively large, which results in a large sensitivity of the particular method to inexact matrix-vector products. The choice of the expansion basis, rather than the extraction method, determines the sensitivity to errors in the matrix-vector products. In the following we will make this more precise.

3.1 The general case

We study the size of the elements of $\|\mathbf{z}_j\|_2 |e_{j+1}^* S_k^{-1} e_1|$ by assuming *exact* matrix-vector products for the moment, i.e., (2) holds. This problem was studied in related formulation in [6, 8]. We first have to introduce some notation. Let $\mathbf{M}$ and $\mathbf{N}$ be Hermitian, positive definite, n dimensional matrices. We define

$$\delta_{\mathbf{M}\to\mathbf{N}} \equiv \max_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{y}\|_{\mathbf{M}}}{\|\mathbf{y}\|_{\mathbf{N}}}$$

which gives the following norm equivalence

$$(\delta_{\mathbf{N}\to\mathbf{M}})^{-1}\|\mathbf{y}\|_{\mathbf{N}} \leq \|\mathbf{y}\|_{\mathbf{M}} \leq \delta_{\mathbf{M}\to\mathbf{N}}\|\mathbf{y}\|_{\mathbf{N}}. \tag{5}$$

We furthermore define the inner product $<\mathbf{z},\mathbf{y}>_{\mathbf{M}}\equiv \mathbf{z}^*\mathbf{M}\mathbf{y}$ and assume that $\mathbf{Z}_k$ is an $\mathbf{M}$-orthogonal basis. Now we have for all $\tilde{\mathbf{x}}_j \in \mathcal{K}_j$

$$\begin{aligned} |e_{j+1}^* S_k^{-1} e_1| \|\mathbf{z}_j\|_{\mathbf{M}}^2 &= |<\mathbf{z}_j, \mathbf{x}_k>_{\mathbf{M}}| \\ &= |<\mathbf{z}_j, \mathbf{x}_k - \tilde{\mathbf{x}}_j>_{\mathbf{M}}| \leq \|\mathbf{x}_k - \tilde{\mathbf{x}}_j\|_{\mathbf{M}} \|\mathbf{z}_j\|_{\mathbf{M}}. \end{aligned} \tag{6}$$

Here we have made use of the fact that $<\mathbf{z}_j, \tilde{\mathbf{x}}_j>_{\mathbf{M}}= 0$. Recall that $\mathbf{x}_j^{\mathrm{MR}}$ is defined as the approximation from the space $\mathcal{K}_j$ that minimizes the error in $\mathbf{A}^*\mathbf{A}$-norm, or, equivalently, minimizes the 2-norm of the residual $\mathbf{r}_j^{\mathrm{MR}} = \mathbf{b} - \mathbf{A}\mathbf{x}_j^{\mathrm{MR}}$. With this definition and (6), we get the bound

$$\|\mathbf{A}\|_2\|\mathbf{z}_j\|_2|e_{j+1}^* S_k^{-1} e_1| \leq \|\mathbf{A}\|_2 \frac{\|\mathbf{z}_j\|_2}{\|\mathbf{z}_j\|_{\mathbf{M}}} \left(\|\mathbf{x} - \mathbf{x}_j^{\mathrm{MR}}\|_{\mathbf{M}} + \|\mathbf{x} - \mathbf{x}_k\|_{\mathbf{M}}\right) \tag{7}$$

$$\leq \|\mathbf{A}\|_2\, \delta_{\mathbf{I}\to\mathbf{M}}\, \delta_{\mathbf{M}\to\mathbf{A}^*\mathbf{A}} \left(\|\mathbf{r}_j^{\mathrm{MR}}\|_2 + \|\mathbf{r}_k\|_2\right). \tag{8}$$

This simple argument, combined with the bound on the residual gap (4), suggests that, if the inexact Krylov subspace method is terminated as soon as $\|\mathbf{r}_k\|_2 \leq \epsilon$, then the size of the residual gap is essentially bounded by a constant times the norm of the residuals corresponding to a minimal residual extraction. Notice that these residuals form a monotonically decreasing sequence and are therefore bounded.

If, for some $\mathbf{N}$, the particular Krylov subspace methods minimizes the error in $\mathbf{N}$-inner product, then we can even remove the $\|\mathbf{r}_k^{\mathrm{MR}}\|_2$ term in (8). In this case we have that $\|\mathbf{x} - \tilde{\mathbf{x}}_j\|_{\mathbf{N}}^2 = \|\mathbf{x}_k - \tilde{\mathbf{x}}_j\|_{\mathbf{N}}^2 + \|\mathbf{x}_k - \mathbf{x}\|_{\mathbf{N}}^2$. Using this we get

$$\|\mathbf{x}_k - \tilde{\mathbf{x}}_j\|_{\mathbf{M}} \leq \delta_{\mathbf{M}\to\mathbf{N}}\|\mathbf{x}_k - \tilde{\mathbf{x}}_j\|_{\mathbf{N}} \leq \delta_{\mathbf{M}\to\mathbf{N}}\|\mathbf{x} - \tilde{\mathbf{x}}_j\|_{\mathbf{N}} \leq \delta_{\mathbf{M}\to\mathbf{N}}\, \delta_{\mathbf{N}\to\mathbf{A}^*\mathbf{A}}\|\mathbf{r}_j^{\mathrm{MR}}\|_2,$$

which leads to the bound

$$\|\mathbf{A}\|_2\|\mathbf{z}_j\|_2|e_{j+1}^* S_k^{-1} e_1| \leq \|\mathbf{A}\|_2\, \delta_{\mathbf{I}\to\mathbf{M}}\, \delta_{\mathbf{M}\to\mathbf{N}}\, \delta_{\mathbf{N}\to\mathbf{A}^*\mathbf{A}}\|\mathbf{r}_j^{\mathrm{MR}}\|_2. \tag{9}$$

For several well-known Krylov subspace methods we have summarized the relevant quantities in Table 1. Substituting these values into (9) finally shows, for all methods mentioned in the table, that

$$\|\mathbf{A}\|_2\|\mathbf{z}_j\|_2|e_{j+1}^* S_k^{-1} e_1| \leq \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2\|\mathbf{r}_j^{\mathrm{MR}}\|_2. \tag{10}$$

From our discussion it is clear that the optimality properties of the iterates can simplify the bound (8) somewhat. Since we terminate as soon as $\|\mathbf{r}_k\|_2 \leq \epsilon$, it follows that the impact of the choice of the optimality properties (i.e., the $\mathbf{N}$-inner product) for the iterates is small. (However, it can be large during some iteration steps of the iterative process.) A more important factor in the sensitivity for approximate matrix-vector products is the conditioning of the basis $\mathbf{z}_0,\ldots,\mathbf{z}_{k-1}$ which is determined by the choice of the matrix $\mathbf{M}$. For example, if $\mathbf{M} = \mathbf{A}$ and $\mathbf{A}$ is indefinite then the basis can be almost linear dependent. See [8] for a different point of view, analysis and examples. We will focus on this in the next section

Table 1: Values for various Krylov subspace methods assuming that $\mathbf{M} = \mathbf{M}^* > 0$ and $\mathbf{N} = \mathbf{N}^* > 0$.

Example method	$\mathbf{M}$	$\mathbf{N}$	$\delta_{\mathbf{I}\to\mathbf{M}}$	$\delta_{\mathbf{M}\to\mathbf{N}}$	$\delta_{\mathbf{N}\to\mathbf{A}^*\mathbf{A}}$
ORTHORES	$\mathbf{I}$	$\mathbf{A}$	1	$\sqrt{\|\mathbf{A}^{-1}\|_2}$	$\sqrt{\|\mathbf{A}^{-1}\|_2}$
GMRES	$\mathbf{I}$	$\mathbf{A}^*\mathbf{A}$	1	$\|\mathbf{A}^{-1}\|_2$	1
CG	$\mathbf{A}$	$\mathbf{A}$	$\sqrt{\|\mathbf{A}^{-1}\|_2}$	1	$\sqrt{\|\mathbf{A}^{-1}\|_2}$
GCR	$\mathbf{A}$	$\mathbf{A}^*\mathbf{A}$	$\sqrt{\|\mathbf{A}^{-1}\|_2}$	$\sqrt{\|\mathbf{A}^{-1}\|_2}$	1

3.2 The matrix $\mathbf{Z}_k$ is $\mathbf{A}$-orthogonal

In this section we consider the situation that $\mathbf{z}_i^*\mathbf{A}\mathbf{z}_j = 0$ for $i < j$. Or in other words, $\mathbf{Z}_k^*\mathbf{A}\mathbf{Z}_k$ is upper triangular which reduces to a diagonal matrix in case $\mathbf{A}$ is Hermitian. If the matrix $\mathbf{A}$ is Hermitian positive definite then the vectors $\mathbf{z}_j$ form an orthogonal basis with respect to the $\mathbf{A}$-inner product and, therefore, $\mathbf{Z}_k$ is orthogonal with respect to a well-defined inner product. Consequently, we do not expect that the elements of the vector $|S_k^{-1}e_1|$ can be arbitrary large. In fact, from (10) in the previous section, we have (with either Galerkin or minimal residual extraction)

$$|e_{j+1}^* S_k^{-1} e_1| \leq \|\mathbf{A}^{-1}\|_2.$$

For general matrices $\mathbf{A}$, the situation is more problematic. In the appendix we prove for minimal residual extraction the following, reasonably sharp, estimate in this case:

$$|e_{j+1}^* S_k^{-1} e_1| \leq \|\mathbf{A}^{-1}\|_2 \left(\frac{\|\mathbf{r}_j^{\mathrm{GAL}}\|_2}{\|\mathbf{r}_j^{\mathrm{MR}}\|_2} + \frac{\|\mathbf{r}_{j+1}^{\mathrm{GAL}}\|_2}{\|\mathbf{r}_j^{\mathrm{MR}}\|_2} \frac{\|\mathbf{r}_{j+1}^{\mathrm{MR}}\|_2}{\|\mathbf{r}_j^{\mathrm{MR}}\|_2} \right). \tag{11}$$

This shows that the j-th element of the vector $|S_k^{-1}e_1|$ might be large if the Galerkin residual is very large in the $j-1$-th or j-th step.

To illustrate the previous observations, we have included the results of a simple numerical experiment in Figure 1 where the matrix is diagonal with elements $\{1, 2, \ldots, 100\} - 5.2025$ and the right-hand side has all components equal. The matrix is constructed such that the Galerkin residual is very large in the fifth step with as a result that the $\mathbf{A}$-orthogonal basis is ill conditioned. Approximate matrix-vector products are simulated by adding random vectors of relative norm 10^{-10} to the exact product. For the methods mentioned in Table 1 we have included the results in this figure.

The left picture shows the convergence curves of the true residuals of the CG and GCR methods. Both methods apply their matrix-vector products to, in the exact case, an $\mathbf{A}$-orthogonal basis. As predicted by (11) and the expression for the gap (4), we see that the height of the largest peak of the Galerkin residuals (which coincide with the CG residuals) determines the precision that can be achieved with the GCR method. The stagnation level for the CG method is not very different, as expected. In the right picture we have two methods, ORTHORES and GMRES, that apply their matrix-vector products to an, in the exact case, orthogonal basis. The height of the largest peak is here not of importance and, in fact, the attainable precision is close to 10^{-10}. Notice that the choice of the extraction technique is not of influence on the stagnation level. Both GCR and GMRES employ minimal residual extraction, while CG and ORTHORES use the Galerkin approach. The use of short recurrences does not play a role either. In contrast to GMRES, ORTHORES relies on short recurrences.

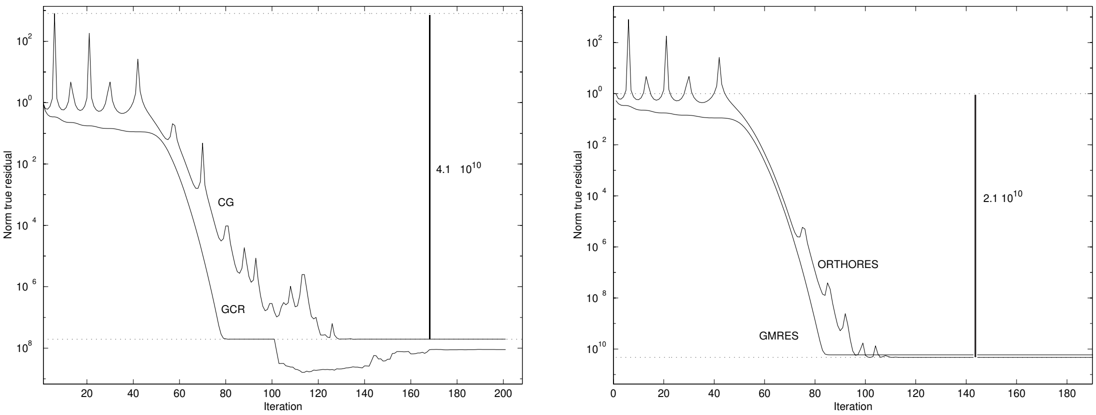

Figure 1: Illustration that the use of "**A**-orthogonal" vectors $\mathbf{z}_j$ can have a negative impact on the stagnation level of a method in case of very large intermediate Galerkin residuals.

4 Discussion

In the previous section we considered the size of the quantities $\|\mathbf{z}_j\|_2|e_{j+1}^* S_k^{-1} e_1|$ in exact iterative methods. This allowed us to give an abstract and general discussion and identify the main sources of sensitivity towards errors in the matrix-vector products. This gives a fresh point of view on stability issues in iterative linear system solvers. For example, in rounding error analyses of the *conjugate gradient* method, traditionally the sensitivity of this method in case of a large intermediate residual is contributed to instabilities in the Cholesky decomposition implicitly made in the CG method, e.g., [2]. Here we argue that it can be explained by the fact that we work with an ill-conditioned basis. This also explains why we see precisely the same stagnation level in the GCR method, which does not involve a Cholesky decomposition and, moreover, uses full orthogonalization and, in contrast to CG, extracts its iterates such that they have minimizing residuals. Moreover, our heuristical framework also gives an alternative explanation for the observations in [5] that the impact of rounding errors on the attainable accuracy of Krylov methods is not influenced essentially by the smoothing the convergence curve of the residuals. Notice, that possible instabilities are caused by the choice of an inappropriate solution method and are not part of the problem to be solved itself and can be easily circumvented by switching to a different method. For example, instead of CG for indefinite problems one can use ORTHORES.

We considered the size of the quantities in the exact case. This is in some sense an optimal scenario: if these elements are large, then they are not expected to be small in the perturbed case. Moreover, they give a good understanding of what we see happen in practice and certainly provides a good guideline for the selection of a suitable Krylov subspace as a solver. For perturbed matrix-vector products (the situation of Equation (3)) we can derive for most methods analogous results by interpreting the vector $S_k^{-1}e_1$ as constructed by an exact process applied to a Hessenberg matrix with starting vector e_1 which then proves the bound that is given in [8, 6] for the inexact GMRES method.

5 Practical consequences: relax to the max

We gave insight into the effect of approximate matrix-vector products on Krylov subspace solvers for linear systems in the previous sections. An important practical consequence of this work should be the construction of efficient strategies for controlling the error in the products such that the overall cost of approximating the products is as small as possible. From a practical point of view this means that we should allow the errors in the matrix-vector products to be as large as possible. In [8] strategies for choosing the η_j are derived by bounding each summand of the sum in (4) on a small, appropriate multiple of ϵ. Equation (10) suggests, in the relevant circumstances, to use a relative precision for the matrix-vector product in step $j+1$ that is bounded by

$$\eta_j = \frac{\epsilon}{\|\mathbf{r}_j^{\mathrm{MR}}\|_2}. \tag{12}$$

In some iterative methods we have only available the length of $\mathbf{r}_j^{\mathrm{GAL}}$ instead of $\mathbf{r}_j^{\mathrm{MR}}$, as for example the CG method. However, we notice that

$$\|\mathbf{r}_j^{\mathrm{MR}}\|_2 = \left(\sum_{i=0}^{j} \|\mathbf{r}_i^{\mathrm{GAL}}\|_2^{-2}\right)^{-1/2}.$$

An interesting property of (12) is that it requires very accurate matrix-vector products in the beginning of the process, and the precision is relaxed as soon as the method starts converging, that is, the residuals become increasingly smaller. This justifies the term *relaxation strategy* for this choice of η_j which was introduced by Bouras and Frayssé who reported various numerical results for the GMRES method in [3]. For an impressive list of numerical experiments they observe that the GMRES method with tolerance (12) converges roughly as fast as the unperturbed version, despite the, sometimes large, perturbations. Furthermore, the norm of the true residual ($\|\mathbf{b} - \mathbf{A}\mathbf{x}_j\|_2$) seemed to stagnate around a value of $\mathcal{O}(\epsilon)$.

Despite the recent efforts, the theoretical understanding of the effect of perturbations of the matrix-vector products is still not complete. Specially concerning the effect of the perturbations on the convergence speed. (Practical experience with these strategies is however very promising.) We stress that this aspect can be cheaply monitored during the iteration process whereas the residual gap can only be computed using an expensive matrix-vector product.

Acknowledgments

Part of the work of J. van den Eshof was financially supported by the Dutch scientific organization (NWO) through project 613.002.035.

A A technical result

For convenience we assume in this appendix that a vector is appended with additional zeros if this is necessary to make dimensions match. In this appendix we prove (11). In its proof we need the *Arnoldi relation*

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\underline{T}_k, \text{ with } \mathbf{V}_k e_1 = \mathbf{b},$$

and the fact that the minimal residuals are equal to the GMRES residuals given our assumption of exact arithmetic and matrix-vector products. Assume that $\underline{T}_k$ has full rank and define the vector $\overline{\gamma}_k = (\gamma_0, \ldots, \gamma_k)^* \in \mathbb{R}^{k+1}$ such that $\overline{\gamma}_k^* \underline{T}_k = \overline{0}^*$ and $e_1^* \overline{\gamma}_k = 1$. It was shown in [8] that

$$\mathbf{r}_k^{\mathrm{MR}} = \|\overline{\gamma}_k\|_2^{-2} \mathbf{V}_{k+1} \overline{\gamma}_k \quad \text{and} \quad \mathbf{r}_k^{\mathrm{GAL}} = \gamma_k^{-1} \mathbf{V}_{k+1} e_{k+1}. \tag{13}$$

The relation between the vector $\overline{\gamma}_k$ and the residuals can be expressed for the residuals $\mathbf{r}_j^{\mathrm{MR}}$ with $j = 0, \ldots, k-1$ by the relation

$$\mathbf{R}_k^{\mathrm{MR}} = \mathbf{V}_k \Upsilon_k \Theta_k^{-2} \quad \text{with} \quad \Upsilon_k e_{j+1} = \overline{\gamma}_j \text{ and } \Theta_k = \mathrm{diag}(\|\overline{\gamma}_0\|_2, \ldots, \|\overline{\gamma}_{k-1}\|_2).$$

This gives us a QR-decomposition for the matrix $\mathbf{R}_k^{\mathrm{MR}}$ which we need in the following lemma.

Lemma A.1. *For exact minimal residual approximations without preconditioning we have that*

$$S_k^{-1} e_1 = (\mathbf{A}\mathbf{R}_k^{MR})^\dagger \mathbf{r}_0^{MR} = \Theta_k^2 \Upsilon_k^{-1} \underline{T}_k^\dagger e_1 \tag{14}$$

$$e_{j+1}^* \Theta_k^2 \Upsilon_k^{-1} = \frac{\|\overline{\gamma}_j\|_2^2}{\gamma_j} e_{j+1}^* - \frac{\|\overline{\gamma}_j\|_2^2}{\gamma_{j+1}} e_{j+2}^*. \tag{15}$$

Proof. As observed we have that $\mathbf{A}\mathbf{R}_k^{\mathrm{MR}} = \mathbf{A}\mathbf{V}_k \Upsilon_k \Theta_k^{-2}$. This gives

$$(\mathbf{A}\mathbf{R}_k^{\mathrm{MR}})^\dagger \mathbf{r}_0^{\mathrm{MR}} = (\mathbf{A}\mathbf{V}_k \Upsilon_k \Theta_k^{-2})^\dagger \mathbf{r}_0^{\mathrm{MR}} = \Theta_k^2 \Upsilon_k^{-1} (\mathbf{A}\mathbf{V}_k)^\dagger \mathbf{r}_0^{\mathrm{MR}} = \Theta_k^2 \Upsilon_k^{-1} \underline{T}_k^\dagger e_1.$$

Equality (15) follows from the observation that $\Upsilon_k = \mathrm{diag}(\overline{\gamma}_{k-1}) J_k^{-1}$ where J_k is lower bidiagonal with -1 on its subdiagonal and 1 on its diagonal. $\square$

To bound the elements of the vector $\underline{T}_k^\dagger e_1$ we can use the observation that the Hessenberg matrix $\underline{T}_k$ is equal to the generated Hessenberg matrix for an exact GMRES process applied to the matrix $\underline{T}_k$ with starting vector e_1. Now we can use, with some additional work, the presented bounds in Section 3.1 (or the equivalent ones from [8, 6] for the GMRES method). In combination with (14) and (15) this gives, for general matrices $\mathbf{A}$,

$$\begin{aligned} |e_{j+1}^* S_k^{-1} e_1| &\leq \|\underline{T}_k^\dagger\|_2 \left(\frac{\|\overline{\gamma}_j\|_2}{|\gamma_j|} + \frac{\|\overline{\gamma}_j\|_2}{|\gamma_{j+1}|} \frac{\|\overline{\gamma}_j\|_2}{\|\overline{\gamma}_{j+1}\|_2} \right) \\ &\leq \|\mathbf{A}^{-1}\|_2 \left(\frac{\|\mathbf{r}_j^{\mathrm{GAL}}\|_2}{\|\mathbf{r}_j^{\mathrm{MR}}\|_2} + \frac{\|\mathbf{r}_{j+1}^{\mathrm{GAL}}\|_2}{\|\mathbf{r}_j^{\mathrm{MR}}\|_2} \frac{\|\mathbf{r}_{j+1}^{\mathrm{MR}}\|_2}{\|\mathbf{r}_j^{\mathrm{MR}}\|_2} \right). \end{aligned}$$

In the last inequality we have used (13).

References

[1] G. Arnold, N. Cundy, J. van den Eshof, A. Frommer, S. Krieg, Th. Lippert, and K. Schäfer. Numerical methods for the QCD overlap operator: III. Nested iterations. To appear.

[2] Randolph E. Bank and Tony F. Chan. A composite step bi-conjugate gradient algorithm for nonsymmetric linear systems. *Numer. Algorithms*, 7(1):1–16, 1994.

[3] A. Bouras and V. Fraysse. A relaxation strategy for inexact matrix-vector products for Krylov methods. Technical Report TR/PA/00/15, CERFACS, France, 2000.

[4] A. Bouras, V. Fraysse, and L. Giraud. A relaxation strategy for inner-outer linear solvers in domain decomposition methods. Technical Report TR/PA/00/17, CERFACS, France, 2000.

[5] Martin H. Gutknecht and Miroslav Rozložnik. Residual smoothing techniques: do they improve the limiting accuracy of iterative solvers? *BIT*, 41(1):86–114, 2001.

[6] Valeria Simoncini and Daniel Szyld. Theory of inexact Krylov subspace methods and applications to scientific computing. Technical Report 02-4-12, Department of Mathematics, Temple University, 2002. Revised version November 2002.

[7] Gerard L. G. Sleijpen, Henk A. van der Vorst, and Diederik R. Fokkema. BiCGstab(ℓ) and other hybrid Bi-CG methods. *Numer. Algorithms*, 7(1):75–109, 1994.

[8] Jasper van den Eshof and Gerard L. G. Sleijpen. Inexact Krylov subspace methods for linear systems. Preprint 1224, Dep. Math., University Utrecht, Utrecht, the Netherlands, February 2002. To appear in SIMAX.

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