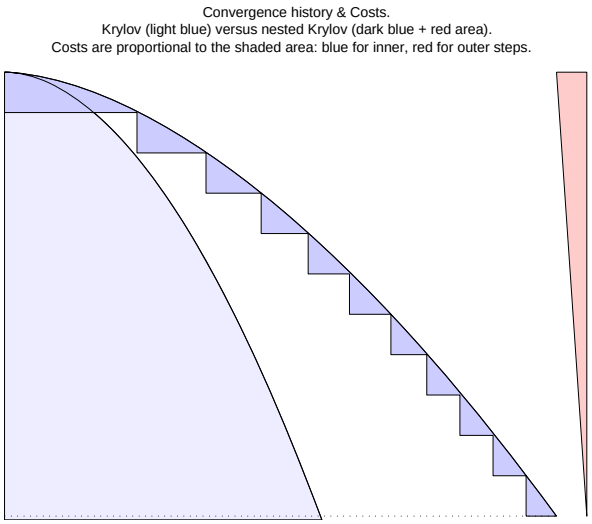

Universiteit Utrecht

Department
of Mathematics

Relaxation strategies for nested Krylov methods

by

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

Preprint

nr. 1268

March 4, 2003

RELAXATION STRATEGIES FOR NESTED KRYLOV METHODS

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

Abstract

There are classes of linear problems for which the matrix-vector product is a time consuming operation because an expensive approximation method is required to compute it to a given accuracy. In recent years different authors have investigated the use of, what is called, relaxation strategies for various Krylov subspace methods. These relaxation strategies aim to minimize the amount of work that is spent in the computation of the matrix-vector product without compromising the accuracy of the method or the convergence speed too much. In order to achieve this goal, the accuracy of the matrix-vector product is decreased when the iterative process comes closer to the solution. In this paper we show that a further significant reduction in computing time can be obtained by combining a relaxation strategy with the nesting of inexact Krylov methods. Flexible Krylov subspace methods allow variable preconditioning and therefore can be used in the outer most loop of our overall method. We analyze for several flexible Krylov methods strategies for controlling the accuracy of both the inexact matrix-vector products and of the inner iterations. The results of our analysis will be illustrated with an example that models global ocean circulation.

1 Introduction

There are classes of linear problems for which the matrix-vector product is a time consuming operation because an expensive approximation method is required to compute it to a given accuracy. Examples of such type of problems include simulations in quantum chromodynamics [21], electromagnetic applications [5, 16] and the solution of Schur complement systems [3, 19, 27]. Obviously, the more accurate the matrix-vector product is approximated the more expensive or time consuming the overall process becomes. In previous technical reports different authors [2, 3, 19, 22] have investigated the use of *relaxation strategies* for Krylov subspace methods for linear systems. The goal of these relaxation strategies is, given a required residual precision ϵ, to minimize the amount of work that is spent in the computation of the matrix-vector product. From a practical point of view this means that these strategies try to allow the error in the product to be as large as possible without compromising the accuracy of the method or its convergence speed too much (with respect to ϵ).

In [22] we derived relaxation strategies for various Krylov subspace solvers by bounding the gap between the *true residual* and the *computed residual*. This approach confirmed the empirical observations by Bouras et al. [2, 3] that a very accurate matrix-vector product is necessary in the initial phase but the precision can be relaxed as soon as the methods starts converging. Even though the work that is spent in the final iterations is very small, the gain of a relaxation strategy for practical problems is often modest compared to using a fixed tolerance. One of the reasons is the often slow convergence of Krylov subspace methods in the

*Department of Mathematics, Utrecht University, P.O. Box 80.010, NL-3508 TA Utrecht, The Netherlands. E-mail: eshof@math.uu.nl, sleijpen@math.uu.nl.

†CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse Cedex 01, France. E-mail: gijzen@cerfacs.fr.

early iterations: we are required to compute matrix-vector products to full accuracy despite the slow progress of the method at this point.

In this report we focus on the drawbacks and advantages of relaxation strategies for practical problems. After reviewing the main points of the effect of inexact matrix-vector products on Krylov methods in Section 2, we discuss in Section 3 the computational gain of using a relaxation strategy. We argue that this is often modest for practical instances. As an alternative strategy we propose to precondition an inexact Krylov subspace methods by another inexact Krylov subspace method set to a larger tolerance. We discuss several choices for the outer iteration in Section 4. Our observations are illustrated for a Schur complement problem that stems from an ocean circulation model for steady barotropic flow as described in [25].

2 Relaxation strategies for inexact Krylov subspace methods

The central problem in this report is to find a vector $\mathbf{x}'$ that approximately satisfies the equation

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \text{such that} \quad \|\mathbf{b} - \mathbf{A}\mathbf{x}'\|_2 < \epsilon, \tag{2.1}$$

for some user specified, predefined value of ϵ. Without loss of generality we assume that the vector $\mathbf{b}$ is of unit length. An important class of iterative solvers for linear systems are *Krylov subspace solvers* in which in each step only basic linear algebra operations are required including the matrix-vector product. In an *inexact* Krylov subspace method, instead of the exact matrix-vector product, we have available some device that computes an approximation $\mathcal{A}_\eta(\mathbf{v})$ to the matrix-vector product $\mathbf{A}\mathbf{v}$ to a relative precision η as

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

For obvious reasons we assume here that the computation of this vector becomes more costly when the relative precision η is picked smaller. We now summarize and discuss some known results that we need in the remainder of this paper concerning relaxation strategies and on the impact of inexact matrix-vector products on Krylov subspace solvers in general.

Bouras and Frayssé reported various numerical results for GMRES in [2] with a relative precision for the matrix-vector product in step $j+1$ that was essentially given by

$$\eta_j = \frac{\epsilon}{\|\mathbf{r}_j\|_2}. \tag{2.2}$$

The vector $\mathbf{r}_j$ is the last computed residual vector in GMRES. An interesting property of this empirical choice for η_j 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 when the residuals become increasingly smaller. This justifies the term *relaxation strategy*. For an impressive list of numerical experiments they observe that the GMRES method with tolerance (2.2) 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 $\mathcal{O}(\epsilon)$. Two recent publications [22] and [19] analyze these remarkable observations. In [22] Van den Eshof and Sleijpen explained them by bounding the gap between the true and the recursively computed residual and, separately, studying the convergence of the computed residuals. We will review the main observations from this paper now. To this purpose we will use a slightly different but more straightforward approach.

In Krylov subspace methods the *Krylov subspace* is expanded by applying the matrix-vector product 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 $\mathcal{K}_k$ defined as the span of $\{\mathbf{b}, \mathbf{Ab}, \ldots, \mathbf{A}^{k-1}\mathbf{b}\}$. (Notice that in this paper we assume that the starting vector of the iterative methods is the zero vector.) The choice of this basis plays an important role in the sensitivity of Krylov methods for perturbations on the matrix-vector product. Other quantities of interests are the *iterates*, $\mathbf{x}_j$, and the *residuals*, $\mathbf{r}_j = \mathbf{b} - \mathbf{A}\mathbf{x}_j$. In general the following relation can be identified that links together the quantities of interest:

$$\mathbf{A}\mathbf{Z}_k = \mathbf{R}_{k+1}\underline{S}_k \quad \text{and} \quad \mathbf{x}_k = \mathbf{Z}_k S_k^{-1} e_1, \tag{2.3}$$

with $\underline{S}_k$ being a $(k+1) \times k$ upper Hessenberg matrix and S_k the $k \times k$ upper block of $\underline{S}_k$. Throughout this paper capital letters 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, $\mathbf{R}_k e_{j+1} = \mathbf{r}_j$.

In case the matrix-vector product in the Krylov method is approximated to some relative precision η_j in step $j+1$, we assume that (2.3) becomes

$$\mathbf{A}\mathbf{Z}_k + \mathbf{F}_k = \mathbf{R}_{k+1}\underline{S}_k \quad \text{and} \quad \mathbf{x}_k = \mathbf{Z}_k S_k^{-1} e_1. \tag{2.4}$$

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 (2.4) causes that $\mathbf{r}_k$ is not a residual for the vector $\mathbf{x}_k$ defined by the second relation. Similarly to the work in [22] it follows from (2.4) that 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*, $\mathbf{r}_k$, is bounded by

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 = \|\mathbf{F}_k S_k^{-1} e_1\|_2 \leq \sum_{j=0}^{k-1} \eta_j \|\mathbf{A}\|_2 \|\mathbf{z}_j\|_2 |e_{j+1}^* S_k^{-1} e_1|. \tag{2.5}$$

The idea of analyzing the residual gap is not uncommon in theoretical analyses of the attainable accuracy of iterative methods in the finite precision context, see e.g., [20]. 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 inexact 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. In [22] strategies for choosing the η_j are derived by bounding each summand of the sum in (2.5) on a small, appropriate multiple of ϵ which reduces the problem to bounding the elements of the vector $|S_k^{-1} e_1|$. Because it is known that $\mathbf{x}_k = \mathbf{Z}_k S_k^{-1} e_1$, 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 $\mathbf{z}_j$.

We study the size of the elements of $S_k^{-1} e_1$ by assuming exact matrix-vector products for the moment, i.e., (2.3) holds. This problem was studied in related formulation in [19, 22]. We first have to introduce some notation. Let $\mathbf{M}$ and $\mathbf{N}$ be Hermitian, positive definite, n dimensional matrices. We define

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

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	$\sqrt{\|\mathbf{A}^{-1}\|_2}$
CR	$\mathbf{A}$	$\mathbf{A}^*\mathbf{A}$	$\sqrt{\|\mathbf{A}^{-1}\|_2}$	$\sqrt{\|\mathbf{A}^{-1}\|_2}$	1

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

which gives the following norm equivalence

$$(\delta_{\mathbf{N}\to\mathbf{M}})^{-1}\|\mathbf{y}\|_{\mathbf{N}} \le \|\mathbf{y}\|_{\mathbf{M}} \le \delta_{\mathbf{M}\to\mathbf{N}}\|\mathbf{y}\|_{\mathbf{N}}. \tag{2.6}$$

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 $\tilde{\mathbf{x}}_j \in \mathcal{K}_j$

$$|e_{j+1}^* S_k^{-1} e_1| \|\mathbf{z}_j\|_{\mathbf{M}}^2 = |\, <\mathbf{z}_j, \mathbf{x}_k>_{\mathbf{M}} \,| = |\, <\mathbf{z}_j, \mathbf{x}_k - \tilde{\mathbf{x}}_j>_{\mathbf{M}} \,| \le \|\mathbf{x}_k - \tilde{\mathbf{x}}_j\|_{\mathbf{M}} \|\mathbf{z}_j\|_{\mathbf{M}}. \tag{2.7}$$

Here we have made use of the fact that $<\mathbf{z}_j, \tilde{\mathbf{x}}_j>_{\mathbf{M}} = 0$. We define $\mathbf{x}_j^{\mathrm{MR}}$ 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 (2.7) we get the bound

$$\begin{aligned} \|\mathbf{A}\|_2 \|\mathbf{z}_j\|_2 |e_{j+1}^* S_k^{-1} e_1| &\le \|\mathbf{A}\|_2 \frac{\|\mathbf{z}_j\|_2}{\|\mathbf{z}_j\|_{\mathbf{M}}} \left(\|\mathbf{x} - \mathbf{x}_j^{\mathrm{MR}}\|_{\mathbf{M}} + \|\mathbf{x} - \mathbf{x}_k\|_{\mathbf{M}} \right) && (2.8) \\ &\le \|\mathbf{A}\|_2 \, \delta_{\mathbf{I}\to\mathbf{M}} \, \delta_{\mathbf{M}\to\mathbf{A}^*\mathbf{A}} \left(\|\mathbf{r}_j^{\mathrm{MR}}\|_2 + \|\mathbf{r}_k\|_2 \right). && (2.9) \end{aligned}$$

This simple argument shows that, if the inexact Krylov subspace method is terminated as soon as $\|\mathbf{r}_k\|_2 \le \epsilon$, then the size of the residual gap is essentially bounded by the norm of the minimal residuals times some constant.

If the particular Krylov subspace methods minimizes the error in $\mathbf{N}$-norm for some $\mathbf{N}$, then we can even remove the $\|\mathbf{r}_k\|_2$ term in (2.9). 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

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

which leads to the bound

$$\|\mathbf{A}\|_2 \|\mathbf{z}_j\|_2 |e_{j+1}^* S_k^{-1} e_1| \le \|\mathbf{A}\|_2 \, \delta_{\mathbf{I}\to\mathbf{M}} \, \delta_{\mathbf{M}\to\mathbf{N}} \, \delta_{\mathbf{N}\to\mathbf{A}^*\mathbf{A}} \|\mathbf{r}_j^{\mathrm{MR}}\|_2. \tag{2.10}$$

For several well-known Krylov subspace methods we have summarized the relevant quantities in Table 1. Substituting these values into (2.10) finally shows, for all methods mentioned in the table, that

$$\|\mathbf{A}\|_2 \|\mathbf{z}_j\|_2 |e_{j+1}^* S_k^{-1} e_1| \le \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2 \|\mathbf{r}_j^{\mathrm{MR}}\|_2. \tag{2.11}$$

Recall that this bound holds for exact matrix-vector products only. For most methods in case of perturbed matrix-vector products, i.e., we are in the situation of (2.4), analogous results can be derived 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

[22, 19] for the inexact GMRES method. We notice that the norm of the minimal residual approximation can often be cheaply computed in a Krylov method from available information. Finally, we remark that there can be advantages of relaxing on the **N**-norm of the error if this quantity is available. We have not followed this direction.

From our discussion it is clear that the optimality properties of the iterates can simplify the bound (2.9) 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 **N**-norm) for the iterates is small. (However, it can be large during some iteration steps of the iterative process.) This is comparable with the conclusions in [13] for the impact of rounding errors on the attainable accuracy of Krylov methods. 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 **M**. For example, if $\mathbf{M} = \mathbf{A}$ and **A** is indefinite then the basis can be ill conditioned and this might result in the necessity of very accurate matrix-vector products to achieve the required precision, see [22] for analysis and examples. It is important to realize that this is caused by the choice for the solution method and is not part of the problem to be solved itself.

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.

3 Practical aspects of relaxation

We try to get some insight into the expected gain of using a relaxation strategy. For the problems that we have in mind the matrix-vector product is often approximated with an iterative solver that converges linearly. An example is the computation of the matrix sign function in quantum chromodynamics with a Chebyshev series, see [21]. Therefore, we use $-\log(\eta)$ as a model for the amount of work for computing the matrix-vector product with a relative accuracy of η. It is also reasonable to assume that in every step the cost of the matrix-vector product dominates the other costs.

With this simple assumption, we have that the cost for k steps inexact GMRES with a fixed tolerance $\eta_j = \epsilon$ and a relaxed tolerance as in (2.2) are respectively given by

$$C_f = -\sum_{j=0}^{k-1} \log(\epsilon), \quad C_r = -\sum_{j=0}^{k-1} \left(\log(\epsilon) - \log(\|\mathbf{r}_j\|_2)\right). \tag{3.1}$$

It is standard practice to visualize the convergence history of iterative solvers by making a log-plot of the norms of the residuals versus the iteration number. Equation (3.1) shows that the cost of using the relaxation strategy is approximately proportional to the area between the convergence curve and the constant line ϵ whereas for the fixed strategy the cost is (approximately) proportional to the size of the area between the lines $\|\mathbf{r}_0\|_2$ and ϵ. We give a simple illustration in Figure 1 for a matrix from the Matrix Market [1].

To interpret the obtained estimates let us assume that $\|\mathbf{r}_j\|_2 = \alpha^{j^\beta}$ for some $0 < \alpha < 1$ and $\beta > 0$. Then the number of required iterations is approximately $k = (\log(\epsilon)/\log(\alpha))^{1/\beta}$ and we get the following estimate for the ratio of the cost of both strategies, assuming that

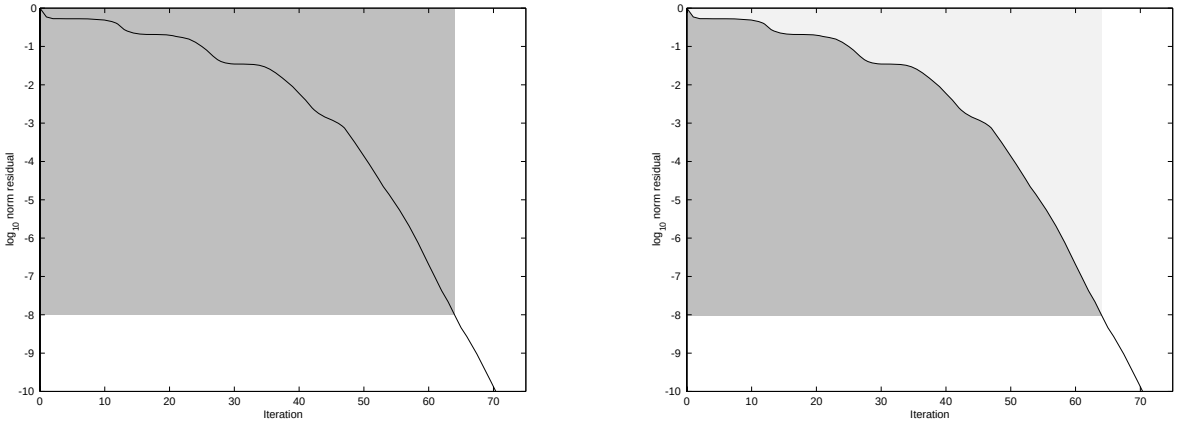

Figure 1: Convergence history GMRES for the matrix GRE115 from the Matrix Market where the dark grey area is proportional to the expected cost of the inexact Krylov subspace method under the assumption that the convergence curve is not different for the inexact method. Left picture: using a fixed accuracy of $\epsilon = 10^{-8}$. Right picture: using the relaxation strategy (2.2). The light grey area in the right picture is proportional to the expected saving.

there is no change in convergence behavior,

$$C_r/C_f \approx \int_0^k \log(\epsilon/\alpha^{x^\beta})dx / \int_0^k \log(\epsilon)dx = \frac{\beta}{1+\beta}.$$

This shows that if convergence is linear ($\beta = 1$) then the improvement is about a factor two. In case inexact GMRES converges *superlinearly*, which means that the residual decrease in the beginning is much smaller than at the final iterations (i.e., $\beta > 1$), then the advantage of relaxed GMRES becomes smaller. Superlinear convergence is, fortunately, often observed in convergence plots of the GMRES method in practical applications and if we use in this case a relaxed tolerance for the matrix-vector product then a lot of effort is spent in the first few iterations, despite the fact that progress is slow at this stage. It is tempting to reduce the accuracy in the first few iterations. However, as is reasonably well understood from the previous section, this does not work.

We derived relaxation strategies by bounding the norm of each summand in (2.5) by ϵ. Since the number of summands corresponds to the number of iterations, the accumulation of the errors can be considerable if the number of iterations to reach the required precision is large. This problem can be compensated by working with a smaller tolerance on the matrix-vector products which is the approach taken by Simoncini and Szyld [19]. Of course, this comes at some cost.

The two sketched drawbacks of relaxed GMRES are also relevant for other relaxed Krylov subspace methods. In the next section we try to reduce the effect of both problems.

4 Nested inexact Krylov subspace method

In order to reduce the number of necessary iterations of the inexact Krylov method, we propose the idea of 'preconditioning' the inexact Krylov subspace method by another inexact Krylov subspace method set to the larger tolerance of ξ_j in step $j + 1$. The idea behind

reducing the number of iterations is to keep the number of highly accurate matrix-vector products small and, furthermore, to reduce the effect of the accumulation of errors as we discussed in the previous section. We will frequently refer to the inexact Krylov method and its variable preconditioner as the *outer iteration* and *inner iteration* respectively. Methods that can be used for the outer iteration are the so-called *flexible methods*. These are methods that are specially designed for dealing with variable preconditioning, e.g., [11, 18, 23] which we combine with an approximate matrix-vector product.

We discuss a few choices for the outer iteration in the remainder of this section. For this purpose, we need the following notation: when an inexact Krylov subspace method is used for solving the linear system $\mathbf{Az} = \mathbf{y}$ with a relative residual precision of at least ξ, we will write

$$\mathbf{z} = \mathcal{P}_\xi(\mathbf{y}), \quad \text{where } \mathbf{z} \text{ such that } \|\mathbf{y} - \mathbf{Az}\|_2 \leq \xi \|\mathbf{y}\|_2.$$

Notice that this is a so-called *flexible* preconditioner that may change every step depending not only on ξ but also on $\mathbf{y}$.

In our nested algorithm the necessary 'new' information about the true matrix is introduced in the outer iteration and the accuracy of the matrix-vector products in the inner iteration is only modest. A drawback of this method is that in general nesting Krylov subspace methods affects, usually negatively, the efficiency with respect to the total number of matrix-vector products (the 'convergence speed'). The goal is to make a trade-off between choosing the ξ_j small (leading to a small k in this case) and thereby avoiding the computation of many accurate matrix-vector products in the outer iteration and, on the other hand, keeping the ξ_j large which is expected to reduce the total number of matrix-vector products necessary (since the optimality of the outer iteration becomes more effective). This issue is discussed in more detail in Section 4.4.

4.1 The outer iteration: Richardson iteration

The *nested* inexact Krylov subspace method with Richardson iteration as outer iteration is, for $j = 1, 2, \ldots, k$, defined by the following recurrences

$$\begin{array}{lcl} \mathbf{z}_{j-1} & = & \mathcal{P}_{\xi_{j-1}}(\mathbf{r}_{j-1}) \\ \mathbf{x}_j & = & \mathbf{x}_{j-1} + \mathbf{z}_{j-1} \\ \mathbf{r}_j & = & \mathbf{r}_{j-1} - \mathcal{A}_{\eta_{j-1}}(\mathbf{z}_{j-1}). \end{array} \tag{4.1}$$

It can be easily checked that this method fits into our general relation (2.3):

$$\mathbf{AZ}_k + \mathbf{F}_k = \mathbf{R}_{k+1}\underline{S}_k, \quad \text{and} \quad \mathbf{x}_k = \mathbf{Z}_k S_k^{-1} e_1, \tag{4.2}$$

with $\underline{S}_k$ lower bidiagonal with one on its diagonal and minus one on its subdiagonal. Therefore, $S_k^{-1}e_1 = \vec{1}$ and using, furthermore, the estimate $\|\mathbf{z}_j\|_2 \leq \|\mathbf{A}^{-1}\|_2 \|\mathbf{r}_j\|_2(1 + \xi_j)$, we find with (2.5) the following bound on the norm of the residual gap:

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{Ax}_k)\|_2 \leq \|\mathbf{A}\|_2 \sum_{j=0}^{k-1} \eta_j \|\mathbf{z}_j\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2 \sum_{j=0}^{k-1} \eta_j \|\mathbf{r}_j\|_2 (1 + \xi_j).$$

This suggests to pick the tolerance η_j equal to $\epsilon / \|\mathbf{r}_j\|_2$ in step $j + 1$ which is consistent with the in [22] proposed strategy for inexact Richardson iteration in the unpreconditioned case.

This version of Richardson iteration uses recursively computed residuals which is essential for a relaxation strategy to be feasible as discussed in [22]. We can alternatively compute the residuals directly, i.e., the residuals are instead computed as

$$\mathbf{r}_j = \mathbf{b} - \mathcal{A}_{\eta_j}(\mathbf{x}_j). \tag{4.3}$$

To derive a suitable strategy for choosing the η_j for this version we can exploit that, with (4.3), there is no propagation of errors throughout the process. Deriving a strategy by bounding the residual gap is therefore for this method not very useful. We have that

$$\begin{aligned} \|\mathbf{b} - \mathbf{A}\mathbf{x}_j\|_2 &= \|\mathbf{b} - \mathbf{A}(\mathbf{x}_{j-1} + \mathbf{z}_{j-1})\|_2 \\ &\leq \|\mathbf{r}_{j-1} - (\mathbf{b} - \mathbf{A}\mathbf{x}_{j-1})\|_2 + \|\mathbf{r}_{j-1} - \mathbf{A}\mathbf{z}_{j-1}\|_2 \\ &\leq \eta_{j-1}\|\mathbf{A}\|_2\|\mathbf{x}_{j-1}\|_2 + \xi_{j-1}\|\mathbf{r}_{j-1}\|_2. \end{aligned}$$

This suggests to pick $\eta_j = \xi_j\|\mathbf{r}_j\|_2$ such that we, roughly, have that

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}_j\|_2 \lesssim \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2\,\xi_{j-1}\|\mathbf{r}_{j-1}\|_2.$$

In other words, the precision of the matrix-vector product is chosen equal to the current residual precision times the expected residual reduction. This latter version of Richardson iteration can viewed as making periodic restarts of the Krylov subspace method used in the inner iteration. This gives relations with strategies for dealing with approximate matrix-vector products that have been proposed in literature. These connections are discussed in Section 4.5.

The total work that is spent in the outer iteration on the matrix-vector products is for both versions of Richardson iteration approximately the same. The difference is that we work with an increasingly tighter tolerance in the outer loop for the latter strategy whereas for the first discussed version of Richardson iteration we relax the tolerance more in the later iteration steps. An advantage of the use of the directly computed residuals is that there is no 'memory' in the iterative method and the precision ϵ has not to be decided a priori. Furthermore, there is no accumulation of errors (so the error is independent of the number of iterations). For this reason using directly computed residuals can be necessary in some applications, see e.g., [9] where the authors discuss the approximate solution of infinite dimensional systems. On the other hand, the advantage of the recursively computed residual is that we do not have to estimate the residual reduction in the coming step. (Notice that this is not precisely ξ_{j-1} in practice.) Finally, the recursively computed residual is an essential ingredient of optimal methods like, for example, flexible GMRES [18] and GMRESR [23].

4.2 The outer iteration: flexible GMRES

The *flexible* GMRES method by Saad [18] is a variant of the GMRES method that can deal with variable preconditioning. It constructs an orthogonal basis $\mathbf{V}_k$ for $\mathbf{A}\mathbf{Z}_k$ where $\mathbf{z}_j = \mathcal{P}_{\xi_j}(\mathbf{v}_j)$ which, with inexact matrix-vector product, can be summarized by the relations

$$\mathbf{A}\mathbf{Z}_k + \mathbf{F}_k = \mathbf{V}_{k+1}\underline{T}_k, \quad \mathbf{x}_k = \mathbf{Z}_k\underline{T}_k^{\dagger}e_1 \quad \text{and } \mathbf{r}_k = \mathbf{V}_{k+1}(I - \underline{T}_k\underline{T}_k^{\dagger})e_1. \tag{4.4}$$

With some small effort these relations can be transformed into our general relations given in (2.4). We will not give the details here. In order to bound the norm of the gap we can also

use the relations (4.4) directly as in [22, Section 7]. This shows that

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 = \|\mathbf{F}_k \underline{T}_k^\dagger e_1\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2 \|\underline{T}_k^\dagger\|_2 \sum_{j=0}^{k-1} \eta_j \|\mathbf{r}_j\|_2 (1 + \xi_j),$$

where we, moreover, have used [22, Lemma 3.1] and the estimate for the $\|\mathbf{z}_j\|_2$ as mentioned in the previous section. As for the standard inexact GMRES method analyzed in [22, 19], the norm of $\underline{T}_k^\dagger$ is difficult to bound a priori. In the exact case (i.e., $\eta_j = 0$ for all j), we have, with $\sigma_{\min}(B)$ denoting the smallest singular value of the matrix B, for small enough precisions ξ_j:

$$\sigma_{\min}(\underline{T}_k) = \sigma_{\min}(\mathbf{V}_k + (\mathbf{A}\mathbf{Z}_k - \mathbf{V}_k)) \geq 1 - \sqrt{k} \max_j \xi_j.$$

For this reason we assume that $\|\underline{T}_k^\dagger\|_2$ is bounded by a modest constant in the remainder of this section. This indicates that the relaxation strategy for inexact GMRES given by (2.2) is also useful in the flexible context and leads to

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 \leq k\epsilon \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2 \|\underline{T}_k^\dagger\|_2 (1 + \max_j \xi_j).$$

A notable difference with Richardson iteration described in the previous section is that, in flexible GMRES, the preconditioner is applied to multiples of the residuals of the corresponding Galerkin approximations (also referred to as the approximations from the associated flexible FOM process). For this reason the residual reduction of flexible GMRES can be less than ξ_j in step $j+1$ if the residual of the corresponding Galerkin approximations is large. This observation is due to Vuik [26] and is not difficult to understand in case of exact matrix-vector products: if we define $\alpha \equiv e_{j+1}^*(I - \underline{T}_j T_j^{-1})e_1$, then, using the optimality of flexible GMRES, we find with $y_{j+1} \equiv [(T_j^{-1}e_1)^{\mathrm{T}},\, \alpha]^{\mathrm{T}}$ that

$$\begin{aligned} \|\mathbf{r}_{j+1}\|_2 &\leq \|\mathbf{b} - \mathbf{A}\mathbf{Z}_{j+1} y_{j+1}\|_2 = \|\mathbf{b} - \mathbf{A}\mathbf{Z}_j T_j^{-1} e_1 - \mathbf{A}\mathbf{z}_{j+1}\alpha\|_2 \\ &= \|(\mathbf{v}_{j+1} - \mathbf{A}\mathbf{z}_{j+1})\alpha\|_2 \leq \xi_j \alpha. \end{aligned} \tag{4.5}$$

These type of results can be found in [26]. It shows that in case the associated FOM process suffers from near breakdowns, i.e., its residuals are very large, flexible GMRES might not be a suitable choice. However, if we have that $\xi_j = \xi$ for all j and ξ is small enough this is no serious problem as the following lemma demonstrates for the case of exact matrix-vector products. We mention that for Richardson iteration, with exact matrix-vector products, we have that $\|\mathbf{r}_k\|_2 \leq \prod_{i=0}^{k-1} \xi_i$.

Lemma 4.1. *Let $\xi_j = \xi < 1/2$ and $\eta_j = 0$ for all j. Then,*

$$\|\mathbf{r}_k\|_2^2 \leq \xi^{-2k} \frac{1}{\beta} \left(\left(\frac{1+\beta}{2} \right)^{k+1} - \left(\frac{1-\beta}{2} \right)^{k+1} \right) \qquad \textit{with } \beta \equiv \sqrt{1 - 4\xi^2}.$$

Proof. We have that

$$\frac{1}{\|\mathbf{r}_{k+1}\|_2^2} \geq \frac{1}{\xi^2} \frac{1}{\alpha^2} = \frac{1}{\xi^2} \left(\frac{1}{\|\mathbf{r}_k\|_2^2} - \frac{1}{\|\mathbf{r}_{k-1}\|_2^2} \right), \tag{4.6}$$

where for the last equation we have used a well-known relation due to Brown [4], see also [22]. We define $\gamma_k = \xi^{2k}/\|\mathbf{r}_k\|_2^2$. From (4.6) it follows that $\gamma_{k+1} \geq \gamma_k - \xi^2 \gamma_{k-1}$ and $\gamma_0 = \gamma_1 = 1$.

Furthermore, we introduce the quantity ρ_k that satisfies the recursion $\rho_{k+1} = \rho_k - \xi^2\rho_{k-1}$ with $\rho_0 = \rho_1 = 1$. Our first step is to show that for all k we have that $0 \leq \rho_k \leq \gamma_k$. If $\xi \leq 1/2$ then there exist constants $\alpha, \beta \in [0,1]$ such that $\alpha + \beta = 1$ and $\alpha\beta = \xi^2$. Hence,

$$\begin{array}{rcl} (\gamma_{k+1} - \alpha\gamma_k) & \geq & \beta(\gamma_k - \alpha\gamma_{k-1}), \quad \gamma_1 - \alpha\gamma_0 = \rho_1 - \alpha\rho_0 = 1 - \alpha \geq 0 \\ (\rho_{k+1} - \alpha\rho_k) & = & \beta(\rho_k - \alpha\rho_{k-1}). \end{array}$$

This shows by induction that $(\gamma_k - \alpha\gamma_{k-1}) \geq (\rho_k - \alpha\rho_{k-1}) \geq 0$. Hence, $\gamma_k \geq \alpha\gamma_{k-1} + (\rho_k - \alpha\rho_{k-1})$. Again with induction, we find that $\gamma_k \geq \alpha\rho_{k-1} + (\rho_k - \alpha\rho_{k-1}) = \rho_k \geq 0$. The proof is concluded by solving the recursion for the ρ_k using standard techniques. □

This lemma demonstrates that $\xi^{-k}\|\mathbf{r}_k\|_2$ approaches a value smaller than or equal to one for ξ going to zero and, therefore, the disadvantage of flexible GMRES that the residual reduction can be much less than ξ does not play an important role in our context where we work with a constantly modest value of ξ. For a numerical illustration of this observation we refer the reader to our numerical experiments in Section 5.

4.3 The outer iteration: GMRESR

The *GMRESR* method of Van der Vorst and Vuik [23] is another variant of GMRES that allows for flexible preconditioning. The authors propose for the inner iteration (or, in other words, as flexible preconditioner) to use a few steps of the GMRES method. This makes this method also very interesting as outer iteration method in our context and we consider this option in this section. In the GMRESR method the flexible preconditioner is directly applied to the last computed residual. This means that we have $\mathbf{z}_j = \mathcal{P}_{\xi_j}(\mathbf{r}_j)$. The GMRESR method minimizes the residual by constructing its iterates as a suitable linear combination of all previously computed vectors $\mathbf{z}_j$. Therefore, a simple argument shows that this guarantees in the exact case a residual reduction of at least ξ_j in step $j+1$ which can be an advantage over the use of flexible GMRES.

We now discuss the matrix formulation of the GMRESR method with inexact matrix-vector product. In the inexact GMRESR method, in every step decompositions are updated such that

$$\mathbf{A}\mathbf{Z}_k + \mathbf{F}_k = \mathbf{C}_k B_k, \quad \mathbf{Z}_k = \mathbf{U}_k B_k, \quad \text{with } \mathbf{C}_k^*\mathbf{C}_k = I_k \text{ and } B_k \text{ upper triangular.} \tag{4.7}$$

In the second part of the iteration step, the residual and iterate are updated as follows

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{u}_{k-1}(\mathbf{c}_{k-1}^*\mathbf{r}_{k-1}), \quad \mathbf{r}_k = \mathbf{r}_{k-1} - \mathbf{c}_{k-1}(\mathbf{c}_{k-1}^*\mathbf{r}_{k-1}). \tag{4.8}$$

Notice that, in contrast to a standard implementation of GMRESR, we assumed here that the vectors $\mathbf{c}_j$ are normalized which is not an essential restriction. The last two relations can be expressed by the matrix relations

$$\mathbf{C}_k D_k = \mathbf{R}_{k+1}\underline{J}_k \quad \text{and} \quad \mathbf{U}_k D_k = \mathbf{X}_{k+1}\underline{J}_k,$$

where $D_k \equiv \mathrm{diag}(\mathbf{c}_0^*\mathbf{r}_0, \ldots, \mathbf{c}_{k-1}^*\mathbf{r}_{k-1})$, $\underline{J}_k$ is the $(k+1) \times k$ matrix with one on its diagonal and minus one on its subdiagonal and J_k is the upper $k \times k$ block of $\underline{J}_k$. Substituting all relations and using that $D_k J_k^{-1} e_1 = D_k\vec{1} = \mathbf{C}_k^*\mathbf{r}_0$ (since $\mathbf{c}_j^*\mathbf{r}_j = \mathbf{c}_j^*\mathbf{r}_0$), we find that

$$\mathbf{A}\mathbf{Z}_k + \mathbf{F}_k = \mathbf{R}_{k+1}(\underline{J}_k D_k^{-1} B_k) \quad \text{and} \quad \mathbf{x}_k = \mathbf{U}_k(\mathbf{C}_k^*\mathbf{r}_0) = \mathbf{Z}_k(J_k D_k^{-1} B_k)^{-1} e_1.$$

This shows that GMRESR with inexact matrix-vector product fits into the general framework of this paper given by (2.4). The residual gap is therefore determined by the elements of the vector $S_k^{-1}e_1$, see (2.5). For this vector we have that

$$S_k^{-1}e_1 = B_k^{-1}D_kJ_k^{-1}e_1 = B_k^{-1}\mathbf{C}_k^*\mathbf{r}_0 = (\mathbf{A}\mathbf{Z}_k + \mathbf{F}_k)^\dagger\mathbf{r}_0.$$

The size of the elements of this vector are difficult to assess. In general small errors in the matrix-vector product can have a considerable influence on the residual gap. This happens if $\mathbf{A}\mathbf{z}_k + \mathbf{f}_k$ is for a large part contained in the span of the previously computed vectors, in this case the resulting vector after the orthogonalization procedure is for a relatively large part contaminated by the error in the matrix-vector product, i.e., $\|\mathbf{A}\mathbf{u}_k - \mathbf{c}_k\|_2$ is large. Such an unfortunate cancellation is also reflected by large elements in the vector $S_k^{-1}e_1$. We are interested in the size of the residual gap and we will investigate the size of the elements of the vector $S_k^{-1}e_1$ in the remainder of this section. We will restrict ourselves to the case of exact matrix-vector products for simplicity. First we study the situation in which there is no preconditioning, i.e., $\mathcal{P}_\xi(\mathbf{y}) = \mathbf{y}$ for all $\mathbf{y}$. Then GMRESR reduces to GCR [8].

If the matrix $\mathbf{A}$ is Hermitian positive definite then the residuals $\mathbf{r}_j$ $(= \mathbf{z}_j)$ form an orthogonal basis with respect to the $\mathbf{A}$-inner product and, therefore, $\mathbf{R}_k$ is orthogonal with respect to a well-defined inner product. Since GCR minimizes the residual, the iterate $\mathbf{x}_k = \mathbf{R}_kS_k^{-1}e_1$ must be bounded in norm. Consequently, we do not expect that the elements of the vector $S_k^{-1}e_1$ can be incidentally large due to a near breakdown (or in other words, the effect of cancellation is expected to be limited). In fact, from (2.11) in Section 2 we have for the CR method (or GCR applied to a Hermitian positive definite matrix) that

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

Although we did not investigate the influence of the inexact matrix-vector products, this shows that, for these type of matrices, we may expect good results using the relaxation strategy given in (2.2).

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

$$|e_{j+1}^*S_k^{-1}e_1| \leq \|\mathbf{A}^{-1}\|_2 \left(\frac{\|\mathbf{r}_j^{\mathrm{F}}\|_2}{\|\mathbf{r}_j\|_2} + \frac{\|\mathbf{r}_{j+1}^{\mathrm{F}}\|_2}{\|\mathbf{r}_j\|_2}\frac{\|\mathbf{r}_{j+1}\|_2}{\|\mathbf{r}_j\|_2} \right), \tag{4.9}$$

where the $\mathbf{r}_j^{\mathrm{F}}$ are the residuals of the corresponding FOM process. This shows that the elements of the vector $|S_k^{-1}e_1|$ might be large if the related FOM process has a near breakdown, this is similar to the bounds on the residual gap for the CG method that we have seen in [22]. This is not surprising since both methods construct their iterates from an $\mathbf{A}$-orthogonal basis which can break down. The sensitivity in case of large FOM residuals can also be interpreted differently: if the related FOM residual is large then GCR nearly stagnates which means that the GCR residuals are close in two consecutive iterations which leads to cancellation in the orthogonalization part.

We conclude that, although GCR residuals coincide with GMRES residuals in the exact case, in contrast to GMRES, GCR suffers from FOM peaks (and even FOM peaks in the next step). This shows, again, that it is the choice of the basis $\mathbf{z}_0, \ldots, \mathbf{z}_{k-1}$ that essentially determines the sensitivity of a Krylov method to approximate matrix-vector products.

In our application GMRESR is used in an outer iteration with variable preconditioning which means that we work with a preconditioner that reduces the residual in step $j + 1$

with at least a factor ξ_j. If $\xi_j = \xi$ for all j and ξ is small we expect that the situation is not very different from the case of Richardson iteration in Section 4.1. The techniques from Appendix A can not be applied in this case. Instead we have the following lemma.

Lemma 4.2. *Let $\xi_j = \xi$ for all $j \in \{0, k-1\}$. Then,*

$$|e_{j+1}^* S_k^{-1} e_1| \leq (1-\xi)^{j-k}.$$

Proof. We have in step $j+1$ that $\mathbf{A}\mathbf{z}_j = \mathbf{r}_j + (\mathbf{A}\mathbf{z}_j - \mathbf{r}_j)$ with $\|\mathbf{A}\mathbf{z}_j - \mathbf{r}_j\|_2 \leq \xi \|\mathbf{r}_j\|_2$. Since, $\mathbf{C}_j^* \mathbf{r}_j = \vec{0}$ we find for the diagonal elements $|B_k|_{jj} \geq \|\mathbf{r}_{j-1}\|_2 (1-\xi)$ and for the off-diagonal elements $|B_k|_{ij} \leq \|\mathbf{r}_{j-1}\|_2 \xi$ for $i < j$. We can use this to show that

$$|B_k^{-1}| \leq \frac{1}{1-\xi} \mathrm{diag}(\|\mathbf{r}_0\|_2, \ldots, \|\mathbf{r}_{k-1}\|_2)^{-1} \tilde{B}_k^{-1}.$$

The matrix $\tilde{B}_k$ is upper triangular with one on its diagonal and $-\alpha$ in all its off-diagonal entries with $\alpha \equiv \xi/(1-\xi)$. It is not difficult to prove that $(\tilde{B}_k^{-1})_{ij}$ equals one for $i = j$ and equals $\alpha(1+\alpha)^{j-1-i}$ for $i < j$ (see also [15, Equation 8.3]). The vector $e_{j+1}^* \tilde{B}_k^{-1}$ is zero in the first j components and we can show that

$$\|\tilde{B}_k^{-*} e_{j+1}\|_1 = 1 + \alpha \sum_{i=0}^{k-j-2} (1+\alpha)^i = (1+\alpha)^{k-j-1} = (1-\xi)^{j+1-k}.$$

The proof is completed as follows

$$|e_{j+1}^* S_k^{-1} e_1| \leq |e_{j+1}^* B_k^{-1}| \; |C_k^* \mathbf{r}_0| \leq (1-\xi)^{j-k} \|\mathbf{r}_j\|_2^{-1} \max_{i>j} |\mathbf{c}_i^* \mathbf{r}_0| \leq (1-\xi)^{j-k}.$$

□

We expect that we can safely use the relaxation strategy (2.2) for inexact GMRESR when the preconditioner leads to a residual reduction in every step that is big enough. Notice that this observation is the converse of our findings for flexible GMRES in the previous section. There, we argued that a near breakdown of the associated FOM process results in a residual reduction that can be less than what can be expected from the effort put into the inner iteration. This is, however, not expected to be a problem if the residual reduction in every step is big enough. Conversely, in GMRESR problems can occur with the accuracy of the method in case of a near breakdown, however, if ξ is small enough in every step, this effect is small. We conclude that the differences between both methods are small for ξ small enough. See also our numerical experiments in Section 5.

4.4 The choice of the precisions ξ_j

Making general statements about the optimal sequence of tolerances ξ_j is difficult since this requires detailed knowledge about the convergence behavior of Krylov subspace methods including the case of variable preconditioners. Nevertheless, we want to provide some insight into the problem of selecting the tolerances. To this purpose, we assume that the residual reduction in each step of each inner iteration is constant. This is a reasonable assumption for methods like Richardson iteration. Furthermore, the required residual precision is set to

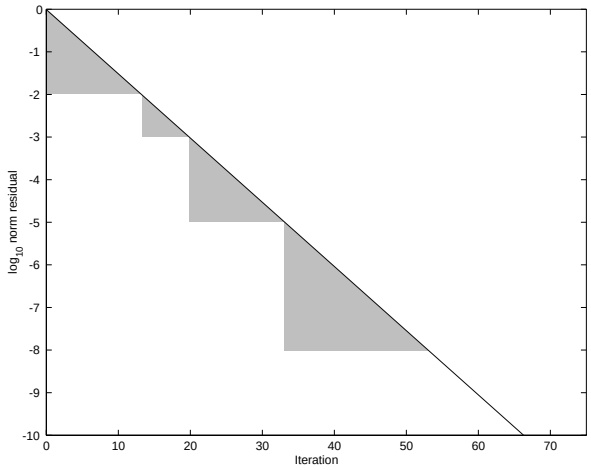

Figure 2: The grey area approximates the cost of the inner iteration when convergence in inner iteration is linear.

$\epsilon = 10^{-l}$ and the residual reduction in the outer iteration is assumed to be precisely ξ_j in step $j+1$. This means that the norm of the residual at the end of step j is equal to ϵ_j with

$$1 = \epsilon_0, \epsilon_1, \ldots, \epsilon_k = \epsilon = 10^{-l}, \quad \text{and } \epsilon_j = \prod_{i=0}^{j-1} \xi_i.$$

With these assumptions and following the same reasoning as in Section 3, we have that the cost for the inner iteration is approximated by the grey area in Figure 2 when a standard relaxation strategy is used in the inner iteration. In this example we have $\xi_0 = 10^{-2}$, $\xi_1 = 10^{-1}$, $\xi_2 = 10^{-2}$, $\xi_3 = 10^{-3}$.

If we define $\epsilon_j = 10^{-t_j}$ (hence $t_0 = 0$ and $t_k = l$) and assume that the total number of inner iterations is equal to m, then we have the following approximation to the cost:

$$\tilde{C}_k = \frac{m}{2l} \sum_{i=1}^{k} (t_i - t_{i-1})^2 + \sum_{i=0}^{k} t_i, \tag{4.10}$$

where the second term represents the cost of the matrix-vector products in the outer iteration for Richardson with directly computed residuals, see Section 4.1.

Lemma 4.3. *For fixed $k < \frac{1}{2}(1 + \sqrt{1+8m})$ the quantity $\tilde{C}_k$ in (4.10) is minimized by*

$$t_i = \frac{l}{2m} i(i-k) + \frac{l}{k} i \quad \text{for } i \in \{0, \ldots, k\}. \tag{4.11}$$

Furthermore, the optimal value of k is given by the largest integer strictly smaller than $\frac{1}{2}(1 + \sqrt{1+8m})$.

Proof. Differentiating (4.10) with respect to t_i (for $i \in \{1, \ldots, k-1\}$) and equating to zero gives the equation

$$\frac{m}{l} \left(-t_{i-1} + 2t_i - t_{i+1}\right) + 1 = 0.$$

Using standard theory for solving recurrences we get that $t_i = \frac{l}{2m}i^2 + \alpha i + \beta$ for some constants α and β. Using the boundary conditions $t_0 = 0$ and $t_k = l$ we find the required expression. It can be easily verified that if $k < \frac{1}{2}(1 + \sqrt{1 + 8m})$ then all the t_i in (4.11) are in the open interval $(0, l)$ and therefore are the optimal points. Furthermore, we have that the minimal value of $\tilde{C}_k$ is smaller than the minimal value of $\tilde{C}_{k-1}$ (to see this, pick for $\tilde{C}_k$ the t_j equal to the optimal points for $\tilde{C}_{k-1}$ and choose the additional point equal to zero). Suppose that k' is the largest integer strictly smaller than $\frac{1}{2}(1 + \sqrt{1 + 8m})$ then the global minimum of $\tilde{C}_{k'+1}$ defined by (4.11) is outside the bounding box and therefore the minimum is attained at the boundary of the box (that is $t_i = 0$ for one or more indices i). The minimal value of $\tilde{C}_{k'}$ equals the minimal value of $\tilde{C}_{k'+1}$. The optimal number of (nonzero) tolerances is given by $k = k'$. □

It is interesting to notice that in the common case that the number of inner iterations, m, is large compared to the number of outer iterations k, we have for the optimal t_i that $t_i \approx il/k$. This implies that keeping ξ_{j-1} constant is almost optimal.

The above analysis is for Richardson's method with directly computed residuals as outer iteration. The analysis for the case when relaxation is employed in the outer iteration is completely analogous and yields the same results. The conclusion that ξ_{j-1} constant is almost optimal also holds for this case. For this reason we will use $\xi_j = \xi$ in our numerical experiments in Section 5.

4.5 Discussion

Some inexact methods proposed in literature have important connections with our nested inexact Krylov subspace methods with Richardson iteration with directly computed residuals in the outer iteration as described in Section 4.1. In that section we also mentioned that this method can be interpreted as frequently restarting the inexact Krylov subspace method at which point the required residual precision is decreased. Alternatively, it can be viewed as an iterative refinement procedure. We discuss some related ideas that can be found in literature.

Golub and his co-workers [12] introduced the *successive Lanczos method* for the computation of eigenvectors with the Lanczos method. For this method, the user has to specify in advance the tolerances ϵ_j and the precise number of iterations for each inner iteration (there appears to be no automatic stopping condition for the inner iterations). At the beginning of each cycle the computed approximation from the previous cycle is used as a starting vector. In the presented numerical experiments the method was not combined with a relaxation strategy for the inner iterations (the possibility is mentioned).

For solving linear systems also related strategies have been applied. Hernández et al. explored in [14] an algorithm very similar to our method in the context of Neuberger fermions, e.g., [17]. They refer to this method as scheme (b). In their approach they also choose a sequence of tolerances ϵ_j and first solve the linear system with a precision ϵ_1 by working with a fixed approximation to the matrix-vector product that has an error roughly proportional to ϵ_1. They use the resulting approximate solution as a starting vector for solving the linear system with a precision ϵ_2, and so on. In their j-th step they use a matrix-vector product that has a relative precision of order ϵ_j. This is an essential difference with our proposal: due to the good starting vector the relative residual reduction in the j-th step is only $\epsilon_j/\epsilon_{j-1}$ which is a bounded quantity and, as we argued, this allows us to use a much cheaper matrix-vector product, except only for the few matrix-vector products in the outer iteration.

The first scheme that they suggest, scheme (a), is also clearly related. To explain this we need to define two approximation to the matrix $\mathbf{A}$,

$$\mathbf{A} \approx \mathbf{A}_1 = \mathbf{A}_2 + \Delta\mathbf{A}_2,$$

where $\mathbf{A}_1$ is a fixed approximation with a relative precision of order ϵ and $\mathbf{A}_2$ a relative precision of order ϵ'. Then they subsequently solve

$$\mathbf{A}_2\mathbf{x}_0 = \mathbf{b}, \quad \mathbf{A}_2\mathbf{x}_1 = \mathbf{b} - \Delta\mathbf{A}_2\mathbf{x}_0, \quad \mathbf{A}_1\mathbf{x}_2 = \mathbf{b},$$

where in the last step they use $\mathbf{x}_1$ as a starting vector. Solving the first two steps can be seen as equivalent to two steps with $\xi_0 = \xi_1 = \epsilon'$. The third step differs from our approach for the same reason as indicated for scheme (b): we essentially work with a precision of $\xi_3 = \epsilon/(\epsilon')^2$ for the matrix-vector product in this last step.

Giusti et al. recently proposed [10, Section 9], what they call, an *adapted-precision inversion algorithm* for problems from quantum chromodynamics which they interpret as some form of iterative refinement. This is essentially equal to the method proposed here with Richardson iteration with directly computed residuals in the outer iterations. The authors do not discuss specific choices for the precision of the matrix-vector product in the outer iteration and use a fixed precision in the inner iteration.

Carpentieri [5] describes experiments with nested Krylov subspace methods that use inexact matrix-vector products. He uses flexible GMRES in the outer iteration and GMRES in the inner iteration. The matrix-vector products, which are computed using a fast multipole technique, are evaluated to a high precision in the outer iteration, whereas the matrix vector-products in the inner iteration are evaluated to a lower, but fixed, precision. In the numerical experiments that are described in the thesis no relaxation strategies are applied.

The use of inner iterations set to variable precisions in the context of flexible or variable preconditioning was investigated by Dekker [7] for a problem from domain decomposition.

5 Numerical experiments

5.1 An example from global ocean circulation

In this section we present our, preliminary, experience with nesting inexact Krylov subspace methods for solving Schur complement systems. The example comes from a finite element discretization of a model that describes the steady barotropic flow in a homogeneous ocean with constant depth and in nearly equilibrium as described in [25]. For convenience of the reader we summarize here the main points. The model is described by the following set of partial differential equations:

$$\begin{aligned} -r\nabla^2\psi - \beta\frac{\partial\psi}{\partial x} - \alpha\nabla^2\zeta &= \nabla \times F \quad \text{in } \Omega \\ \nabla^2\psi + \zeta &= 0, \end{aligned}$$

in which ψ is the stream-function and ζ the vorticity. The domain Ω is the part of the world that is covered by sea. The external force field F is equal to the wind stress τ divided by product of the average water depth H and the water density ρ. The other parameters in these equations are: the lateral viscosity α, the bottom friction r and the Coriolis parameter β.

The above equations have to be complemented by a suitable set of boundary conditions. On the boundary of a continent usually the no-slip condition is imposed, which means that both the normal and the tangential flow are zero. The no-normal-flow boundary condition implies that the stream-function is constant on each continent,

$$\psi = C_k \quad \text{on } \Gamma_k, \quad k = 1, \ldots, n_{\text{con}},$$

where n_{con} is the total number of continents. We can determine the unknown constants by imposing that the water-level is continuous around each continent boundary Γ_k. This yields for each continent an integral condition given by

$$\oint_{\Gamma_k} \alpha \frac{\partial \zeta}{\partial n}\, ds = -\oint_{\Gamma_k} F \cdot s\, ds.$$

The no-tangential-flow condition is given in terms of the stream-function by

$$\frac{\partial \psi}{\partial n} = 0 \quad \text{on } \Gamma_k, \quad k = 1, \ldots, n_{\text{con}}.$$

The equations are commonly expressed in spherical coordinates, which maps the physical domain onto a rectangular domain with periodic boundary conditions

$$\psi(-\pi, \theta) = \psi(\pi, \theta), \quad \zeta(-\pi, \theta) = \zeta(\pi, \theta), \tag{5.1}$$

in which θ is the latitude. To fix the level of the stream-function (which is determined by the above equations only up to an arbitrary constant) one can impose the slip condition $\psi = \zeta = 0$ on the North Pole. An additional advantage of using this slip condition is that the singularity at the North Pole, introduced by the use of spherical coordinates, is excluded.

The above equations are discretized with the method described in [25]. This gives the following linear system of equations

$$\begin{bmatrix} r\mathbf{L} - \mathbf{C} & \alpha\tilde{\mathbf{L}} \\ -\tilde{\mathbf{L}}^* & \mathbf{M} \end{bmatrix} \begin{bmatrix} \boldsymbol{\psi} \\ \boldsymbol{\zeta} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix}. \tag{5.2}$$

In this expression the discrete counterparts of the continuous operators can be recognized, these are

$$-\nabla^2 \to \mathbf{L}, \tilde{\mathbf{L}}, \tilde{\mathbf{L}}^*, \quad \beta\frac{\partial}{\partial x} \to \mathbf{C}, \quad 1 \to \mathbf{M}.$$

The matrices $\mathbf{L}$, $\tilde{\mathbf{L}}$, and $\tilde{\mathbf{L}}^*$ differ due to the incorporation of the no-normal-flow boundary conditions. The mass matrix $\mathbf{M}$ is lumped and, therefore, a diagonal matrix.

The physical parameters are chosen as in Section 7.1 in [25] except for the viscosity parameter α which is 10^5 in this experiment. We have given a contour plot of the stream-function ψ in Figure 3 for these parameters. The precision is set to two degrees which results in a matrix of dimension 26455.

5.2 The Schur complement systems

From equation (5.2) we can eliminate either $\boldsymbol{\psi}$, which gives the Schur complement for $\boldsymbol{\zeta}$:

$$(\mathbf{M} + \alpha\tilde{\mathbf{L}}^*(r\mathbf{L} - \mathbf{C})^{-1}\tilde{\mathbf{L}})\boldsymbol{\zeta} = \tilde{\mathbf{L}}^*(r\mathbf{L} - \mathbf{C})^{-1}\mathbf{f}, \tag{5.3}$$

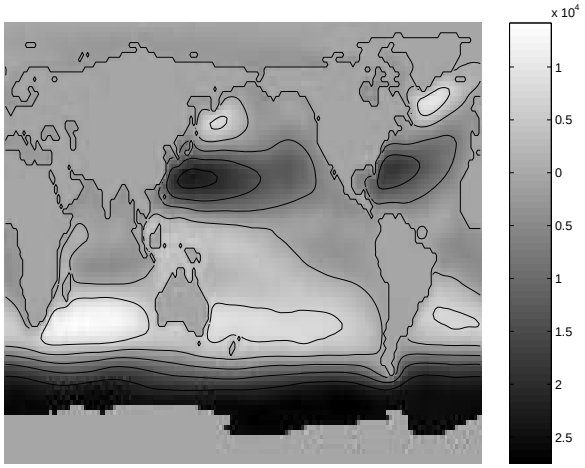

Figure 3: Stream-function, contour/density plot.

or we can eliminate $\boldsymbol{\zeta}$, which gives the Schur complement for $\boldsymbol{\psi}$:

$$((r\mathbf{L} - \mathbf{C}) + \alpha\tilde{\mathbf{L}}\mathbf{M}^{-1}\tilde{\mathbf{L}}^*)\boldsymbol{\psi} = \mathbf{f}. \tag{5.4}$$

The equation for the stream-function has the obvious advantage that, since $\mathbf{M}$ is diagonal and hence trivially invertible, operations with the Schur complement matrix $(r\mathbf{L}-\mathbf{C})+\alpha\tilde{\mathbf{L}}\mathbf{M}^{-1}\tilde{\mathbf{L}}^*$ are relatively cheap. This in contrast to the equation for the vorticity (5.3), where operations with the Schur complement require the solution of a linear system with the matrix $r\mathbf{L} - \mathbf{C}$.

There are, however, reasons why it can be preferable to solve (5.3) instead of (5.4). The Schur complement for the stream-function can be considerably worse conditioned than the Schur complement for the vorticity, in particular if α is large, or if the mesh size h is small. The ill-conditioning of the stream-function Schur complement is caused by the term $\tilde{\mathbf{L}}\mathbf{M}^{-1}\tilde{\mathbf{L}}^*$, which is a discretized biharmonic operator, of which the condition number is $\mathcal{O}(h^{-4})$. In practice it is very difficult to derive effective preconditioners for this operator. On the other hand, the diagonal matrix $\mathbf{M}$ turns out to be a very effective preconditioner for (5.3) and $r\mathbf{L} - \mathbf{C}$ is a discretized convection-diffusion operator for which also reasonably effective preconditioners are readily available. The smaller number of iterations for solving (5.3) in combination with the existence of preconditioners for the convection-diffusion operator may outweigh the extra work for the more costly matrix-vector products. This will be illustrated with the numerical experiments that are described below.

Another advantage of solving for the vorticity is that, since $\mathbf{M}$ is trivially invertible, we can construct solutions $\boldsymbol{\zeta}$ for various values of α by constructing the Krylov subspace only once using ideas from the so-called class of *multi-shift* Krylov subspace methods, e.g., [6] and [19, Section 10].

5.3 Numerical results for the vorticity Schur complement

If (5.3) is solved using a Krylov subspace method then a system with discrete convection-diffusion operator $r\mathbf{L} - \mathbf{C}$ has to be solved for every matrix-vector product with the Schur complement. In our experiments this was done using Bi-CGSTAB with an incomplete LU preconditioner [24]. The Bi-CGSTAB method was terminated when a relative residual precision was achieved of η. Note that it follows from the analysis in [19, Section 8] that this does

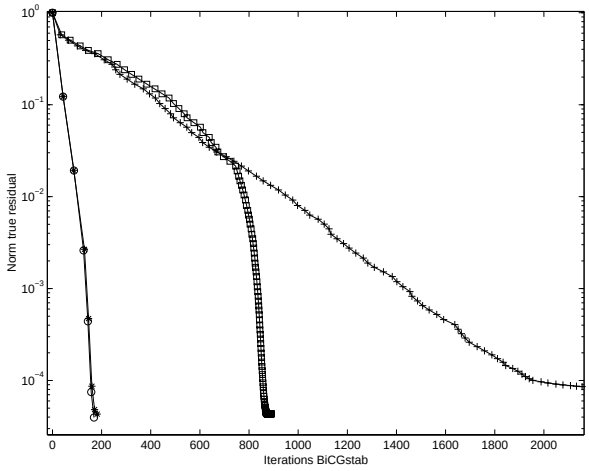

Figure 4: Norm true residual as function of the total number of iterations Bi-CGSTAB for the inexact GMRES method with fixed precision $\epsilon = 10^{-6}$ (+), relaxed GMRES method with $\epsilon = 10^{-7}$ (squares) and relaxed flexible GMRES preconditioned with relaxed GMRES set to a precision 0.1 (*) and the same for GMRESR in outer iteration (o).

not guarantee a relative error of η for the matrix-vector product and ideally an additional factor should be taken into account.

We aim for a residual precision of about 10^{-5} and consider a few different approaches for solving (5.3). In order to be able to make a fair comparison between different methods, for all methods the parameter ϵ was empirically chosen such that the methods achieve about the same accuracy. For the first method we have applied the inexact (full) GMRES method with the relaxation strategy of Bouras and Frayssé (2.2). As a preconditioner we have used the diagonal matrix $\mathbf{M}$. This preconditioner corrects for the adverse scaling effects introduced by the use of spherical coordinates and becomes increasingly more effective for smaller α. The results for this strategy are plotted in Figure 4. On the horizontal axis the total number of iterations with Bi-CGSTAB is given as a measure for the amount of work spent in the matrix-vector product. The vertical axis gives the norm of the true residual. The number of GMRES iterations is large for this problem, about 130, which limits the precision of the inexact Krylov method because of the accumulation of the errors in the matrix-vector product. For this reason we have chosen the empirical value $\epsilon = 10^{-7}$.

The convergence of GMRES is linear for the above example. The gain by applying the relaxation strategy is about a factor of two: the number of iterations Bi-CGSTAB drops from 2000 to about 900. This experimentally observed gain is consistent with the theoretical prediction for the gain that is presented in Section 3 (although the number of necessary Bi-CGSTAB iterations is not proportional to $-\log(\eta)$). As a consequence of the relaxation strategy, the number of Bi-CGSTAB iterations drops from 38.5 for the initial GMRES iterations to 0.5 for the last.

The alternative is to use the method from Section 4.2, that is, we precondition (inexact) flexible GMRES with an inexact GMRES method set to a precision of $\xi = 10^{-1}$. For the GMRES methods in the outer and inner iteration we have both used the Bouras-Frayssé relaxation strategy (2.2) with ϵ respectively given by 10^{-5} and 10^{-1}. The results are given in Figure 4 where every '*' corresponds to the true residual in one step of flexible GMRES

j	Tolerance (η_{j-1})	Bi-CGSTAB iterations	$\|\mathbf{b} - \mathbf{A}\mathbf{x}_j\|_2$	$\|\mathbf{r}_j\|_2$
1	$1.0 \cdot 10^{-5}$	35.5	$1.2 \cdot 10^{-1}$	$1.2 \cdot 10^{-1}$
2	$8.2 \cdot 10^{-5}$	33.5	$1.9 \cdot 10^{-2}$	$1.9 \cdot 10^{-2}$
3	$5.2 \cdot 10^{-4}$	28.5	$2.6 \cdot 10^{-3}$	$2.6 \cdot 10^{-3}$
4	$3.9 \cdot 10^{-3}$	7	$4.4 \cdot 10^{-4}$	$4.3 \cdot 10^{-4}$
5	$2.3 \cdot 10^{-2}$	3	$7.5 \cdot 10^{-5}$	$6.8 \cdot 10^{-5}$
6	$1.4 \cdot 10^{-1}$	1	$4.1 \cdot 10^{-5}$	$9.4 \cdot 10^{-6}$

Table 2: Numerical values concerning the relaxation process of the outer loop of GMRESR for iteration number j.

in the outer iteration. For the nested method only a few outer iterations are necessary and therefore the residual gap is less contaminated by errors in the matrix-vector product which results in a final precision of almost 10^{-5}. The large number of matrix-vector products and the accumulation of the errors now manifests itself in the inner iterations: we do not achieve a residual reduction of 10^{-1} in every outer step. In this picture we have also included the results for a nesting strategy with GMRESR in the outer iteration, the size of the true residuals is indicated with 'o'. The difference between the convergence of flexible GMRES and GMRESR is very small, which confirms the remarks made at the end of the Subsections 4.2 and 4.3.

To further illustrate the relaxation process in the outer iteration we have tabulated in Table 2 some interesting numerical values for the GMRESR-method. For the flexible GMRES method we find very similar numbers, which we have not reported here. The table shows for step j the used (approximate) tolerance for the matrix-vector product η_{j-1}, the number of Bi-CGSTAB iterations to compute the matrix-vector product in step j of the outer iteration and the norm of the true and computed residual at the end of step j (recall that $\mathbf{b}$ is normalized). With discretization step sizes that are more relevant in practice, the norm of the true residual will not be known during the process. The results in this table clearly illustrate that the norm of the true residual stagnates, in contrast to the norm of the updated residual. Another noteworthy observation is the sharp decrease in the number of Bi-CGSTAB iterations if the required tolerance is relaxed from $5.2 \cdot 10^{-4}$ to $3.9 \cdot 10^{-3}$. This is explained by the typical convergence behavior for Bi-CGSTAB that we observed for this example, which exhibits a fast decrease of the residual norm during the first iterations followed by a phase of slow convergence. The transition between fast and slow convergence is typically when the norm of the scaled residual is $\mathcal{O}(10^{-3})$.

A direct consequence of this initial fast convergence behavior of Bi-CGSTAB is that half a Bi-CGSTAB iteration (this is one application of the ILU-preconditioner and one matrix-vector product) is sufficient to reduce the scaled residual norm to below 0.1, which is an upper bound on the criterion for Bi-CGSTAB in the inner-loop. As a result, there is no practical difference between using a relaxation strategy or a fixed precision for the inner-iteration in this example.

5.4 Numerical results for the stream-function Schur complement

Although it is outside the scope of this paper we give, to underline the relevance of solving (5.3) instead of (5.4), also numerical results for the iterative solution of the equation for the stream-function. The solution technique we have used is Bi-CGSTAB in combination with an ILU-preconditioner of $r\mathbf{L} - \mathbf{C}$. The system is solved to about the same precision (for

the stream-function) as is reached if first the equation for the vorticity is solved with one of the methods described above. The iterative solution of equation (5.4) requires about 1000 Bi-CGSTAB iterations.

If we take the number of Bi-CGSTAB iterations as measure for the amount of work we can conclude that the relaxed inner-outer schemes for (5.3) are far more efficient than Bi-CGSTAB for (5.4). Less than 200 Bi-CGSTAB iterations are needed for the relaxed nested schemes for (5.3), while 5 times as many Bi-CGSTAB iterations are needed for solving (5.4). The comparison gives only an indication, since it neglects the overhead for GMRESR (or flexible GMRES), and the matrix-vector multiplications for solving (5.4) are more expensive.

6 Conclusions

In this paper we have analyzed strategies for controlling the accuracy of approximate matrix-vector products in the context of nested Krylov methods. We have demonstrated the benefits of nesting inexact Krylov subspace methods for a Schur complement system that occurs in a model that describes global ocean circulation. As part of the multiplication of the Schur complement with a vector, a linear system has to be solved, for which we used Bi-CGSTAB. The reduction in the total number of Bi-CGSTAB iterations for the nested Krylov methods flexible GMRES and GMRESR, is very large in comparison with normal GMRES. This can (partly) be explained by the fact that the matrix-vector products in the inner iterations can be evaluated to much lower precision than the ones in the outer loop.

Our approach in this paper was motivated by practical considerations and justified with many theoretical results. Although it is reasonably well understood why a relaxation strategy works, there still remain important questions to be answered. Among these we mention the fact that it is not well known how the relaxation method influences the convergence.

Acknowledgments

We thank Luc Giraud for pointing out and providing us with a copy of [16] and Henk van der Vorst for bringing [26] to our attention. J. van den Eshof was financially supported by the Dutch scientific organization (NWO) through project 613.002.035. Part of this research was done during a visit of the first author to CERFACS.

A Appendix

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 (4.9). 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 GCR 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 $\vec{\gamma}_k = (\gamma_0, \ldots, \gamma_k)^* \in \mathbb{R}^{k+1}$ such that $\vec{\gamma}_k^* \underline{T}_k = \vec{0}^*$ and $e_1^* \vec{\gamma}_k = 1$. It was shown in [22] that

$$\mathbf{r}_k = \|\vec{\gamma}_k\|_2^{-2} \mathbf{V}_{k+1} \vec{\gamma}_k \quad \text{and} \quad \mathbf{r}_k^{\mathrm{F}} = \gamma_k^{-1} \mathbf{V}_{k+1} e_{k+1}, \tag{A.1}$$

where $\mathbf{r}_k$ and $\mathbf{r}_k^{\mathrm{F}}$ are, respectively, the residuals of the corresponding GMRES and FOM process. The relation between the vector $\vec{\gamma}_k$ and the residuals can be expressed for the residuals $\mathbf{r}_j$ with $j = 0, \ldots, k-1$ by the relation

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

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

Lemma A.1. *For exact GCR without preconditioning we have that*

$$S_k^{-1} e_1 = (\mathbf{A}\mathbf{R}_k)^\dagger \mathbf{r}_0 = \Theta_k^2 \Upsilon_k^{-1} \underline{T}_k^\dagger e_1 \quad \textit{and} \quad e_{j+1}^* \Theta_k^2 \Upsilon_k^{-1} = \frac{\|\vec{\gamma}_j\|_2^2}{\gamma_j} e_{j+1}^* - \frac{\|\vec{\gamma}_j\|_2^2}{\gamma_{j+1}} e_{j+2}^*. \tag{A.2}$$

If $\mathbf{A} = \mathbf{A}^*$ *then*

$$e_{j+1}^* \Theta_k^2 \Upsilon_k^{-1} = \frac{\|\vec{\gamma}_j\|_2^2}{\vec{\gamma}_j^* T_k \vec{\gamma}_j} \vec{\gamma}_j^* \underline{T}_k \tag{A.3}$$

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

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

The second equality in (A.2) follows using the observation that $\Upsilon_k = \mathrm{diag}(\vec{\gamma}_{k-1}) J_k^{-1}$ where J_k is lower bidiagonal with -1 on its subdiagonal and 1 on its diagonal. The statement for $\mathbf{A} = \mathbf{A}^*$ follows from the observation that in this case the minimal residuals form an $\mathbf{A}$-orthogonal basis, or equivalently, $\Upsilon_k^* T_k \Upsilon = \Delta_k$ where Δ_k is diagonal with $\Delta_{j+1,j+1} = \vec{\gamma}_j^* T_k \vec{\gamma}_j$ and using that $\vec{\gamma}_j^* T_k = \vec{\gamma}_j^* \underline{T}_k$. □

Using this lemma we get for $\mathbf{A} = \mathbf{A}^* > 0$:

$$|e_{j+1}^* S_k^{-1} e_1| = |\frac{\vec{\gamma}_j^* \underline{T}_k \underline{T}_k^\dagger \vec{\gamma}_j}{\vec{\gamma}_j^* T_k \vec{\gamma}_j}| \leq \|\mathbf{A}^{-1}\|_2.$$

This is equal to our already presented bound in Section 2 for CR. 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 the presented bounds in Section 2 (or the equivalent ones from [22, 19]). For general matrices $\mathbf{A}$ this gives

$$|e_{j+1}^* S_k^{-1} e_1| \leq \|\underline{T}_k^\dagger\|_2 \left(\frac{\|\vec{\gamma}_j\|_2}{|\gamma_j|} + \frac{\|\vec{\gamma}_j\|_2}{|\gamma_{j+1}|} \frac{\|\vec{\gamma}_j\|_2}{\|\vec{\gamma}_{j+1}\|_2} \right) \leq \|\mathbf{A}^{-1}\|_2 \left(\frac{\|\mathbf{r}_j^{\mathrm{F}}\|_2}{\|\mathbf{r}_j\|_2} + \frac{\|\mathbf{r}_{j+1}^{\mathrm{F}}\|_2}{\|\mathbf{r}_j\|_2} \frac{\|\mathbf{r}_{j+1}\|_2}{\|\mathbf{r}_j\|_2} \right).$$

In the last inequality we have used (A.1).

References

[1] Ronald F. Boisvert, Roldan Pozo, Karin Remington, Richard Barrett, and Jack J. Dongarra, *The Matrix Market: A web resource for test matrix collections*, Quality of Numerical Software, Assessment and Enhancement (London) (Ronald F. Boisvert, ed.), Chapman & Hall, 1997, pp. 125–137.

[2] 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.

[3] 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.

[4] Peter N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Stat. Comput. **12** (1991), no. 1, 58–78. MR 92e:65035

[5] B. Carpentieri, *Sparse preconditioners for dense complex linear systems in electromagnetic applications*, Ph.D. dissertation, INPT, April 2002, TH/PA/02/48.

[6] Biswa Nath Datta and Youcef Saad, *Arnoldi methods for large Sylvester-like observer matrix equations, and an associated algorithm for partial spectrum assignment*, Linear Algebra Appl. **154/156** (1991), 225–244. MR 92b:65032

[7] Kees Dekker, *Partitioned-GMRES in domain decomposition with approximate subdomain solution*, BIT **41** (2001), no. 5, 924–935.

[8] Stanley C. Eisenstat, Howard C. Elman, and Martin H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal. **20** (1983), no. 2, 345–357. MR 84h:65030

[9] Paola Favati, Grazia Lotti, Ornella Menchi, and Francesco Romani, *Solution of infinite linear systems by automatic adaptive iterations*, Linear Algebra Appl. **318** (2000), no. 1-3, 209–225. MR 2001e:65048

[10] L. Giusti, C. Hoelbling, M. Lüscher, and H. Wittig, *Numerical techniques for lattice QCD in the ϵ–regime*, Preprint hep-lat/0212012, 2002.

[11] Gene H. Golub and Michael L. Overton, *The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems*, Numer. Math. **53** (1988), no. 5, 571–593. MR 90b:65054

[12] Gene H. Golub, Zhenyue Zhang, and Hongyuan Zha, *Large sparse symmetric eigenvalue problems with homogeneous linear constraints: the Lanczos process with inner-outer iterations*, Linear Algebra Appl. **309** (2000), no. 1-3, 289–306. MR 2001e:65060

[13] Martin H. Gutknecht and Miroslav Rozložnik, *Residual smoothing techniques: do they improve the limiting accuracy of iterative solvers?*, BIT **41** (2001), no. 1, 86–114. MR 2002b:65048

[14] P. Hernández, K. Jansen, and L. Lellouch, *A numerical treatment of Neuberger's lattice Dirac operator*, Numerical challenges in Lattice Quantum Chromodynamics (Berlin) (A. Frommer, Th. Lippert, B. Medeke, and K. Schilling, eds.), Springer-Verlag, 2000, pp. 29–39.

[15] Nicholas J. Higham, *Accuracy and stability of numerical algorithms*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. MR 97a:65047

[16] Katherine Mer-NKonga and Francis Collino, *The fast multipole method applied to a mixed integral system for time-harmonic Maxwell's equations*, Tech. report, 2002.

[17] H. Neuberger, *Overlap Dirac operator*, Numerical challenges in Lattice Quantum Chromodynamics (Berlin) (A. Frommer, Th. Lippert, B. Medeke, and K. Schilling, eds.), Springer-Verlag, 2000.

[18] Youcef Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput. **14** (1993), no. 2, 461–469. MR 1204241

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

[20] Gerard L. G. Sleijpen, Henk A. van der Vorst, and Diederik R. Fokkema, *BiCGstab(ℓ) and other hybrid Bi-CG methods*, Numer. Algorithms **7** (1994), no. 1, 75–109. MR 95d:65030

[21] J. van den Eshof, A. Frommer, Th. Lippert, K. Schilling, and H.A. van de Vorst, *Numerical methods for the QCD overlap operator: I. sign-function and error bounds*, Comput. Phys. Comm. **146** (2002), 203–224.

[22] 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, Submitted.

[23] H. A. van der Vorst and C. Vuik, *GMRESR: a family of nested GMRES methods*, Numer. Linear Algebra Appl. **1** (1994), no. 4, 369–386. MR 95j:65034

[24] 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** (1992), no. 2, 631–644. MR 92j:65048

[25] M. B. van Gijzen, C. B. Vreugdenhil, and H. Oksuzoglu, *The finite element discretization for stream-function problems on multiply connected domains*, J. Comput. Phys. **140** (1998), no. 1, 30–46. MR 98k:76088

[26] C. Vuik, *New insights in GMRES-like methods with variable preconditioners*, J. Comput. Appl. Math. **61** (1995), no. 2, 189–204. MR 96h:65054

[27] J. Warsa, M. Benzi, T. Wareing, and J. Morel, *Preconditioning a mixed discontinuous finite element method for radiation diffusion*, Tech. report, 2002, To appear in Numerical Linear Algebra with Applications.