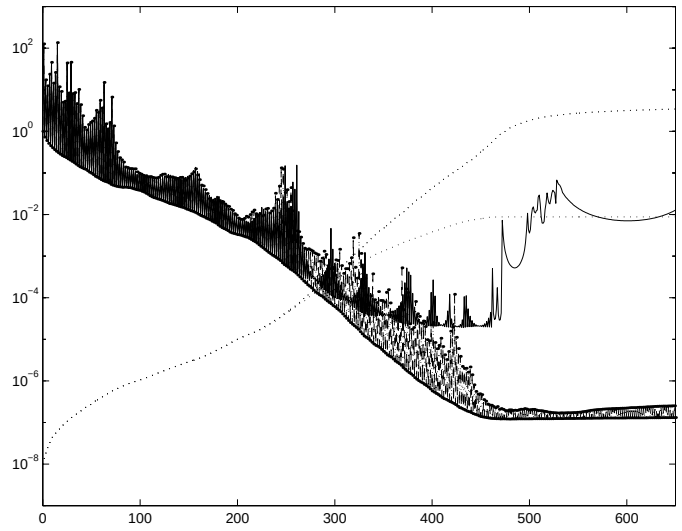

Universiteit Utrecht

Department
of Mathematics

Inexact Krylov subspace methods for linear systems

by

Jasper van den Eshof and Gerard L.G. Sleijpen

Preprint

nr. 1224

February, 2002

INEXACT KRYLOV SUBSPACE METHODS FOR LINEAR SYSTEMS*

JASPER VAN DEN ESHOF† AND GERARD L.G. SLEIJPEN*

Abstract. There is a class of linear problems for which the computation of the matrix-vector product is very expensive since a time consuming approximation method is necessary to compute it with some prescribed relative precision. In this paper we investigate the effect of an approximately computed matrix-vector product on the convergence and accuracy of several Krylov subspace solvers. The obtained insight is used to tune the precision of the matrix-vector product in every iteration so that an overall efficient process is obtained. This gives the empirical relaxation strategy of Bouras and Frayssé proposed in [2]. These strategies can lead to considerable savings over the standard approach of using a fixed relative precision for the matrix-vector product in every step. We will argue that the success of a relaxation strategy depends on the underlying way the Krylov subspace is constructed and not on the optimality properties for the residuals. Our analysis leads to an improved version of a strategy of Bouras, Frayssé, and Giraud [3] for the Conjugate Gradient method in case of Hermitian indefinite matrices.

1. Introduction. In Quantum Chromodynamics (QCD) [7], the overlap formulation has initiated a lot of research in solving linear systems of the form

$$(r\mathbf{\Gamma_5} + \mathrm{sign}(\mathbf{Q}))\mathbf{x} = \mathbf{b} \quad (r \geq 1), \tag{1.1}$$

where $\mathbf{Q}$ and $\mathbf{\Gamma_5}$ are sparse Hermitian indefinite matrices, and $\mathrm{sign}(t)$ is the standard sign-function. Thus, $\mathrm{sign}(\mathbf{Q})$ is essentially the matrix $\mathbf{Q}$ with all positive eigenvalues replaced by one and all negative eigenvalues by minus one. Therefore, the system in (1.1) is dense. Realistic simulations require in the order of one to ten million unknowns.

Usually, Equation (1.1) is solved with a standard Krylov subspace method for linear systems, for example the Conjugate Gradient method (since this matrix is Hermitian). In every step some method is required to compute the product of $\mathrm{sign}(\mathbf{Q})$ and a vector. The usual approach is to construct some polynomial approximation for the sign-function, for example using a Lanczos approximation. For an overview and comparison of methods used in this context we refer to [27].

It is obvious that the accurate computation of the matrix-vector product can be quite time consuming if done to high precision. On the other hand, the accuracy of the matrix-vector product has influence on the Krylov subspace method used for solving the linear system (i.e., the outer iteration). In this paper we investigate the influence of an approximately computed or inexact matrix-vector product on the convergence and accuracy of various Krylov subspace methods. This should lead to a further understanding of, for example, the *relaxation strategy* for the accuracy of the matrix-vector product as introduced by Bouras et al. [2, 3]. For example for GMRES, they propose to compute the matrix-vector product with a precision proportional to the inverse of the norm of the current residual. When the residual decreases the demands on the quality of the computed matrix-vector product are relaxed, thus explaining the term relaxation.

When thinking of inexact Newton methods, e.g., [5] the success of such relaxation strategies for Krylov subspace methods seems puzzling and further analysis is

*version March 5, 2002

† 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.

The research of the first author was financially supported by the Dutch Scientific Organization (NWO), under project number 613.002.035.

necessary on this topic. This is the subject of this paper. We also like to refer to an independent forthcoming paper by Simoncini and Szyld on this subject [23]. Their approach is based on an orthogonality condition, but the fundamental ideas seem to be related. The perhaps counter-intuitive phenomenon that an accurate matrix-vector product is needed in the beginning of the iterative process, instead of at the final iterations has also been observed and analyzed for the Lanczos method for the eigenvalue problem [12].

In this paper we consider the effect of perturbations on the matrix-vector product for various Krylov subspace solvers. This problem is related to rounding error analysis of Krylov subspace methods since in this situations an inexact matrix-vector product is one source of errors. In our analysis we will use an approved method from this area: we try to bound the *residual gap* and separately analyze the behavior of the computed residuals (although this is only possible in a few special cases). The usual way for bounding the gap is based on an inspection of the recursions, e.g., [24, 14, 19, 18]. Our approach differs from the analysis in these papers in the sense that our analysis is based on properties of the upper Hessenberg matrix that arises in the matrix formulation of the Krylov subspace method. Where possible we point out the differences with techniques used in literature and discuss implications for rounding error analysis.

Another related problem is when a variable preconditioner is used in the Krylov subspace method. See [9, 22, 28, 8, 11] for some results.

The outline of this paper is as follows. In Sections 2 and 3 we setup the framework that we need in the rest of this paper. We give an expression for the residual gap for a general Krylov subspace method in Section 3. This general expression is exploited in the remainder of this paper, starting with Richardson iteration in Section 4 and Chebyshev iteration in Section 5. The Conjugate Gradient method is the subject of Section 6. The focus in that section will be mainly on indefinite systems. Inexact GMRES and FOM for general matrices are treated in Section 7 and we conclude with some numerical experiments in Section 8.

2. Krylov subspace methods. This paper is concerned with the approximate solution of the $n \times n$ linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \quad \text{with} \quad \|\mathbf{b}\|_2 = 1. \tag{2.1}$$

Before we continue we define some notation. The vector e_k denotes the kth standard basis vector, i.e., $(e_k)_j = 0$ for all $j \neq k$ and $(e_k)_k = 1$. Furthermore, $\vec{1}$ is the vector with all components one and, similarly, $\vec{0}$ is the vector with all components zero. The dimension of these vectors should be apparent from the context. We warn the reader for some unconventional notation. If we apply a matrix with k columns to an ℓ-vector with $\ell \leq k$, then we assume the vector to be expanded with zeros if necessary (we do the same with other operations and equalities). Finally, we use bold capital letters to denote matrices of length n and use small bold capitals to denote the columns of these matrices where the subscript denotes the column number (starting with 0), so for example, $\mathbf{v}_0 = \mathbf{V}e_1$. The zero vector of length n is denoted by $\mathbf{0}$.

The notion of a *Krylov subspace* plays an important role in the understanding of a large class of iterative methods for solving (2.1). The Krylov subspace of order k is defined as

$$\mathcal{K}_k \equiv \mathcal{K}_k(\mathbf{A}, \mathbf{b}) \equiv \text{span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{k-1}\mathbf{b}\}. \tag{2.2}$$

In this paper we concentrate on iterative methods for which the residual $\mathbf{r}_j$ in step j belongs to the space $\mathcal{K}_{j+1}$ and $\mathbf{r}_0 = \mathbf{b}$. Because $\mathcal{K}_j \subseteq \mathcal{K}_{j+1}$, the residuals provide a sequence that after k steps of the subspace method can be summarized by the following matrix relation

$$\mathbf{A}\mathbf{R}_k = \mathbf{R}_{k+1}\underline{S}_k, \quad \mathbf{R}_k e_1 = \mathbf{b}, \tag{2.3}$$

where $\mathbf{R}_k$ is an n by k matrix with as jth column $\mathbf{r}_{j-1}$, and $\underline{S}_k$ is a $k+1$ by k upper Hessenberg matrix such that $\vec{1}^*\underline{S}_k = \vec{0}^*$.

The vector $\mathbf{r}_j$ is a residual corresponding to some approximate solutions in the Krylov subspace $\mathcal{K}_j$. Indeed, if S_j denotes the matrix $\underline{S}_j$ from which the last row is dropped, then, if S_j is invertible, we have with $\beta \equiv e_{j+1}^*\underline{S}_j e_j$,

$$\vec{0}^* = \vec{1}^*\underline{S}_j = \vec{1}^* S_j + \beta e_j^* \quad \Rightarrow \quad \beta e_j^* S_j^{-1} = -\vec{1}^*,$$

and

$$\underline{S}_j S_j^{-1} e_1 = (S_j + \beta e_{j+1} e_j^*) S_j^{-1} e_1 = e_1 - e_{j+1}. \tag{2.4}$$

Now, let

$$\mathbf{x}_j \equiv \mathbf{R}_k (S_j^{-1} e_1). \tag{2.5}$$

Then we get using (2.3) and (2.4) that

$$\begin{aligned} \mathbf{b} - \mathbf{A}\mathbf{x}_j &= \mathbf{b} - \mathbf{A}\mathbf{R}_k(S_j^{-1}e_1) = \mathbf{b} - \mathbf{A}\mathbf{R}_j(S_j^{-1}e_1) = \mathbf{b} - \mathbf{R}_{j+1}(\underline{S}_j S_j^{-1} e_1) \\ &= \mathbf{b} - \mathbf{R}_{j+1}(e_1 - e_{j+1}) = \mathbf{b} - (\mathbf{r}_0 - \mathbf{r}_j) = \mathbf{r}_j, \end{aligned}$$

and we have that $\mathbf{r}_j = \mathbf{b} - \mathbf{A}\mathbf{x}_j$. We see that if $\mathbf{x}_j \equiv \mathbf{R}_k(S_j^{-1}e_1)$ then the iterate $\mathbf{x}_j$ is *consistent* with the residual $\mathbf{r}_j$.

A recursion for the iterates $\mathbf{x}_j$ follows by substituting $\mathbf{R}_k = \mathbf{b}\vec{1}^* - \mathbf{A}\mathbf{X}_k$ in (2.3); this gives

$$-\mathbf{R}_k = \mathbf{X}_{k+1}\underline{S}_k, \quad \mathbf{X}_k e_1 = \mathbf{0}. \tag{2.6}$$

Some Krylov subspace methods use the recursions in (2.3) or (2.6) in a practical way, for example the Chebyshev method uses (2.6) for constructing the iterates $\mathbf{x}_j$.

We saw that for the $\mathbf{r}_j$ to be a residual corresponding to an approximate solution from $\mathcal{K}_j$ it is sufficient that $\vec{1}^*\underline{S}_k = \vec{0}^*$ and $\underline{S}_k$ is upper Hessenberg. This is also a necessary condition, see [17, Section 4.4].

We now summarize some properties and relations that we use in the remainder of this paper.

LEMMA 2.1. *If the matrix S_j is invertible for $j \le k$, then the LU-decomposition of S_k and the one of $\underline{S}_k$ exists and is unique. Furthermore,*

$$S_k = J_k U_k \quad \textit{and} \quad \underline{S}_k = \underline{J}_k U_k, \tag{2.7}$$

where $\underline{J}_k$ is lower bidiagonal with $(\underline{J}_k)_{j,j} = 1$ and $(\underline{J}_k)_{j+1,j} = -1$ and and U_k is upper triangular with $(U_k)_{i,j} = \sum_{l=1}^{i}(\underline{S}_k)_{l,j}$ for $i \le j$.

Proof. The existence and uniqueness of the LU-decomposition of S_k follows from [10, Theorem 3.2.1]. The matrix J_k^{-1} is lower triangular with all components one.

Therefore, it follows that $J_k^{-1} S_k = U_k$. This proves the first equality in (2.7). The second equality follows by checking that

$$\underline{J}_k U_k = (J_k - e_{k+1} e_k^*) U_k = S_k - e_{k+1} e_k^* U_k = \underline{S}_k. \quad \square$$

The LU-decomposition of $\underline{S}_k$ is used in the construction of some Krylov subspace methods. We will return to this later.

In some cases it is convenient to consider the Krylov decomposition

$$\mathbf{A}\mathbf{C}_k = \mathbf{C}_{k+1} \underline{T}_k, \quad \mathbf{C}_k e_1 = \mathbf{b}, \tag{2.8}$$

where $\mathbf{C}_k$ is an n by k matrix and $\underline{T}_k$ is a $k+1$ by k upper Hessenberg matrix. From this relation different residual sequences (2.3) can be derived depending on the required properties for the $\mathbf{r}_j$. In order to continue our discussion, we assume that $\underline{T}_k$ has full rank, and we define the $k+1$-vector $\vec{\gamma}_k$ as the vector such that $\vec{\gamma}_k^* \underline{T}_k = \vec{0}^*$ and $\vec{\gamma}_k^* = (1, \gamma_1, \ldots, \gamma_k)^*$. Notice that, due to the Hessenberg structure of $\underline{T}_k$, the elements γ_j can be computed using a simple and efficient recursion.

A simple way to derive a residual sequence is to put $\Gamma_k \equiv \mathrm{diag}(\vec{\gamma}_k)$; then we see that the matrices

$$\underline{S}_k \equiv \Gamma_{k+1} \underline{T}_k \Gamma_k^{-1} \quad \text{and} \quad \mathbf{R}_k \equiv \mathbf{C}_k \Gamma_k^{-1} \tag{2.9}$$

satisfy (2.3) (with, indeed, $\vec{1}^* \underline{S}_k = \vec{0}^*$). In this case the residual $\mathbf{r}_j$ is a multiple of the vector $\mathbf{c}_j$. Furthermore, if T_j is invertible, then we have for the residual

$$\mathbf{r}_j = \mathbf{c}_j / \gamma_j = \mathbf{C}_{j+1} (I - \underline{T}_j T_j^{-1}) e_1 = \mathbf{b} - \mathbf{A}\mathbf{C}_j T_j^{-1} e_1, \tag{2.10}$$

where we have used the following lemma and (2.8). (For ease of future reference, we formulate the lemma slightly more general than needed here.)

LEMMA 2.2. *Let $j \leq k$. Then*

$$e_1 - \underline{T}_k (\underline{T}_j^\dagger e_1) = \frac{\vec{\gamma}_j}{\|\vec{\gamma}_j\|_2^2} \quad \textit{and} \quad e_1 - \underline{T}_k (T_j^{-1} e_1) = \frac{e_{j+1}}{\gamma_j}, \tag{2.11}$$

where $\underline{T}_j^\dagger$ denotes the generalized inverse of $\underline{T}_j$ [10, Section 5.5.4]*, and where, for the second expression, T_j is assumed to be invertible.*

Proof. Note that $e_1 - \underline{T}_k (\underline{T}_j^\dagger e_1) = (I - \underline{T}_j \underline{T}_j^\dagger) e_1$. Since $I - \underline{T}_j \underline{T}_j^\dagger$ is the orthogonal projection on $\mathrm{Ker}(\underline{T}_j{}^*) = \mathrm{span}(\vec{\gamma}_j)$, we have that $I - \underline{T}_j \underline{T}_j^\dagger = \|\vec{\gamma}_j\|_2^{-2} \, \vec{\gamma}_j \, \vec{\gamma}_j^*$. This leads to the first expression in (2.11). The second expression follows from a combination of $e_1 - \underline{T}_k (T_j^{-1} e_1) = e_1 - \underline{T}_j (T_j^{-1} e_1) = e_1 - \Gamma_{j+1}^{-1} \underline{S}_j S_j^{-1} \Gamma_j e_1$ and (2.4). $\square$

The lemma also leads to an expression for residuals from an alternative construction:

$$\mathbf{r}_j = \mathbf{b} - \mathbf{A}\mathbf{C}_j \underline{T}_j^\dagger e_1 = \mathbf{C}_{j+1} (I - \underline{T}_j \underline{T}_j^\dagger) e_1 = \frac{1}{\|\vec{\gamma}_j\|_2^2} \mathbf{C}_{j+1} \vec{\gamma}_j. \tag{2.12}$$

If we define

$$\Upsilon_k \equiv [\vec{\gamma}_0, \ldots, \vec{\gamma}_{k-1}], \quad \Theta_k \equiv \mathrm{diag}(\|\vec{\gamma}_0\|_2, \ldots, \|\vec{\gamma}_{k-1}\|_2),$$

then we get

$$\underline{S}_k \equiv (\Upsilon_{k+1} \Theta_{k+1}^{-2})^{-1} \underline{T}_k (\Upsilon_k \Theta_k^{-2}) \quad \text{and} \quad \mathbf{R}_k \equiv \mathbf{C}_k (\Upsilon_k \Theta_k^{-2}). \tag{2.13}$$

It can be easily checked that $\vec{1}^* (\Upsilon_{k+1} \Theta_{k+1}^{-2})^{-1} = \vec{\gamma}_k^*$ and therefore $\vec{1}^* \underline{S}_k = \vec{0}^*$ and also the Hessenberg form is preserved. However, a tridiagonal matrix $\underline{T}_k$ does not guarantee that $\underline{S}_k$ is tridiagonal in general. However, it should be noted that the matrix $(\Upsilon_{k+1} \Theta_{k+1}^{-2})^{-1}$ can be decomposed into simple factors.

3. Inexact Krylov subspace methods. There is a class of applications for which it is very costly to compute the matrix-vector product to high precision. We assume that we are given some device $\mathcal{M}_\eta : \mathbb{C}^n \to \mathbb{C}^n$ with the property that

$$\mathcal{M}_\eta(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{g} \quad \text{with} \quad \|\mathbf{g}\|_2 \leq \eta \|\mathbf{A}\|_2 \, \|\mathbf{y}\|_2, \tag{3.1}$$

and the smaller η the more costly this operation becomes.

In step j of all the iterative methods, that we discuss, it it necessary to compute the product of the matrix $\mathbf{A}$ with some vector, say $\mathbf{y}$, which is done using $\mathcal{M}_\eta$. It is equivalent to consider the exact method where a perturbation $\mathbf{g}_{j-1}$ is added to the matrix-vector product in step j with $\|\mathbf{g}_{j-1}\|_2 \leq \eta_{j-1}\|\mathbf{A}\|_2 \, \|\mathbf{y}\|_2$.

Due to the existence of the $\mathbf{g}_{j-1}$, the space spanned by the residuals is in general not a Krylov subspace generated by $\mathbf{A}$ anymore. This has two consequences: the convergence behavior is altered, and the maximal attainable accuracy of the iterative method is limited. The central question is how large the perturbations can be if one is interested in a solution $\mathbf{x}_k$ such that $\|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|_2 = \mathcal{O}(\varepsilon)$ without altering the convergence behavior too much, or equivalently, how to pick η_{j-1} in step j.

In the remainder of this paper we will see that, for the methods that we consider, the residuals now satisfy the perturbed relation

$$\mathbf{A}\mathbf{R}_k + \mathbf{F}_k = \mathbf{R}_{k+1}\underline{S}_k, \quad \mathbf{R}_k e_1 = \mathbf{b}, \tag{3.2}$$

where the columns of $\mathbf{F}_k$ are a function of the perturbations $\mathbf{g}_j$. Furthermore, $\mathbf{x}_j$ still satisfies (2.5) (or equivalently (2.6)) because of the assumption of exact arithmetic. These two common properties allow a unified analysis.

The vector $\mathbf{r}_k$ is usually not a residual anymore for the approximate solution $\mathbf{x}_k$ due to the perturbation $\mathbf{F}_k$. Therefore, we will refer to the vector $\mathbf{r}_k$ as the *computed residual* in contrast to the *true residual* defined by $\mathbf{b} - \mathbf{A}\mathbf{x}_k$. The goal is to get an expression for $\mathbf{e}_k$, the difference between the computed residual and the true residual. If we, furthermore, show that the computed residuals become small with respect to the residual gap, then the ultimately attainable accuracy is essentially determined by this gap $\mathbf{e}_k$.

We define $\mathbf{E}_{k+1} \equiv \mathbf{R}_{k+1} - (\mathbf{b}\vec{1}^* - \mathbf{A}\mathbf{X}_{k+1})$. If we multiply $\mathbf{E}_{k+1}$ from the right with $\underline{S}_k$ and use (2.6) and (3.2), we get that $\mathbf{E}_{k+1}\underline{S}_k = \mathbf{F}_k$. Under our assumptions, we have that $\mathbf{e}_0 = \mathbf{E}_{k+1}e_1 = \mathbf{r}_0 - \mathbf{b} = \mathbf{0}$. Hence, using (2.4) we get

$$\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k) = \mathbf{e}_k = -\mathbf{E}_{k+1}\underline{S}_k S_k^{-1} e_1 = -\mathbf{F}_k S_k^{-1} e_1 = -\sum_{j=1}^{k} \mathbf{f}_{j-1} e_j^* S_k^{-1} e_1. \tag{3.3}$$

We see that the expression for the gap is a linear combination of the perturbations $\mathbf{f}_{j-1}$ with coefficients $-e_j^* S_k^{-1} e_1$. Our approach for bounding the gap $\mathbf{e}_k$ is based on using properties of the matrix $\underline{S}_k$. We will do this for various Krylov subspace methods in the remainder of this paper. Therefore, the following lemma is convenient and will be frequently used in the remainder of this paper.

LEMMA 3.1. *Assume that $j \leq k$. Let $\underline{S}_k = \Gamma_{k+1}\underline{T}_k\Gamma_k^{-1}$ as in (2.9), then*

$$e_j^* S_k^{-1} e_1 = \gamma_{j-1} e_j^* T_k^{-1} e_1 = e_j^* T_k^{-1} e_j,$$

and

$$|e_j^* \underline{T}_k^\dagger e_1| \leq \frac{\sigma_{\min}(\underline{T}_k)^{-1}}{\|\vec{\gamma}_{j-1}\|_2}, \quad |e_j^* T_k^{-1} e_1| \leq \sigma_{\min}(\underline{T}_k)^{-1} \left(\frac{1}{\|\vec{\gamma}_{j-1}\|_2} + \frac{1}{|\gamma_k|} \right). \tag{3.4}$$

Here $\sigma_{\min}(\underline{T}_k)$ is the smallest singular value of $\underline{T}_k$.

Proof. The relation $\underline{S}_k = \Gamma_{k+1}\underline{T}_k\Gamma_k^{-1}$ in (2.9) implies $e_j^* S_k^{-1} e_1 = \gamma_{j-1} e_j^* T_k^{-1} e_1$ and $e_j^* S_k^{-1} e_j = e_j^* T_k^{-1} e_j$. Using (2.4) we see that $e_j^* S_k^{-1} e_1 = e_j^* S_k^{-1}(e_1 - S_k(S_{j-1}^{-1} e_1)) = e_j^* S_k^{-1}(e_1 - \underline{S}_{j-1}(S_{j-1}^{-1} e_1)) = e_j^* S_k^{-1} e_j$.

To prove (3.4), we observe that $\underline{T}_k^\dagger \underline{T}_k$ is the identity on k-vectors if $\underline{T}_k$ is of rank k. Since $e_j^* \vec{y}_{j-1} = 0$ for any $j-1$-vector $\vec{y}_{j-1}$ we have that

$$\begin{aligned} e_j^* \underline{T}_k^\dagger e_1 &= e_j^* \underline{T}_k^\dagger (e_1 - \underline{T}_k \vec{y}_{j-1}) \quad \text{and} \\ e_j^* T_k^{-1} e_1 &= e_j^* \underline{T}_k^\dagger (e_1 - \underline{T}_k \vec{y}_{j-1}) + e_j^* \underline{T}_k^\dagger (\underline{T}_k (T_k^{-1} e_1) - e_1). \end{aligned}$$

With $\vec{y}_{j-1} = \underline{T}_{j-1}^\dagger e_1$ and $\vec{y}_{j-1} = T_{j-1}^{-1} e_1$, a combination with (2.11) leads to

$$e_j^* \underline{T}_k^\dagger e_1 = e_j^* \underline{T}_k^\dagger \frac{\vec{\gamma}_{j-1}}{\|\vec{\gamma}_{j-1}\|_2^2} = e_j^* \underline{T}_k^\dagger \frac{e_j}{\gamma_{j-1}} \text{ and } e_j^* T_k^{-1} e_1 = e_j^* \underline{T}_k^\dagger e_1 - e_j^* \underline{T}_k^\dagger \frac{e_{k+1}}{\gamma_k}. \tag{3.5}$$

and (3.4) easily follows. ∎

We expressed our estimates in terms of the smallest singular value of $\underline{T}_k$. This value depends monotonically (decreasing) on k, and $\sigma_{\min}(T_m) \leq \sigma_{\min}(\underline{T}_k)$ if $m > k$. The smallest singular value of T_k does not have this attractive property: even if T_m is well-conditioned, there may be a $k < m$ for which T_k is singular or nearly singular.

3.1. Relaxation strategies. In [2], Bouras and Frayssé showed experiments for GMRES with a relative precision η_j in step $j+1$ given by

$$\eta_j = \max\left\{ \frac{\varepsilon}{\|\mathbf{b} - \mathbf{A}\mathbf{x}_j\|_2}, \varepsilon \right\}. \tag{3.6}$$

An interesting property of this choice for η_j is that it requires very accurate matrix-vector products in the beginning of the process, and the precision is relaxed if the methods converges, i.e., the residuals become small. This justifies the term *relaxation strategy* as introduced in [2]. For an impressive list of numerical experiments, they observe that with (3.6) the GMRES method converges roughly as fast as the unperturbed version, despite the, sometimes large, perturbations. Furthermore, the norm of the true residual ($\|\mathbf{b} - \mathbf{A}\mathbf{x}_j\|_2$) seems to stagnate around a value of $\mathcal{O}(\varepsilon)$. Obviously, such a strategy can result in large savings in practical applications. The true residual is unfortunately in general not known, since this would require an exact matrix-vector product. The norm of the computed residual $\|\mathbf{r}_j\|_2$ can serve as an alternative in (3.6).

In the coming sections, we will analyze the effect of inexact matrix-vector products and relaxation strategies as in (3.6) for different Krylov subspace methods by writing the residual relation into the form (3.2) and by analyzing the residual gap using (3.3). If it is additionally shown that the computed residuals $\mathbf{r}_k$ become sufficiently small, then the residual gap will ultimately determine the attainable accuracy. The convergence of the computed residuals is a difficult topic that we can only analyze in some special cases. It should be noticed that for the applications that we have in mind the norm of the computed residuals can be efficiently monitored, while for the true residual or size of the residual gap, it is necessary to compute an accurate matrix-vector product which is not feasible.

For the analysis in this paper, we assume the use of exact arithmetic operations. This is a reasonable assumption, considering that in general the "error" in the matrix-vector product is much larger than machine precision, as in the QCD example (1.1)

mentioned in the beginning of Section 1, where the error in the matrix-vector product is an error resulting from the truncation of an approximation process for the matrix sign-function times a vector.

4. Inexact Richardson iteration. One of the simplest iterative method for linear systems is *Richardson iteration*, e.g., [15]. With a perturbed matrix-vector product, this method is described by the following recursion (with $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{r}_0 = \mathbf{b}$) for $j = 1, \ldots, k$

$$\mathbf{r}_j = \mathbf{r}_{j-1} - \alpha(\mathbf{A}\mathbf{r}_{j-1} + \mathbf{g}_{j-1}) \tag{4.1}$$

$$\mathbf{x}_j = \mathbf{x}_{j-1} + \alpha\mathbf{r}_{j-1} \tag{4.2}$$

and $\|\mathbf{g}_j\| \leq \eta_j \|\mathbf{A}\|_2 \|\mathbf{r}_j\|_2$. For simplicity we restrict our attention to symmetric positive definite matrices $\mathbf{A}$ with an optimal choice for α:

$$\alpha \equiv \frac{2}{\lambda_- + \lambda_+}, \tag{4.3}$$

where λ_- and λ_+ are, respectively the smallest and largest eigenvalue of $\mathbf{A}$.

For this method it is clear that after k steps of the method, the iterates satisfy (2.6) and the residuals satisfy (3.2) with $\mathbf{F}_k = \mathbf{G}_k$ and $\underline{S}_k = \underline{J}_k U_k$ with $U_k = \alpha^{-1} I$. Therefore, we can exploit (3.3) and get using $e_j^* S_k^{-1} e_1 = \alpha$ for the residual gap the following bound

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 = \|\sum_{j=1}^{k} \mathbf{f}_{j-1}\alpha\|_2 \leq \alpha\|\mathbf{A}\|_2 \sum_{j=0}^{k-1} \eta_j \|\mathbf{r}_j\|_2.$$

Recall that we are only interested in an approximate solution $\mathbf{x}_k$ with $\|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|_2 = \mathcal{O}(\varepsilon)$. This suggests to pick $\eta_j = \varepsilon/\|\mathbf{r}_j\|_2$ and we get using (4.3),

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 \leq \varepsilon k\alpha\|\mathbf{A}\|_2 = \varepsilon 2k \frac{\mathcal{C}(\mathbf{A})}{\mathcal{C}(\mathbf{A}) + 1} < \varepsilon 2k,$$

where $\mathcal{C}(\mathbf{A}) \equiv \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2$.

It remains to be shown that the computed residuals become sufficiently small. For inexact Richardson iteration we have the following result which even shows that the computed residuals become small at a speed comparable to the exact process.

LEMMA 4.1. *Let* $\overline{\mathbf{r}}_k$ *satisfy* (4.1) *with* $\eta_j = 0$ *and* $\mathbf{r}_k$ *with* $\eta_j = \varepsilon/\|\mathbf{r}_j\|_2$, *then*

$$\|\mathbf{r}_k - \overline{\mathbf{r}}_k\| \leq \varepsilon \mathcal{C}(\mathbf{A}).$$

Proof. The difference between the two residuals is given by

$$\mathbf{r}_k - \overline{\mathbf{r}}_k = (I - \alpha\mathbf{A})^k \mathbf{b} + \alpha \sum_{j=1}^{k} (I - \alpha\mathbf{A})^{k-j} \mathbf{f}_{j-1} - (I - \alpha\mathbf{A})^k \mathbf{b} = \alpha \sum_{j=1}^{k} (I - \alpha\mathbf{A})^{k-j} \mathbf{f}_{j-1}.$$

For $\eta_j = \varepsilon/\|\mathbf{r}_j\|_2$ we have $\|\mathbf{f}_j\|_2 \leq \eta_j \|\mathbf{A}\|_2 \|\mathbf{r}_j\|_2 = \varepsilon\|\mathbf{A}\|_2$, hence

$$\|\mathbf{r}_k - \overline{\mathbf{r}}_k\|_2 \leq |\alpha| \sum_{j=1}^{k} \|(I - \alpha\mathbf{A})\|_2^{k-j} \varepsilon \|\mathbf{A}\|_2 \leq \varepsilon \|\mathbf{A}\|_2 \|(\alpha\mathbf{A})^{-1}\|_2 |\alpha| = \varepsilon \mathcal{C}(\mathbf{A}). \quad \square$$

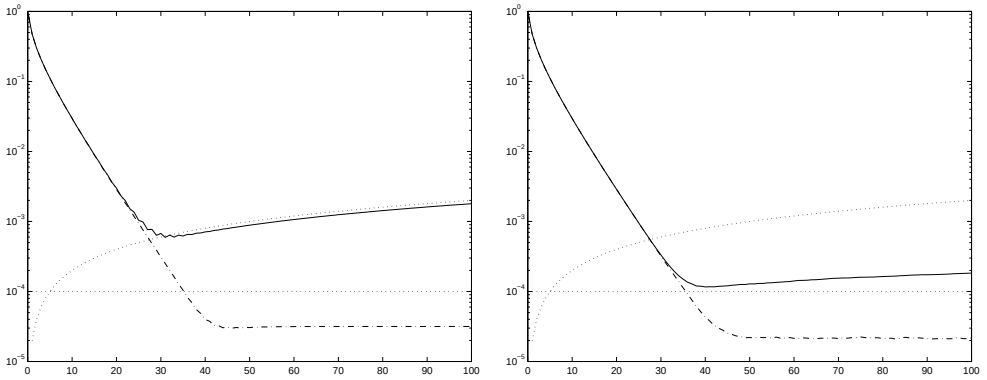

FIG. 4.1: Richardson iteration with $\eta_j = 10^{-5}/\|\mathbf{r}_j\|_2$, true residuals (—), norm computed residual (- ·) and the quantities $10^{-5}\mathcal{C}(\mathbf{A})$, $2k10^{-5}$ (both dotted). The matrix $\mathbf{A}$ has dimension 1000 and $\mathcal{C}(\mathbf{A}) = 10$. Left picture: errors have all components equal. Right picture: random errors.

Since $\bar{\mathbf{r}}_k$ will go to zero for $k \to \infty$, we expect the norm of $\mathbf{r}_k$ ultimately to stagnate at a level below $\varepsilon\mathcal{C}(\mathbf{A})$. This shows that the final residual precision is essentially determined by the residual gap. We give a simple illustration of this in Figure 4.1. We conclude that for Richardson iteration the required precision of the matrix-vector product can be relaxed with a strategy similar as the one proposed for GMRES in (3.6).

4.1. Discussion. One might remark that in practical applications the residual is not computed in an incremental fashion as in (4.1). However, incrementally computed residuals are important for a relaxation strategy to be successful. Furthermore, directly computed residuals are not necessarily more accurate even if using a fixed precision, i.e., $\eta_j = \eta$. In this case a direct computation of the $k+1$th residual yields

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 \leq \eta\|\mathbf{A}\|_2 \, \|\mathbf{x}_k\|_2 = \|(\eta\|\mathbf{A}\|_2\mathbf{R}_k)S_k^{-1}e_1\|_2,$$

whereas an expression for the recursively computed residual follows from (3.3)

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 = \|\mathbf{F}_k S_k^{-1}e_1\|_2.$$

Both $\mathbf{F}_k$ and $\eta\|\mathbf{A}\|_2\mathbf{R}_k$ have a $j+1$th column with a length smaller than $\eta\|\mathbf{A}\|_2 \, \|\mathbf{r}_j\|_2$. Hence, the difference in the upper bounds is determined by the mutual angle between the columns. In case the residuals change slowly and if the $\mathbf{f}_j$ are random, the recursively computed residual can be more accurate. Practical experiments confirm this, although the differences are small. Numerical experiments suggest that in the situation of only finite precision errors an incrementally computed residual is no longer necessarily more accurate than a directly computed residual as is often observed in practice.

5. Inexact Chebyshev iteration. A more advanced method than Richardson iteration is *Chebyshev iteration*, e.g., [10, Section 10.1.5],[6, Chapter 7]. It is more advanced than Richardson iteration in the sense that it employs a three-term recurrence for the residuals for faster convergence. For clarity and in order to establish notation, we start with a short derivation of Chebyshev iteration. Again we assume $\mathbf{A}$ to be symmetric positive definite.

We define $\phi(t) \equiv \alpha t - \beta$ as a function that maps the interval $[\lambda_-, \lambda_+]$ to the interval $[-1, 1]$, so (for example)

$$\alpha \equiv \frac{2}{\lambda_+ - \lambda_-}, \quad \beta \equiv \frac{\lambda_+ + \lambda_-}{\lambda_+ - \lambda_-}. \tag{5.1}$$

The main idea behind the Chebyshev method is to construct the residuals $\mathbf{r}_j$ as multiples of the vectors $\mathbf{c}_j = c_j(\phi(\mathbf{A}))\mathbf{b}$, where $c_j(t)$ is the Chebyshev polynomial of degree j, see for a definition [6, p. 4]. An efficient algorithm comes from the three-term recurrence for the Chebyshev polynomials

$$\mathbf{c}_j = 2\phi(\mathbf{A})\mathbf{c}_{j-1} - \mathbf{c}_{j-2}, \quad \text{with} \quad \mathbf{c}_0 = \mathbf{b},\ \mathbf{c}_1 = \phi(\mathbf{A})\mathbf{b},$$

which reads in matrix formulation for k steps,

$$\mathbf{A}\mathbf{C}_k = \mathbf{C}_k \underline{T}_k \quad \text{with} \quad \underline{T}_k \equiv \begin{bmatrix} \frac{\beta}{\alpha} & \frac{1}{2\alpha} & & \\ \frac{1}{\alpha} & \frac{\beta}{\alpha} & \frac{1}{2\alpha} & \\ & \frac{1}{2\alpha} & \ddots & \ddots \\ & & \ddots & \ddots \\ & & & \frac{1}{2\alpha} \end{bmatrix}. \tag{5.2}$$

Equations (2.3) and (2.9) now give a three-term recurrence for the residuals with $\gamma_j = c_j(\phi_j(0))$. A recursion for the approximate solutions $\mathbf{x}_j$ is given by (2.6). All together we have the following recursion for $j = 2, \ldots, k$ (which we state here for convenience)

$$\mathbf{r}_j = 2\alpha\frac{\gamma_{j-1}}{\gamma_j}(\mathbf{A}\mathbf{r}_{j-1} + \mathbf{g}_{j-1}) - 2\beta\frac{\gamma_{j-1}}{\gamma_j}\mathbf{r}_{j-1} - \frac{\gamma_{j-2}}{\gamma_j}\mathbf{r}_{j-2}, \tag{5.3}$$

$$\mathbf{x}_j = -2\alpha\frac{\gamma_{j-1}}{\gamma_j}\mathbf{r}_{j-1} - 2\beta\frac{\gamma_{j-1}}{\gamma_j}\mathbf{x}_{j-1} - \frac{\gamma_{j-2}}{\gamma_j}\mathbf{x}_{j-2}, \tag{5.4}$$

with $\mathbf{r}_0 = \mathbf{b}$, $\mathbf{r}_1 = \alpha\frac{\gamma_0}{\gamma_1}(\mathbf{A}\mathbf{r}_0 + \mathbf{g}_0) - \beta\frac{\gamma_0}{\gamma_1}\mathbf{r}_0$, $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{x}_1 = -\alpha\frac{\gamma_0}{\gamma_1}\mathbf{r}_0$. In this recursion we have already replaced the matrix-vector product in (5.3) with a perturbed version. It easily follows that the residuals for inexact Chebyshev satisfy (3.2) with $\mathbf{F}_k = \mathbf{G}_k$ and therefore $\|\mathbf{f}_j\|_2 \leq \eta_j \|\mathbf{A}\|_2 \|\mathbf{r}_j\|_2$.

In order to bound the residual gap with (3.3) we have to bound $e_j^* S_k^{-1} e_1$, this is done using the following lemma.

LEMMA 5.1. *Let T_k be as in (5.2), and let α and β be as (5.1). Then*

$$|e_j^* T_k^{-1} e_j| \leq \frac{2\alpha}{\sqrt{\beta^2 - 1}} = \frac{2}{\sqrt{\lambda_+\lambda_-}} = 2\frac{\sqrt{\mathcal{C}(\mathbf{A})}}{\|\mathbf{A}\|_2}. \tag{5.5}$$

Proof. The matrix T_k is given by $T_k = \frac{\beta}{\alpha}(I + \frac{1}{2\beta}\Delta)$, where Δ is the k by k matrix with zeros entries everywhere except at the positions $(i-1, i)$ and $(i, i-1)$, where it has the value one and the $(2,1)$ element is 2. To obtain the estimate for $e_j^* T_k^{-1} e_j$, we express $(I + \frac{1}{2\beta}\Delta)^{-1}$ as a Neumann series, and check that $e_j^* \Delta^{2i-1} e_j = 0$. With some effort it can be shown that $|e_j^* \Delta^{2i} e_j| \leq 2\frac{(2i)!}{(i!)^2}$ for all $i = 1, 2, \ldots$. Now use for $t = 1/\beta^2$ that

$$\frac{1}{\sqrt{1-t}} = \sum_{i=0}^{\infty} \frac{(2i)!}{(2^i i!)^2} t^i \quad \text{if} \quad |t| < 1.$$

This leads to the estimate in (5.5).
∎

A combination of Lemma 5.1, the relation $e_j^* S_k^{-1} e_1 = e_j^* T_k^{-1} e_j$ from Lemma 3.1, and (3.3) gives the following bound on the residual gap

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 \leq 2\sqrt{\mathcal{C}(\mathbf{A})}/\|\mathbf{A}\|_2 \sum_{j=0}^{k-1} \|\mathbf{f}_j\|_2 \leq 2\sqrt{\mathcal{C}(\mathbf{A})} \sum_{j=0}^{k-1} \eta_j \|\mathbf{r}_j\|_2.$$

Given the fact that we are interested in a residual precision of only $\mathcal{O}(\varepsilon)$ we propose the same relaxation strategy as for Richardson iteration in Section 4, i.e., pick $\eta_j = \varepsilon/\|\mathbf{r}_j\|_2$. The gap for this strategy can then be bounded as

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 \leq 2k\varepsilon\sqrt{\mathcal{C}(\mathbf{A})}. \tag{5.6}$$

The proposed relaxation strategy allows very large perturbations when the residuals are small. Nevertheless, the following lemma shows that also the convergence of the computed residuals for this strategy is close to that of the exact method. Furthermore, the computed residuals become in the end sufficiently small for (5.6) to be meaningful as measure for the attainable accuracy.

LEMMA 5.2. *Let* $\bar{\mathbf{r}}_k$ *satisfy* (5.3) *with* $\eta_j = 0$ *and* $\mathbf{r}_k$ *with* $\eta_j = \varepsilon/\|\mathbf{r}_j\|_2$. *Then*

$$\|\mathbf{r}_k - \bar{\mathbf{r}}_k\|_2 \leq \varepsilon(1 - |\gamma_k|^{-1})\mathcal{C}(\mathbf{A}).$$

Proof. If we subtract (2.3) from (3.2), then we get

$$\mathbf{A}(\mathbf{R}_k - \overline{\mathbf{R}}_k) + \mathbf{F}_k = (\mathbf{R}_{k+1} - \overline{\mathbf{R}}_{k+1})\underline{S}_k, \quad (\mathbf{R}_0 - \overline{\mathbf{R}}_0)e_1 = \mathbf{0}. \tag{5.7}$$

Let $\mathbf{v}_-$ be the normalized eigenvector of $\mathbf{A}$ corresponding to λ_-. We will show that $\|\bar{\mathbf{r}}_k - \mathbf{r}_k\|_2$ is maximal when for all perturbations we have $\mathbf{f}_j = \varepsilon\|\mathbf{A}\|_2\mathbf{v}_-$ (or $\mathbf{F}_k = \varepsilon\|\mathbf{A}\|_2\mathbf{v}_-\vec{1}^*$). Subsequently, we will solve (5.7) for these perturbations from which our claim follows.

With (2.9) we rewrite (5.7) as

$$\mathbf{A}\mathbf{D}_k + \mathbf{F}_k\Gamma_k = \mathbf{D}_{k+1}\underline{T}_k,$$

with $\mathbf{d}_j \equiv (\mathbf{r}_j - \bar{\mathbf{r}}_j)\gamma_j$. Written as a three-term recurrence this reads

$$\mathbf{d}_j = 2\phi(\mathbf{A})\mathbf{d}_{j-1} - \mathbf{d}_{j-2} + 2\alpha\mathbf{f}_{j-1}\gamma_{j-1},$$

with $\mathbf{d}_0 = \mathbf{0}$, $\mathbf{d}_1 = \alpha\mathbf{f}_0$. This recurrence can be solved using standard techniques (e.g., [6, p.58],[9, Section 2]), which gives

$$\mathbf{d}_k = \alpha u_k(\phi(\mathbf{A}))\mathbf{f}_0\gamma_0 + \sum_{j=1}^{k-1} 2\alpha u_{k-j}(\phi(\mathbf{A}))\mathbf{f}_j\gamma_j,$$

where u_j is the so-called *Chebyshev polynomial of the second kind* (e.g., [6]), i.e., $u_{j+1}(t) = 2tu_j(t) - u_{j-1}(t)$, $u_0(t) = 0$ and $u_1(t) = 1$.

Realizing that $|u_j(t)| \leq j$ for $t \in [-1, 1]$, $u_j(-1) = (-1)^j j$ and $\text{sign}(\gamma_j) = (-1)^j$ it follows that

$$\|\mathbf{d}_k\|_2 \leq \left| \varepsilon\alpha\|\mathbf{A}\|_2 \left(u_k(\phi(\lambda_-))\gamma_0 + \sum_{j=1}^{k-1} 2u_{k-j}(\phi(\lambda_-))\gamma_j \right) \right|.$$

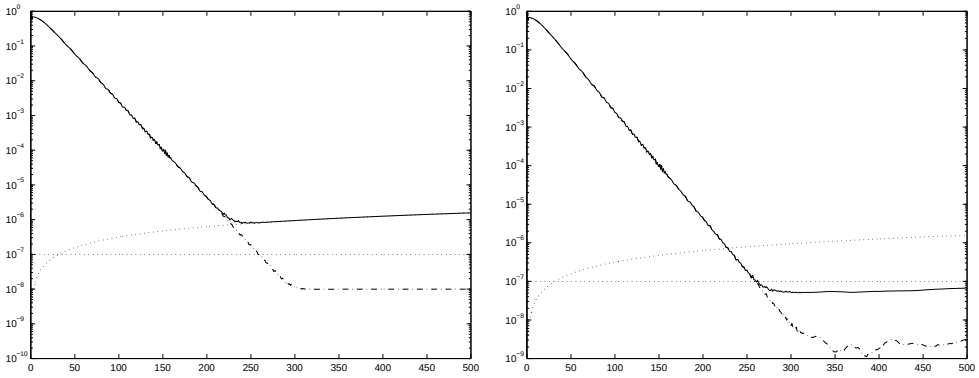

FIG. 5.1: Chebyshev iteration with $\eta_j = 10^{-10}/\|\mathbf{r}_j\|$, true residuals (—), norm computed residual (- ·) and the quantities $10^{-10}\mathcal{C}(\mathbf{A})$, $2k10^{-10}\sqrt{\mathcal{C}(\mathbf{A})}$ (both dotted). The matrix $\mathbf{A}$ has dimension 100 and $\mathcal{C}(\mathbf{A}) = 1000$. Left picture: errors have all components equal. Right picture: random errors.

This shows that the error is maximal if all perturbations are $\varepsilon\|\mathbf{A}\|_2\mathbf{v}_-$.

In order to solve (5.7) with $\mathbf{F}_k = \varepsilon\|\mathbf{A}\|_2\mathbf{v}_-\,\vec{1}^*$, we use a relation for the iterates which follows from substituting $\mathbf{R}_k = \mathbf{b}\vec{1}^* - \mathbf{A}\mathbf{X}_k$ in (2.6):

$$\mathbf{A}\mathbf{X}_k - \mathbf{b}\vec{1}^* = \mathbf{X}_{k+1}\underline{S}_k, \quad \mathbf{X}_0 e_1 = \mathbf{0}. \tag{5.8}$$

Comparing (5.8) with (5.7) shows that $\|\mathbf{r}_k - \overline{\mathbf{r}}_k\|_2$ is bounded by the norm of the $k+1$th approximate solution of Chebyshev iteration when the right hand side is $\varepsilon\|\mathbf{A}\|_2\mathbf{v}_-$, which is

$$\varepsilon\|\mathbf{A}\|_2\frac{1 - c_k(-1)/\gamma_k}{\lambda_-}\mathbf{v}_-.$$

By noting that $0 \le c_k(-1)/\gamma_k \le 1$ and $|c_k(-1)| = 1$ the proof can be concluded. □

In Figure 5.1 we give an illustration of our relaxation strategy for Chebyshev iteration as we did for Richardson iteration in Section 4.

5.1. Discussion. The effect of perturbations on the Chebyshev method has been investigated in literature. In [30], Woźniakowski analyzes the effect of finite precision arithmetic on the Chebyshev method. He describes a variant of the Chebyshev method where the residuals are computed directly and concludes that this method is stable. He, furthermore, points out this method is not well-behaved: the residuals for this method can stagnate at a level of $\mathcal{C}(\mathbf{A})\|\mathbf{A}\|_2\,\|\mathbf{A}^{-1}\mathbf{b}\|_2$ times the machine precision (it is interesting to note that a similar observation has been made for MINRES [25]). A method is *well-behaved* if the true residuals decrease below the level of $\|\mathbf{A}\|_2\,\|\mathbf{A}^{-1}\mathbf{b}\|_2$ times the machine precision.

Gutknecht et al. [19] analyze the residual gap for general Krylov subspace methods that use two three-term recurrences (one for the residuals and one for the approximate solutions). This analysis is applied in [18] in a qualitative discussion on the residual gap for the Chebyshev method. The approach from [18] differs essentially from ours in that we are using properties of the matrix $\underline{S}_k$ to bound the gap instead of a close inspection of the recursion as in [19]. The advantage is that it is easier to derive bounds in terms of global properties (as in Lemma 5.1) and our approach is not restricted to a certain type of recursion. Similar expressions as in [19] can be obtained from (3.3)

by writing out $e_j^* S_k^{-1} e_1$ using the LU-decomposition from Lemma 2.1. A difference is that, due to a different context, we do not consider perturbations on the recursion for the iterates but an analysis as in the previous sections can be easily extended to this case.

6. The Inexact Conjugate Gradient method. In this section we discuss relaxation strategies for the *Conjugate Gradient method* [20] and some of its variants although, strictly speaking, not all variants that we discuss use conjugate gradients. The Conjugate Gradient method (CG) is intimately connected with the Lanczos method, e.g., [10, Chapter 9]. This method exploits an efficient three-term recurrence for the construction of an orthogonal basis $\mathbf{v}_0, \ldots, \mathbf{v}_k$ for the Krylov subspace $\mathcal{K}_{k+1}$. The Lanczos method can be summarized as

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\underline{T}_k, \quad \mathbf{V}_k e_1 = \mathbf{b}, \tag{6.1}$$

where $\underline{T}_k$ is a $k+1$ by k tridiagonal Hermitian matrix and $\mathbf{V}_k$ is an orthonormal matrix whose columns span $\mathcal{K}_k$. The Conjugate Gradient method can be derived from the Lanczos method as is known from the work of Lanczos.

The most popular formulation of the CG method uses three coupled two-term recurrences, e.g., [20, Section 3]. For $j = 1, \ldots, k$ we have

$$\mathbf{c} = \mathbf{A}\mathbf{p}_{j-1} + \mathbf{g}_{j-1} \tag{6.2}$$

$$\mathbf{r}_j = \mathbf{r}_{j-1} - \alpha_{j-1}\mathbf{c} \tag{6.3}$$

$$\mathbf{x}_j = \mathbf{x}_{j-1} + \alpha_{j-1}\mathbf{p}_{j-1} \tag{6.4}$$

$$\mathbf{p}_j = \mathbf{r}_j + \beta_{j-1}\mathbf{p}_{j-1}, \tag{6.5}$$

with

$$\alpha_{j-1} \equiv \frac{\|\mathbf{r}_{j-1}\|_2^2}{\mathbf{p}_{j-1}^*\mathbf{c}} \quad \text{and} \quad \beta_{j-1} \equiv \frac{\|\mathbf{r}_j\|_2^2}{\|\mathbf{r}_{j-1}\|_2^2}, \tag{6.6}$$

and $\mathbf{p}_0 = \mathbf{r}_0 = \mathbf{b}$ and $\mathbf{x}_0 = \mathbf{0}$. We have added a perturbation, $\mathbf{g}_{j-1}$, to the matrix-vector product in (6.3) to obtain the inexact version with $\|\mathbf{g}_{j-1}\|_2 \leq \eta_{j-1}\|\mathbf{A}\|_2\,\|\mathbf{p}_{j-1}\|_2$.

The goal is, again, to obtain a final residual precision of ε. Therefore, we want to investigate the influence of the η_j on the residual gap and we make the assumption that the computed residuals become sufficiently small.

If we define $\Delta_k \equiv \mathrm{diag}(\alpha_0, \ldots, \alpha_{k-1})$ and

$$\widetilde{U}_k \equiv \begin{bmatrix} 1 & -\beta_0 & & & \\ & 1 & -\beta_1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & -\beta_{k-2} \\ & & & & 1 \end{bmatrix},$$

then we get the following equivalent matrix formulations of (6.3), (6.4), and (6.5),

$$\mathbf{A}\mathbf{P}_k + \mathbf{G}_k = \mathbf{R}_{k+1}\underline{J}_k\Delta_{k+1}^{-1}, \quad \mathbf{X}_{k+1}\underline{J}_k = -\mathbf{P}_k\Delta_k, \quad \mathbf{R}_k = \mathbf{P}_k\widetilde{U}_k. \tag{6.7}$$

Combining these relations gives

$$\mathbf{A}\mathbf{R}_k + (\mathbf{G}_k\widetilde{U}_k) = \mathbf{R}_{k+1}(\underline{J}_k\Delta_k^{-1}\widetilde{U}_k) \quad \text{and} \quad -\mathbf{R}_k = \mathbf{X}_{k+1}(\underline{J}_k\Delta_k^{-1}\widetilde{U}_k).$$

We see that (3.2) and (2.6) are satisfied for this method with $\underline{S}_k \equiv \underline{J}_k \Delta_k^{-1} \widetilde{U}_k$ and $\mathbf{F}_k \equiv \mathbf{G}_k \widetilde{U}_k$. The residuals of inexact CG method are now multiples (γ_j^{-1}) of the Lanczos vectors of an inexact Lanczos process given by

$$\mathbf{A}\mathbf{V}_k + \widetilde{\mathbf{F}}_k = \mathbf{V}_{k+1}\underline{T}_k, \tag{6.8}$$

where $\underline{T}_k = \Gamma_{k+1}^{-1}\underline{S}_k\Gamma_k$, $\Gamma_k = \mathrm{diag}(\vec{\gamma}_k)$, $\gamma_j = (-1)^j \|\mathbf{r}_j\|_2^{-1}$, $\mathbf{V}_k = \mathbf{R}_k\Gamma_k$ and $\widetilde{\mathbf{F}}_k = \mathbf{F}_k\Gamma_k$.

We can use (3.3) to get an expression for the residual gap

$$\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k) = -\mathbf{F}_k S_k^{-1} e_1 = -\mathbf{G}_k \widetilde{U}_k S_k^{-1} e_1 = -\mathbf{G}_k \Delta_k J_k^{-1} e_1 = -\sum_{j=0}^{k-1} \alpha_j \mathbf{g}_j.$$

From $\|\mathbf{g}_j\|_2 \leq \eta_j \|\mathbf{A}\|_2 \, \|\mathbf{p}_j\|_2$, we get

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 \leq \sum_{j=0}^{k-1} \eta_j |\alpha_j| \|\mathbf{A}\|_2 \, \|\mathbf{p}_j\|_2 = \sum_{j=0}^{k-1} \eta_j \|\mathbf{A}\|_2 \, \|\mathbf{x}_{j+1} - \mathbf{x}_j\|_2. \tag{6.9}$$

This type of expression is well-known from the work in [24, 14]. Since we have from (6.6) that $\sqrt{\beta_j}\|\mathbf{r}_j\|_2 = \|\mathbf{r}_{j+1}\|_2$, we can bound this as

$$\begin{aligned} \|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 &\leq \mathcal{C}(\mathbf{A}) \sum_{j=0}^{k-1} \eta_j (\|\mathbf{r}_{j+1}\|_2 + \|\mathbf{r}_j\|_2) \\ &= \mathcal{C}(\mathbf{A}) \sum_{j=0}^{k-1} \eta_j \|\mathbf{r}_j\|_2 (1 + \sqrt{\beta_j}). \end{aligned} \tag{6.10}$$

Following their work for inexact GMRES and (3.6), Bouras, Frayssé, and Giraud proposed in [3] for CG a relaxation strategy where

$$\eta_j = \max\left\{ \frac{\varepsilon}{\|\mathbf{r}_j\|_2}, \varepsilon \right\}. \tag{6.11}$$

If we take the slightly weaker $\eta_j = \frac{\varepsilon}{\|\mathbf{r}_j\|_2}$, then we get for this strategy using (6.10)

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 \leq \varepsilon k \mathcal{C}(\mathbf{A}) \max_{0 \leq j < k} (1 + \sqrt{\beta_j}). \tag{6.12}$$

As long as the β_j are not too large, this strategy can work very well. Indeed, numerical experiments with symmetric positive definite matrices $\mathbf{A}$ confirm this.

Practical problems often lead to a matrix $\mathbf{A}$ that is indefinite, for instance the QCD example mentioned in the introduction. In this case there can be very large intermediate residuals caused by an eigenvalue of T_k being "accidentally" close to zero. The CG method is still used in practice for solving Hermitian indefinite systems, despite its lack of robustness. One reason is that, although the tridiagonal matrix can be ill-conditioned in one iteration, this can never happen for two consecutive iterations, e.g., [1, 16]. The situation of an eigenvalue of T_k close to zero is in literature often referred to as a *near breakdown*. It results in a value of β_j that is very large and it follows from (6.12) that the proposed strategy in (6.11) may fail in achieving the required residual precision.

From (6.10) it follows that picking $\eta_j = \varepsilon/(\|\mathbf{r}_{j+1}\|_2 + \|\mathbf{r}_j\|_2)$ is a better strategy in this case. However, this is not practical since $\mathbf{r}_{j+1}$ is not known yet. An alternative is to consider the first bound in Equation (6.9) and pick

$$\eta_j = \frac{\varepsilon}{\alpha_j \|\mathbf{p}_j\|_2}.$$

If the approximation of the matrix-vector product is computed with an incremental method the inner-product of $\mathbf{p}_j$ and the "current" approximation for the product can be monitored at the cost of an additional inner-product and from this α_j can be estimated. Nevertheless, in case of a near breakdown a very accurate matrix-vector product is still necessary. We will therefore consider variants of the Conjugate Gradient method in Section 6.1.

Studying the convergence of the computed residuals is a much more difficult topic. Greenbaum [13] showed that the convergence of a slightly perturbed CG process can be related to that of an enlarged matrix with eigenvalues in small clusters around the eigenvalues of the original matrix. The width of these clusters is determined by the size of the perturbation of the Lanczos process. Unfortunately, this analysis does not apply in our situation since it does not explain why the accuracy of the matrix-vector product can be relaxed when the CG method converges as was the case for Richardson iteration and Chebyshev iteration in the previous sections. Still, when $\widetilde{\mathbf{F}}_k$ in (6.8) has some very large columns, we can expect that the convergence behavior is severely altered if the method is not converged. We know that for the jth column of $\widetilde{\mathbf{F}}_k$ we have $\|\widetilde{\mathbf{f}}_{j-1}\| = \|\mathbf{g}_{j-1} - \beta_{j-2}\mathbf{g}_{j-2}\|_2/\|\mathbf{r}_{j-1}\|_2$. We will bound the length of this vector. A simple analysis shows that

$$\|\mathbf{p}_{j-1}\|_2 = \|\mathbf{R}_k \widetilde{U}_k^{-1} e_j\|_2 \leq \|\mathbf{R}_k \Gamma_k\|_2 \, \|\Gamma_k^{-1} \widetilde{U}_k^{-1} e_j\|_2 = \|\mathbf{V}_k\|_2 \frac{\|\mathbf{r}_{j-1}\|_2^2}{\rho_{j-1}},$$

where

$$\rho_j \equiv \left(\sum_{i=0}^{j} \|\mathbf{r}_i\|^{-2} \right)^{-1/2}. \tag{6.13}$$

Note that ρ_j can be interpreted as the norm of a smoothed residual, see e.g., [20, Section 7]. We have the following upper bound for the norm of the jth column of $\widetilde{\mathbf{F}}_k$

$$\|\widetilde{\mathbf{f}}_{j-1}\| = \|\mathbf{g}_{j-1} - \beta_{j-2}\mathbf{g}_{j-2}\|_2/\|\mathbf{r}_{j-1}\|_2 \leq \|\mathbf{A}\|_2 \, \|\mathbf{V}_k\|_2 \, \|\mathbf{r}_{j-1}\|_2 \left(\frac{\eta_{j-1}}{\rho_{j-1}} + \frac{\eta_{j-2}}{\rho_{j-2}} \right).$$

The ratio $\|\mathbf{r}_{j-1}\|_2/\rho_{j-1}$ is large in case of a near breakdown, since we then have that $\rho_{j-1} \ll \|\mathbf{r}_{j-1}\|_2$. We find that when there is a near breakdown there can be a very large perturbation of the Lanczos relation. One consequence is a large residual gap (as discussed). Another effect is a potential delay in convergence (or even worse). A simple numerical example of this is given in the next section.

6.1. Variants of the Conjugate Gradient method. Mathematically equivalent variants of the CG method can be derived from the Lanczos method in (6.1). In this section we will consider two alternatives for the CG method in the previous section. These methods are based on a three-term recurrence instead of a coupled two-term recurrence. We start with a short derivation.

Since the CG residuals are multiples of the Lanczos vectors we can derive the coefficients for the recursion (2.3) from (6.1) by virtue of (2.9). Therefore, we write

$$\underline{T}_k \equiv \begin{bmatrix} \alpha_0 & \beta_0 & & & \\ \beta_0 & \alpha_1 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{k-2} \\ & & & \beta_{k-2} & \alpha_{k-1} \\ & & & & \beta_{k-1} \end{bmatrix}, \quad \underline{S}_k \equiv \begin{bmatrix} \eta_0 & \delta_0 & & & \\ \tau_0 & \eta_1 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \delta_{k-2} \\ & & & \tau_{k-2} & \eta_{k-1} \\ & & & & \tau_{k-1} \end{bmatrix}.$$

From the property that $\vec{1}^* \underline{S}_k = \vec{0}$, it immediately follows that $\tau_j = -(\eta_j + \delta_{j-1})$ (with $\delta_{-1} = 0$). Using (2.9) we see that $\eta_j = \beta_j$, $\delta_j = \beta_j(\gamma_j/\gamma_{j+1})$ and $\tau_j = \beta_j(\gamma_{j+1}/\gamma_j)$. Eliminating β_j gives that $\delta_j = \tau_j(\gamma_j/\gamma_{j+1})^2$. With $\delta_{-1} = 0$ we get, using (2.12),

$$\delta_j = \tau_j \frac{\|\mathbf{r}_{j+1}\|_2^2}{\|\mathbf{r}_j\|_2^2}, \quad \eta_j = \frac{\mathbf{r}_j^* \mathbf{A} \mathbf{r}_j}{\|\mathbf{r}_j\|_2^2}, \quad \tau_j = -(\eta_j + \delta_{j-1}).$$

Computing the residuals and iterates with these coefficients using the recursions described by (2.3) and (2.6) gives a variant of CG known as *Orthores* (we use the nomenclature from [19]).

Rutishauser's variant of this method is obtained by introducing auxiliary variables $\Delta\mathbf{x}_j$ and $\Delta\mathbf{r}_j$ using the LU-decomposition, $\underline{S}_k = \underline{J}_k U_k$, from Lemma 2.1 where $(U_k)_{j,j} = -\tau_{j-1}$ and $(U_k)_{j+1,j} = \delta_{j-1}$. This gives

$$\begin{aligned} \mathbf{R}_{k+1}\underline{J}_k &= \Delta\mathbf{R}_k, \quad \Delta\mathbf{R}_k U_k = \mathbf{A}\mathbf{R}_k \quad \text{and} \\ \mathbf{X}_{k+1}\underline{J}_k &= \Delta\mathbf{X}_k, \quad \Delta\mathbf{X}_k U_k = -\mathbf{R}_k. \end{aligned} \tag{6.14}$$

Now that we have defined the two methods, we shift our attention to the inexact case. In *inexact Orthores* the matrix-vector product is perturbed in step j with a term $\mathbf{g}_{j-1}$. This gives the (familiar) perturbed residual relation

$$\mathbf{A}\mathbf{R}_k + \mathbf{F}_k = \mathbf{R}_{k+1}\underline{S}_k, \quad \mathbf{R}_k e_1 = \mathbf{b},$$

where $\mathbf{F}_k = \mathbf{G}_k$ and, therefore, $\|\mathbf{f}_j\|_2 \leq \eta_j \|\mathbf{A}\|_2 \|\mathbf{r}_j\|_2$. For the inexact version of Rutishauser's method we have $\Delta\mathbf{R}_k U_k = \mathbf{A}\mathbf{R}_k + \mathbf{G}_k$, and it follows that for the same perturbations, the inexact version of Orthores and Rutishauser's variant are equivalent under the assumption of exact arithmetic and, hence, the same upper bounds apply.

We want to bound the gap for the discussed methods and derive a suitable relaxation strategy. Therefore, we note that the residuals of inexact Orthores are now multiples (γ_j^{-1}) of the Lanczos vectors of an inexact Lanczos process given by (6.8) with $\underline{T}_k = \Gamma_{k+1}^{-1}\underline{S}_k\Gamma_k$, $\Gamma_k = \mathrm{diag}(\vec{\gamma}_k)$ and $\gamma_j = (-1)^j\|\mathbf{r}_j\|_2^{-1}$. Combining this with Lemma 3.1 shows that

$$|e_j^* S_k^{-1} e_1| \leq \sigma_{\min}(\underline{T}_k)^{-1} \frac{1}{\|\mathbf{r}_{j-1}\|_2} \left(\rho_{j-1} + \|\mathbf{r}_k\|_2\right), \tag{6.15}$$

with ρ_{j-1} as defined in (6.13).

From (3.3) we get the following bound on the residual gap

$$\begin{aligned}\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 &\leq \sigma_{\min}(\underline{T}_k)^{-1} \sum_{j=0}^{k-1} \|\mathbf{r}_j\|_2^{-1} \left(\rho_j + \|\mathbf{r}_k\|_2\right) \|\mathbf{f}_j\|_2 \\ &\leq \|\mathbf{A}\|_2 \, \sigma_{\min}(\underline{T}_k)^{-1} \sum_{j=0}^{k-1} \eta_j \left(\rho_j + \|\mathbf{r}_k\|_2\right).\end{aligned}$$

Recall that we assume that the computed residuals ultimately become small enough. Now, assume that we terminate the iterative process for $\|\mathbf{r}_k\|_2 = \mathcal{O}(\varepsilon)$. In this case we see that the size of the gap is essentially determined by the ρ_j, the η_j, and $\sigma_{\min}(\underline{T}_k)$. Unfortunately, we have no a priori knowledge about the size of $\sigma_{\min}(\underline{T}_k)$. We hope that this quantity is in the order of $\sigma_{\min}(\mathbf{A})$. For inexact Orthores (and Rutishauser's variant) we propose the following relaxation strategy

$$\eta_j = \frac{\varepsilon}{\rho_j}, \tag{6.16}$$

with ρ_j as in (6.15). Note that ρ_j can be computed at little additional cost. With this relaxation strategy we get for the residual gap

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 \leq \varepsilon k \|\mathbf{A}\|_2 \, \sigma_{\min}(\underline{T}_k)^{-1} \left(1 + \frac{\|\mathbf{r}_k\|_2}{\rho_k}\right).$$

This shows that the distance between the computed and true residual can be large when there is a near breakdown but when the process is terminated, if $\|\mathbf{r}_k\|_2 = \mathcal{O}(\varepsilon)$, the gap is hopefully $\mathcal{O}(\varepsilon)$.

To summarize: if we consider the upper bounds on the residual gap, we see that for the two discussed variants based on a three-term recurrence there is no need in computing the matrix-vector product more accurately in case of a near breakdown in contrast to CG. As seen, we can exploit this in our relaxation strategy. For indefinite matrices $\mathbf{A}$, where the convergence behavior of the residuals is highly irregular, the alternative CG methods and relaxation strategy in this section can offer advantages over CG and the relaxation strategy by Bouras et al. in (6.11) .

Furthermore, for the three-term recurrences a near breakdown does not lead to a large perturbation of the (implicit) Lanczos relation. Hence, we expect the effect of loss of convergence speed caused by near breakdowns less dramatic than for CG.

In Figure 6.1 we give a simple illustration. The right-hand side has all components equal and the matrix is $\mathbf{A} = \text{diag}(1:100) - 5.2025\,\mathbf{I}$. The shift causes a large intermediate residual in the fifth step. The figure illustrates that Orthores and Rutishauser's variant perform equal and better than the CG method with respect to accuracy and convergence speed. Here, we prefer to use the three-term recurrence variants over the coupled two-term recurrences.

6.2. Discussion. For positive definite systems, the standard CG method seems the most appropriate in the inexact setting. Although this is not apparent from our analysis since the residuals are not monotonically decreasing. In [14] the stability of CG is explained by using the fact that the errors in CG ($\|\mathbf{x}_k - \mathbf{x}\|_2$) are monotonically decreasing in combination with an expression as in (6.9). We refer to [14, 19] for more information and a comparison of the discussed variants in the context of rounding errors and positive definite $\mathbf{A}$.

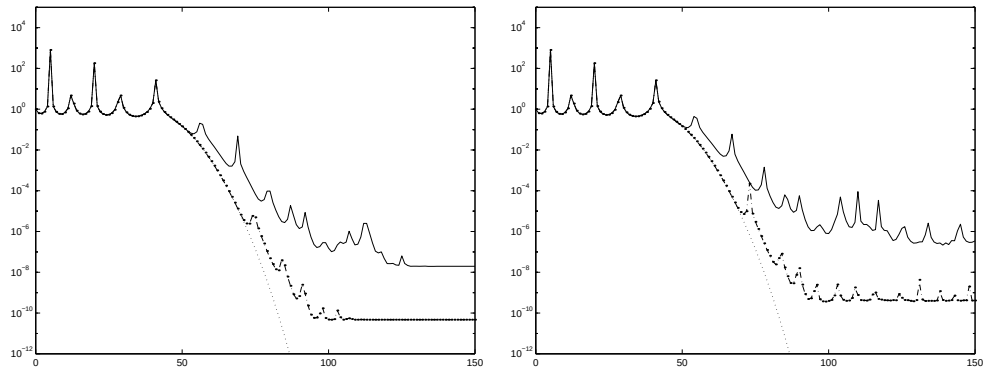

FIG. 6.1: exact FOM (dotted), CG (—), Orthores (- ·), Rutishauser's variant (dots). In both pictures $\varepsilon = 10^{-10}$. Left picture: $\eta_j = \varepsilon$. Right picture: $\eta_j = \varepsilon/\rho_j$.

The observations in the previous section show that (in the inexact setting) the use of a three-term recurrence for solving Hermitian indefinite systems can offer advantages over the standard CG implementation. Specially, when the matrix $\mathbf{A}$ is not too ill-conditioned and convergence is irregular. For numerical experiments we refer the reader to Section 8.

Numerical experiments (not reported here) suggest that this is not necessarily the case when floating point errors are the only source of errors. For example, near-breakdowns also influence the attainable precision of Rutishauser's variant of the CG method, just as for standard CG. Orthores, on the other hand seems not sensitive to peaks but appears to be, like Chebyshev iteration and MINRES, not well-behaved (cf. Section 5.1). Our analysis can be extended for making a rounding error analysis of several variants of the CG method for indefinite systems. This can help identifying the different aspects of the CG method that influence the accuracy.

Studying the behavior of the computed residuals is a much more difficult subject. In general we observe in numerical experiments that the computed residuals become small enough for the residual gap to be a meaningful indicator for the attainable residual precision. Nevertheless, small perturbations of the matrix-vector product can seriously delay convergence for the CG method and its variants. This also is the case for inexact GMRES that we discuss in the next section and we refer to this section for a numerical example.

As a final remark we note that we could have proposed inexact MINRES as the alternative for indefinite systems. We have not done this here for two reasons. A simple analysis of inexact MINRES shows that essentially the same bound applies as for inexact Orthores and therefore the same relaxation strategy is appropriate. Secondly, we want to illustrate that the underlying mechanism for constructing the Krylov subspace is important and *not* the fact if the residuals are smoothed. This is also illustrated in the next section in our discussion about inexact FOM and GMRES.

7. Inexact FOM and GMRES. The Lanczos method is a starting point for the derivation of a large class iterative methods for Hermitian matrices $\mathbf{A}$. For non-Hermitian systems, the *Arnoldi method* [10, Section 9.4] can be used for constructing an orthonormal basis $\mathbf{v}_0, \ldots, \mathbf{v}_k$ for $\mathcal{K}_{k+1}$ and can therefore serve as a starting point. The Arnoldi method can be summarized by the following relation

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\underline{T}_k, \quad \mathbf{V}_k e_1 = \mathbf{b}, \tag{7.1}$$

where $\underline{T}_k$ is $k+1$ by k upper Hessenberg and $\mathbf{V}_k$ is n by k and orthogonal.

If in step j of the Arnoldi method the matrix-vector product is done approximately, i.e., a perturbation $\mathbf{g}_{j-1}$ is added to the matrix-vector product $\mathbf{A}\mathbf{v}_{j-1}$, then we obtain the *inexact Arnoldi method*. This latter method satisfies the following perturbed Arnoldi relation

$$\mathbf{A}\mathbf{V}_k + \widetilde{\mathbf{F}}_k = \mathbf{V}_{k+1}\underline{T}_k, \quad \mathbf{V}_k e_1 = \mathbf{b}/\|\mathbf{b}\|_2 = \mathbf{b}, \tag{7.2}$$

where $\widetilde{\mathbf{F}}_k = \mathbf{G}_k$ and, therefore, $\|\widetilde{\mathbf{f}}_j\| \leq \eta_j \|\mathbf{A}\|_2 \|\mathbf{v}_j\|_2 = \eta_j \|\mathbf{A}\|_2$. An interesting observation is that $\mathbf{V}_k$ is still an orthogonal matrix, but now the columns span the Krylov subspace $\mathcal{K}_k(\widehat{\mathbf{A}}_k, \mathbf{b})$, with $\widehat{\mathbf{A}}_k \equiv \mathbf{A} + \widetilde{\mathbf{F}}_k \mathbf{V}_k^*$. We will assume in this section that T_j is invertible and $\underline{T}_j$ has full rank for $j \leq k$.

The *inexact FOM* and *inexact GMRES method* [2] use the Arnoldi relation explicitly for constructing iterates of the form

$$y_j^{\mathrm{F}} = T_j^{-1} e_1, \quad \mathbf{x}_j^{\mathrm{F}} = \mathbf{V}_j y_j^{\mathrm{F}} \quad \text{and} \quad y_j^{\mathrm{G}} = \underline{T}_j^{\dagger} e_1, \quad \mathbf{x}_j^{\mathrm{G}} = \mathbf{V}_j y_j^{\mathrm{G}}.$$

The corresponding computed residuals are given by

$$\mathbf{r}_j^{\mathrm{F}} = \mathbf{V}_{j+1}(I - \underline{T}_j T_j^{-1}) e_1 \quad \text{and} \quad \mathbf{r}_j^{\mathrm{G}} = \mathbf{V}_{j+1}(I - \underline{T}_j \underline{T}_j^{\dagger}) e_1.$$

These expressions are a special case of Equations (2.10) and (2.12) and, therefore, we get from Lemma 2.2 that $\mathbf{r}_j^{\mathrm{F}} = \mathbf{v}_j/\gamma_j$ and $\mathbf{r}_j^{\mathrm{G}} = \|\vec{\gamma}_j\|_2^{-2} \mathbf{V}_j \vec{\gamma}_j$, where $\vec{\gamma}_k$ is as defined in Section 2, i.e., $\gamma_k^* \underline{T}_k = \vec{0}^*$ and $\vec{\gamma}_k^* e_1 = 1$. This gives the following relation between the norms of the computed residuals of inexact FOM and GMRES

$$\rho_j \equiv \|\mathbf{r}_j^{\mathrm{G}}\|_2 = \left(\sum_{i=0}^{j} \|\mathbf{r}_i^{\mathrm{F}}\|^{-2} \right)^{-1/2}. \tag{7.3}$$

The same result is well-known for exact FOM and GMRES from the work of Brown [4].

Note that an alternative expression for the residuals is given by $\mathbf{r}_j^{\mathrm{F}} = \mathbf{b} - \widehat{\mathbf{A}}_j \mathbf{x}_j^{\mathrm{F}}$ and similarly for inexact GMRES. Hence, inexact FOM/GMRES is equivalent to exact (or ideal) FOM/GMRES applied to the linear system $\widehat{\mathbf{A}}_n \mathbf{x} = \mathbf{b}$. Hence, the computed residuals are monotonically decreasing and after at most n steps the method terminates with $\mathbf{x}_n^{\mathrm{F}} = \mathbf{x}_n^{\mathrm{G}} = (\mathbf{A} + \widetilde{\mathbf{F}}_n \mathbf{V}_n^*)^{-1} \mathbf{b}$. In the remainder of this section we will drop the superscripts F or G in expressions that are valid for both methods.

To bound the residual gap in step k we use an expression for the gap that is equivalent to (3.3) but is expressed in terms of the matrix $\widetilde{\mathbf{F}}_k$ (this simplifies the analysis in this section somewhat). We have

$$\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k) = \mathbf{r}_k - (\mathbf{b} - (\widehat{\mathbf{A}}_k - \widetilde{\mathbf{F}}_k \mathbf{V}_k^*)\mathbf{x}_k) = -\widetilde{\mathbf{F}}_k y_k. \tag{7.4}$$

Hence

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 = \|\widetilde{\mathbf{F}}_k y_k\|_2 \leq \|\mathbf{A}\|_2 \sum_{j=0}^{k-1} \eta_j |e_{j+1}^* y_k|. \tag{7.5}$$

Since, the iterates of inexact FOM and GMRES ultimately will approach the same vector $\widehat{\mathbf{A}}_n^{-1}\mathbf{b}$ it is evident from (7.4) that an appropriate relaxation strategy for inexact GMRES is also suitable for inexact FOM, and vice versa.

If we plug (3.4) into (7.5), then we get the following bound for the residual gap of inexact FOM,

$$\|\mathbf{r}_k^{\mathrm{F}} - (\mathbf{b} - \mathbf{A}\mathbf{x}_k^{\mathrm{F}})\|_2 \leq \|\mathbf{A}\|_2\, \sigma_{\min}(\underline{T}_k)^{-1} \sum_{j=0}^{k-1} \eta_j (\|\mathbf{r}_j^{\mathrm{G}}\|_2 + \|\mathbf{r}_k^{\mathrm{F}}\|_2), \tag{7.6}$$

and for inexact GMRES we get

$$\|\mathbf{r}_k^{\mathrm{G}} - (\mathbf{b} - \mathbf{A}\mathbf{x}_k^{\mathrm{G}})\|_2 \leq \|\mathbf{A}\|_2\, \sigma_{\min}(\underline{T}_k)^{-1} \sum_{j=0}^{k-1} \eta_j \|\mathbf{r}_j^{\mathrm{G}}\|_2. \tag{7.7}$$

We follow the same approach as for Orthores in Section 6.1 and assume that we terminate inexact FOM/GMRES when $\|\mathbf{r}_k\|_2 = \mathcal{O}(\varepsilon)$, where ε is again the required residual precision. We see that in step k the residual gap is essentially determined by the η_j, the $\|\mathbf{r}_j^{\mathrm{G}}\|_2$ (or ρ_j) and the smallest singular value of $\underline{T}_k$. Again, the size of the smallest singular value of the Hessenberg matrix is difficult to estimate a priori (we can however monitor it during the iterations). Assume that this singular value is not getting too small, this suggest again that relaxation is possible with $\eta_j = \varepsilon/\rho_j$. This results for inexact FOM in the bound

$$\|\mathbf{r}_k^{\mathrm{F}} - (\mathbf{b} - \mathbf{A}\mathbf{x}_k^{\mathrm{F}})\|_2 \leq \varepsilon k \sigma_{\min}(\underline{T}_k)^{-1} \|\mathbf{A}\|_2 \left(1 + \frac{\|\mathbf{r}_k^{\mathrm{F}}\|_2}{\rho_k}\right), \tag{7.8}$$

and for inexact GMRES we get

$$\|\mathbf{r}_k^{\mathrm{G}} - (\mathbf{b} - \mathbf{A}\mathbf{x}_k^{\mathrm{G}})\|_2 \leq \varepsilon k \sigma_{\min}(\underline{T}_k)^{-1} \|\mathbf{A}\|_2. \tag{7.9}$$

We see that the relaxation strategy derived from the bounds on the residual gap confirms the empirical choice of Bouras et al. in (3.6) for GMRES and can explain the success of this approach. See also the numerical experiments in [2].

For inexact GMRES we know that the size of the computed residuals monotonically decrease and $\mathbf{r}_n = \mathbf{0}$ and therefore the gap provides in the end useful information about the attainable accuracy. However, this does not say anything about the speed of convergence of the perturbed process. In some cases it can be shown that convergence of the relaxed process is approximately as fast as for the unperturbed process (similar to what we have seen for Chebyshev iteration). This is for example the case when the perturbation $\widetilde{\mathbf{f}}_j \in \mathcal{K}_{j+2}$, which means that only the Hessenberg matrix differs from the Hessenberg matrix of the unperturbed process and $\mathbf{V}_k$ remains the same. (For notational convenience we consider a perturbation in just one step.)

To show this we consider two inexact Arnoldi processes, both with $\eta_i = 0$ for $i < k$, except for the second process we take $\eta_j = \eta$ for some $j < k-1$. The corresponding quantities for the first process (the unperturbed process) are denoted with lines on top of them. Suppose we have $\|\overline{\mathbf{r}_k}\|_2 = \mathcal{O}(\varepsilon)$. The goal is now to show that $\|\mathbf{r}_k\|_2 = \mathcal{O}(\varepsilon)$ if the proposed relaxation strategy is applied.

Let $g \equiv \mathbf{V}_k^* \widetilde{\mathbf{f}}_j$. Then we have that $\underline{T}_k = \overline{\underline{T}}_k + g\, e_{j+1}^*$ and for inexact GMRES we can apply perturbation theory for the least squares problem. For example with Theorem 19.1 in [21], we get

$$\left| \|\overline{\mathbf{r}_k^{\mathrm{G}}}\|_2 - \|\mathbf{r}_k^{\mathrm{G}}\|_2 \right| \leq \|\overline{\mathbf{r}_k^{\mathrm{G}}} - \mathbf{r}_k^{\mathrm{G}}\| \leq (1 + 2\mathcal{C}(\mathbf{A}))\|\widetilde{\mathbf{f}}_j\|_2.$$

We see that this is not sufficient for explaining the fast convergence of the perturbed process. By generalizing Theorem 19.1 in [21], we get the following lemma.

LEMMA 7.1. *Let $j < k-1$. Then*

$$\|\overline{\mathbf{r}_k^{\mathrm{F}}} - \mathbf{r}_k^{\mathrm{F}}\|_2 \leq \|\mathbf{A}^{-1}\|_2 \, \| \left(\mathbf{I} - \widehat{\mathbf{A}}_k \mathbf{V}_k (\mathbf{V}_k^* \widehat{\mathbf{A}}_k \mathbf{V}_k)^{-1} \mathbf{V}_k^*\right) \widetilde{\mathbf{f}}_j \|_2 \, (\|\overline{\mathbf{r}_j^{\mathrm{G}}}\|_2 + \|\overline{\mathbf{r}_k^{\mathrm{F}}}\|_2),$$
$$\|\overline{\mathbf{r}_k^{\mathrm{G}}} - \mathbf{r}_k^{\mathrm{G}}\|_2 \leq \|\mathbf{A}^{-1}\|_2 \, \| \left(\mathbf{I} - \widehat{\mathbf{A}}_k \mathbf{V}_k (\mathbf{V}_k^* \widehat{\mathbf{A}}_k^* \widehat{\mathbf{A}}_k \mathbf{V}_k)^{-1} (\widehat{\mathbf{A}}_k \mathbf{V}_k)^*\right) \widetilde{\mathbf{f}}_j \|_2 \, (\|\overline{\mathbf{r}_j^{\mathrm{G}}}\|_2 + \|\overline{\mathbf{r}_k^{\mathrm{G}}}\|_2).$$

Proof. For inexact FOM we have

$$\|\overline{\mathbf{r}_k^{\mathrm{F}}} - \mathbf{r}_k^{\mathrm{F}}\|_2 = |e_{k+1}^* \underline{T}_k e_k| \, |e_k^* (\overline{T}_k^{-1} - T_k^{-1}) e_1| = |e_{k+1}^* \underline{T}_k e_k| \, |e_k^* T_k^{-1} g| \, |e_{j+1}^* \overline{y_k^{\mathrm{F}}}|.$$

The first two terms can be rewritten with

$$|e_{k+1}^* \underline{T}_k e_k| \, |e_k^* T_k^{-1} g| = \| \left(\mathbf{I} - \widehat{\mathbf{A}}_k \mathbf{V}_k (\mathbf{V}_k^* \widehat{\mathbf{A}}_k \mathbf{V}_k)^{-1} \mathbf{V}_k^*\right) \widetilde{\mathbf{f}}_j \|_2,$$

and the last term can be bounded using Lemma 3.1. This proves the first statement.

Now, we turn to the proof of the second statement. Define $z_k \equiv (I - \underline{T}_k \underline{T}_k^\dagger) e_1$ and $\overline{z}_k \equiv (I - \overline{\underline{T}}_k \overline{\underline{T}}_k^\dagger) e_1$. We have to bound $\|\overline{\mathbf{r}_k^{\mathrm{G}}} - \mathbf{r}_k^{\mathrm{G}}\|_2 = \|\overline{z}_k - z_k\|_2$.

$$\begin{aligned} \overline{z}_k - z_k &= (I - \underline{T}_k \underline{T}_k^\dagger)(\overline{z}_k - z_k) + \underline{T}_k \underline{T}_k^\dagger (\overline{z}_k - z_k) \\ &= (I - \underline{T}_k \underline{T}_k^\dagger) g \; e_{j+1}^* \overline{y_k^{\mathrm{G}}} + \underline{T}_k \underline{T}_k^\dagger (I - \overline{\underline{T}}_k \overline{\underline{T}}_k^\dagger) \overline{z}_k. \end{aligned}$$

For the norm of the first term we have

$$\|(I - \underline{T}_k \underline{T}_k^\dagger) g \, e_{j+1}^* \overline{y_k^{\mathrm{G}}}\|_2 = \|(I - \underline{T}_k \underline{T}_k^\dagger) g\|_2 \, |e_{j+1}^* \overline{y_k^{\mathrm{G}}}|.$$

This expression is bounded using Lemma 3.1. For the second term we have

$$\|\underline{T}_k \underline{T}_k^\dagger (I - \overline{\underline{T}}_k \overline{\underline{T}}_k^\dagger) \overline{z}_k\|_2 \leq \|\underline{T}_k \underline{T}_k^\dagger (I - \overline{\underline{T}}_k \overline{\underline{T}}_k^\dagger)\|_2 \, \|\overline{z}_k\|_2 = \|\underline{T}_k \underline{T}_k^\dagger (I - \overline{\underline{T}}_k \overline{\underline{T}}_k^\dagger)\|_2 \, \|\overline{\mathbf{r}_k^{\mathrm{G}}}\|_2.$$

We know from [26] that $\|\underline{T}_k \underline{T}_k^\dagger (I - \overline{\underline{T}}_k \overline{\underline{T}}_k^\dagger)\|_2 = \|(I - \underline{T}_k \underline{T}_k^\dagger) \overline{\underline{T}}_k \overline{\underline{T}}_k^\dagger\|_2$ and

$$\|(I - \underline{T}_k \underline{T}_k^\dagger) \overline{\underline{T}}_k \overline{\underline{T}}_k^\dagger\|_2 = \|(I - \underline{T}_k \underline{T}_k^\dagger)(\underline{T}_k - g \, e_{j+1}^*) \overline{\underline{T}}_k^\dagger\|_2 \leq \|(I - \underline{T}_k \underline{T}_k^\dagger) g\|_2 \, \|\overline{\underline{T}}_k^\dagger\|_2.$$

The proof is concluded by noting that

$$\|(I - \underline{T}_k \underline{T}_k^\dagger) g\|_2 = \| \left(\mathbf{I} - \widehat{\mathbf{A}}_k \mathbf{V}_k (\mathbf{V}_k^* \widehat{\mathbf{A}}_k^* \widehat{\mathbf{A}}_k \mathbf{V}_k)^{-1} (\widehat{\mathbf{A}}_k \mathbf{V}_k)^*\right) \widetilde{\mathbf{f}}_j \|_2. \quad \square$$

We conclude from this lemma that, if $\widetilde{\mathbf{f}}_j \in \mathcal{K}_{j+2}$, then the convergence of the relaxed method is as fast as for the unperturbed method. The proof in Lemma 7.1 essentially used that the change in the Hessenberg matrix is in the order of the size of the perturbation $\widetilde{\mathbf{f}}_j$. Therefore, this theorem is difficult to extend to more general perturbations, since the Hessenberg reduction is not forward stable, see [29]. In fact, small perturbations of the matrix-vector product can indeed severely delay convergence. This is illustrated by the following experiment with inexact FOM. (Note that the convergence of the computed residuals of inexact FOM and GMRES are related through the vector $\vec{\gamma}_k$.)

The matrix $\mathbf{A} \in \mathbb{R}^{100 \times 100}$ is lower bidiagonal with diagonal elements $(\mathbf{A})_{j,j} = j$ and and has ones on its lower bidiagonal. For the the right-hand side we have taken

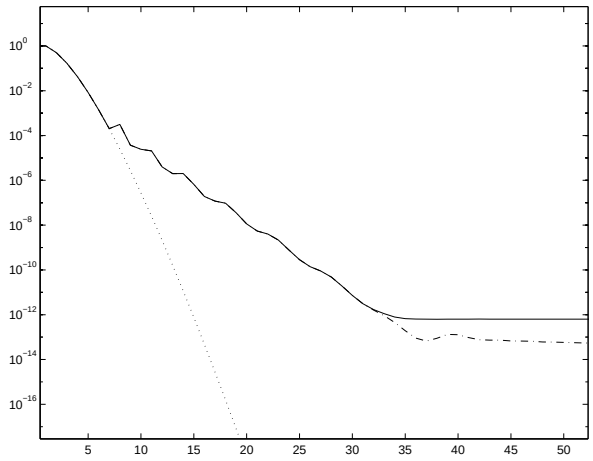

FIG. 7.1: Convergence inexact FOM with $\eta_j = \varepsilon = 10^{-12}$: true residual (—), computed residual (- ·) and $1/j!$ (dotted).

$\mathbf{b} = e_1$. It easily follows for this example that $\overline{T}_n = \mathbf{A}$ and the corresponding vector $\vec{\gamma}_j$ with $\vec{\gamma}_j^* \underline{T}_j = \vec{0}^*$ and $\vec{\gamma}_j^* e_1 = 1$ is given by $\gamma_j = (-1)^j j!$. Therefore we have that $\|\overline{\mathbf{r}^{\mathrm{F}}}_j\|_2 = 1/j!$. Figure 7.1 shows the convergence history of inexact FOM with $\eta_j = \varepsilon = 10^{-12}$. Although, the accuracy requirement is achieved (as expected), for the inexact method many more iterations are necessary to reach the required precision. An explanation is offered by the fact that the right-hand side is very close to an eigenvector of $\mathbf{A}$ and convergence for general right-hand sides is much slower.

8. Numerical experiments. In this section we conduct an experiment with inexact CG and it's variants from Section 6. For experiments with inexact GMRES we refer the reader to [2]. All experiments are done in Matlab.

The linear system comes from the computation of quark propagators using Wilson fermions in Quantum Chromodynamics. The matrix $\mathbf{D}_W$ is `CONF6.0-0.0014x4.2000` from the Matrix Market. This matrix is complex valued and contains 3072 unknowns. The matrix has the following property, e.g., [7], $\mathbf{\Gamma_5}\,\mathbf{D}_W = \mathbf{D}_W^*\,\mathbf{\Gamma_5}$ with $\mathbf{\Gamma_5} \equiv \mathbf{I} \otimes (\gamma_5 \otimes \mathbf{I}_3)$ and

$$\gamma_5 \equiv \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

The Hermitian matrix $\mathbf{A}$ is now given by $\mathbf{A} = \mathbf{\Gamma_5}\,\mathbf{D}_W$. This matrix is highly indefinite. For the right-hand side we have taken a complex random vector of unit length. To simulate an inexact matrix-vector product we have added in step j of CG, a random complex vector (it can be proven that $\|\mathbf{A}\|_2 \leq 8$ and we have not taken into account the norm of $\mathbf{A}$ in our experiments).

Figure 8.1 shows the results for inexact CG, Orthores and Rutishauser's variant when a residual precision of $\mathcal{O}(\varepsilon)$ is required with $\varepsilon = 10^{-8}$. The left picture shows the results for a constant precision ($\eta_j = \varepsilon$) and the right picture for the relaxation strategy from Section 6.1 ($\eta_j = \varepsilon/\rho_j$).

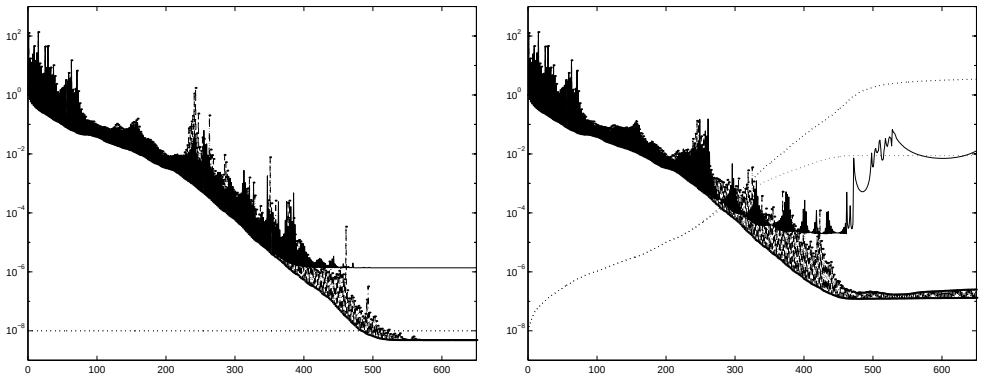

FIG. 8.1: True residuals CG (solid), Orthores (- ·), Rutishauser's variant (dots), η_j (dotted). In both pictures $\varepsilon = 10^{-8}$. Left picture: $\eta_j = \varepsilon$. Right picture: $\eta_j = \varepsilon/\rho_j$.

For $\eta_j = 10^{-8}$ we see that the three-term recurrence is superior to the coupled two-term recurrence. This can be explained by our analysis and the large residuals in the initial steps. This advantage remains if we apply the relaxation strategy from Section 6.1 (although we lose some additional digits compared to the constant precision case).

9. Conclusions and outlook. In this paper we have investigated the effect of approximately computed matrix-vector products on the convergence and accuracy of various Krylov subspace methods. This analysis was used to derive suitable relaxation strategies for these methods. It confirms the empirical results of Bouras et al. in [2, 3]. Furthermore, it was shown that for the Conjugate Gradient method the three-term recurrence can offer advantages over the standard coupled two-term recurrence in case the matrix is indefinite and suffers from large intermediate residuals or peaks in the convergence curve. This was illustrated in Section 8.

For methods like Richardson iteration and Chebyshev iteration it is necessary that the residuals are computed in an incremental matter for a successful relaxation strategy. We illustrated, by the example of CG versus Orthores for indefinite problems, that it is the underlying way the Krylov subspace is constructed that is of importance. By comparing inexact FOM and inexact GMRES we saw that the optimality properties of the residuals are not of influence on the attainable accuracy in the end. Therefore, a relaxation strategy for GMRES should also work for FOM, since the Krylov subspace is constructed in the same matter, i.e., using inexact Arnoldi.

Studying the convergence of the inexact methods is a more difficult problem. Stationary methods construct residual polynomials that are small everywhere on a predefined interval. For these types of methods we could prove that with our relaxation strategies convergence is as fast as for the exact method. For GMRES and CG this is a much more difficult problem.

In future work we want to apply the observations in this paper to the simulation of overlap fermions (as mentioned in the beginning of the introduction) and combine this with the work in [27] for the computation of the matrix sign-function. Furthermore, we plan to extend the analysis in this paper to a rounding error analysis for the different variants of CG for indefinite Hermitian systems (and the BiCG method) in order to understand the effect of the different types of breakdown on the residual gap.

Acknowledgment. The authors are thankful to Valeria Simoncini for providing them with a copy of the slides from [23].

References

[1] R. E. Bank and T. F. Chan, *An analysis of the composite step biconjugate gradient method*, Numer. Math., 66 (1993), pp. 295–319. 13

[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. 1, 6, 18, 19, 21, 22

[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. 1, 13, 22

[4] P. N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 58–78. 18

[5] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408. 1

[6] L. Fox and I. B. Parker, *Chebyshev polynomials in numerical analysis*, Oxford University Press, London, 1972. 8, 9, 10

[7] A. Frommer, T. Lippert, B. Medeke, and K. Schilling, eds., *Numerical Challenges in Lattice Quantum Chromodynamics*, Lecture Notes in Computational Science and Engineering, Springer Verlag, Heidelberg, 2000. Proceedings of the International Workshop, University of Wuppertal, August 22-24, 1999. 1, 21

[8] E. Giladi, G. H. Golub, and J. B. Keller, *Inner and outer iterations for the Chebyshev algorithm*, SIAM J. Numer. Anal., 35 (1998), pp. 300–319 (electronic). 2

[9] G. H. Golub and M. L. Overton, *The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems*, Numer. Math., 53 (1988), pp. 571–593. 2, 10

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore, London, 3rd ed., 1996. 3, 4, 8, 12, 17

[11] G. H. Golub and Q. Ye, *Inexact preconditioned conjugate gradient method with inner-outer iteration*, SIAM J. Sci. Comput., 21 (1999), pp. 1305–1320. 2

[12] G. H. Golub, Z. Zhang, and H. Zha, *Large sparse symmetric eigenvalue problems with homogeneous linear constraints: the Lanczos process with inner-outer iterations*, in Proceedings of the International Workshop on Accurate Solution of Eigenvalue Problems (University Park, PA, 1998), vol. 309, 2000, pp. 289–306. 2

[13] A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63. 14

[14] ———, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–551. 2, 13, 16

[15] ———, *Iterative Methods for Solving Linear Systems*, vol. 17 of Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997. 7

[16] A. Greenbaum, V. L. Druskin, and L. A. Knizhnerman, *On solving indefinite symmetric linear systems by means of the Lanczos method*, Zh. Vychisl. Mat. Mat. Fiz., 39 (1999), pp. 371–377. 13

[17] M. H. Gutknecht, *Lanczos-type solvers for nonsymmetric linear systems of equations*, in Acta Numerica, 1997, Cambridge Univ. Press, Cambridge, 1997, pp. 271–397. 3

[18] M. H. Gutknecht and S. Röllin, *The Chebyshev iteration revisited*, preprint submitted to elsevier science, July 2001. 2, 11

[19] M. H. Gutknecht and Z. Strakoš, *Accuracy of two three-term and three two-term recurrences for Krylov space solvers*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 213–229 (electronic). 2, 11, 15, 16

[20] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436 (1953). 12, 14

[21] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. 19, 20

[22] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput., 14 (1993), pp. 461–469. 2

[23] V. Simoncini and D. B. Szyld, *Flexible inner-outer Krylov methods (and inexact Krylov methods)*. presentation, Zürich, February 20 2002. 2, 23

[24] G. L. G. Sleijpen, H. A. van der Vorst, and D. R. Fokkema, *BiCGstab(ℓ) and other hybrid Bi-CG methods*, Numer. Algorithms, 7 (1994), pp. 75–109. 2, 13

[25] G. L. G. Sleijpen, H. A. van der Vorst, and J. Modersitzki, *Differences in the effects of rounding errors in Krylov solvers for symmetric indefinite linear systems*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 726–751 (electronic). 11

[26] G. W. Stewart and J.-g. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, California, 1990. 20

[27] J. van den Eshof, A. Frommer, T. Lippert, K. Schilling, and H. van de Vorst, *Numerical methods for the QCD overlap operator: I. sign-function and error bounds*, Comput. Phys. Comm., 146 (2002), pp. 203–224. 1, 22

[28] H. A. van der Vorst and C. Vuik, *GMRESR: a family of nested GMRES methods*, Numer. Linear Algebra Appl., 1 (1994), pp. 369–386. 2

[29] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965. 20

[30] H. Woźniakowski, *Numerical stability of the Chebyshev method for the solution of large linear systems*, Numer. Math., 28 (1977), pp. 191–209. 11

1

Contents.