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 impact of approximately computed matrix-vector products on the convergence and attainable accuracy of several Krylov subspace solvers. 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 of the particular method. The obtained insight is used to tune the precision of the matrix-vector product in every iteration step in such a way that an overall efficient process is obtained. Our analysis confirms the empirically found relaxation strategy of Bouras and Frayssé for the GMRES method proposed in [2]. Furthermore, we give an improved version of a strategy of Bouras, Frayssé, and Giraud [3] for the Conjugate Gradient method.

1. Introduction. There is a class of linear problems where the coefficient matrix cannot be stored explicitly in computer memory but where the matrix-vector products can be computed relatively cheaply using an approximation technique. For this type of problems direct methods are not attractive. Krylov subspace methods for solving linear systems of equations require, in every iteration step, basic linear algebra operations, like adding vectors and doing inner products, and, usually, one or two matrix-vector products. This makes this class of solution methods very attractive for this class of problems since we can very easily replace the matrix-vector product in a particular Krylov subspace method with some approximation.

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. In this paper we investigate the impact of approximately computed, or inexact, matrix-vector products on the convergence and attainable accuracy of various Krylov subspace methods. Our analysis should provide further insight into the *relaxation strategies* 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 which explains the term relaxation. Various researchers have reported that this strategy works remarkably well for practical problems.

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]. We also like to refer to independent work of Simoncini and Szyld presented in [25]. Their approach is based on an orthogonality condition, but the fundamental ideas seem to be related (for more details, see below §9).

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

*Submitted March 5, 2002. This is a revised version dated June 5, 2003

† 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.

of Krylov subspace methods since, in the latter case, 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 norm of 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 recurrences, e.g., [26, 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, 23, 30, 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. 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

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

In this section we define the general class of *Krylov subspace methods* for solving this linear system and we collect some of their properties in terms of matrix formulations.

Before we continue we define some notation. The vector e_k denotes the k-th 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 analysis and derivation of a large class of iterative methods for solving (2.1). The Krylov subspace of order k (generated by the matrix $\mathbf{A}$ and the vector $\mathbf{b}$) 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 solution methods for which the residual, $\mathbf{r}_j$, corresponding to the constructed iterate in step j belongs to the space $\mathcal{K}_{j+1}$ and $\mathbf{r}_0 = \mathbf{b}$. Iterative solution methods with this property are called Krylov subspace methods. If we, furthermore, assume for all $j \leq k$ that the first j residuals span the j-th Krylov space, then the residuals provide a sequence that after k steps of the subspace method can be summarized by the following matrix relation

$$\mathbf{A}\mathbf{R}_k = \mathbf{R}_{k+1}\underline{S}_k, \quad \mathbf{R}_k e_1 = \mathbf{b}, \quad \text{with} \quad \vec{1}^*\underline{S}_k = \vec{0}^*. \tag{2.3}$$

Here, the matrix $\mathbf{R}_k$ is an n by k matrix with as j-th column $\mathbf{r}_{j-1}$, and $\underline{S}_k$ is a $k+1$ by k upper Hessenberg matrix. The last condition in (2.3) for the Hessenberg matrix is a necessary condition if $\mathbf{r}_j$ is a residual that corresponds to some approximate solution from the space $\mathcal{K}_j$, see [17, Section 4.4]. 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

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

Now, let

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

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

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

This shows that $\mathbf{r}_j = \mathbf{b} - \mathbf{A}\mathbf{x}_j$ if $\mathbf{x}_j$ is as in (2.5). Hence, for this choice the iterate $\mathbf{x}_j$ is *consistent* with the residual $\mathbf{r}_j$.

Moreover, we can get a recursion for the iterates $\mathbf{x}_j$ by substituting $\mathbf{R}_k = \mathbf{b}\vec{1}^* - \mathbf{A}\mathbf{X}_k$ in (2.3). This shows that

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

Some Krylov subspace methods use the recursions in (2.3) or (2.6) explicitly in their implementation. An example is the Chebyshev method where the iterates are computed with the, in this case, three-term relation in (2.6), see also Section 5.

It is common to view Krylov subspace methods as polynomial based iteration methods where the residuals are characterized as matrix polynomials in $\mathbf{A}$ that act on the vector $\mathbf{b}$, see e.g., [5]. This viewpoint plays an important role in the convergence analysis of a large number of Krylov subspace methods. The property of $\underline{S}_k$ that the columns sum up to zero, is equivalent to the fact that the residual polynomials have the interpolatory constraint that they are one in zero. We will, however, not use this polynomial interpretations and mostly use the matrix formulation and exploit algebraic properties of the matrix $\underline{S}_k$.

We conclude this section with a useful property of the Hessenberg matrix $\underline{S}_k$ that we will frequently use in the remainder of this paper.

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

Proof. The existence of the LU-decomposition of S_k follows from the fact that each principal submatrix of S_k is nonsingular, see, for instance, [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

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

2.1. Derivation from Krylov decompositions. For theoretical purposes and future convenience, we summarize in this section some facts about a, so-called, *Krylov decomposition* given by

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

where $\mathbf{C}_k$ is an n by k matrix and $\underline{T}_k$ is a $k+1$ by k upper Hessenberg matrix. The column space of $\mathbf{C}_k$ is a subspace of the Krylov space $\mathcal{K}_k$ but the columns, $\mathbf{c}_j$, are not necessarily residuals corresponding to approximations from $\mathcal{K}_j$. However, 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 $\Gamma_k \equiv \mathrm{diag}(\vec{\gamma}_{k-1})$; 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 $\mathbf{c}_j$. In terms of the polynomial interpretation of Krylov subspace methods, this construction of the residual sequence can be viewed as obtaining the residual polynomials by scaling the polynomials, generated by the coefficients in $\underline{T}_k$, such that they are one in zero. 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 (2.8) and the first statement of the following lemma. (For ease of future reference, we formulate the lemma slightly more general than needed here.)

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

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

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

Proof. The first expression follows from a combination of $e_1 - \underline{T}_j(T_j^{-1}e_1) = e_1 - \Gamma_{j+1}^{-1}\underline{S}_j S_j^{-1}\Gamma_j e_1$ and (2.4). For the second expression we notice that $I - \underline{T}_j\underline{T}_j^{\dagger}$ 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 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. It should be noted that the matrix $(\Upsilon_{k+1}\Theta_{k+1}^{-2})^{-1}$ can be decomposed into simple factors since $\Upsilon_{k+1} = \Gamma_{k+1}J_{k+1}^{-1}$. These latter observations are related to the well-known fact, see e.g., [5, Section 2.5], that *minimal residual* polynomials, or *Kernel* polynomials, can be generated efficiently using coupled recurrences.

3. Inexact Krylov subspace methods. In the previous section we collected some general properties of Krylov subspace methods. There is a class of applications for which it is very costly to compute the matrix-vector product to high precision. The original motivation for the research in this paper was a linear system that occurs in simulations in quantum chromodynamics (QCD) [7]. In this area the so-called *overlap formulation* has initiated a lot of research in solving linear systems of the form

$$(r\mathbf{\Gamma_5} + \mathrm{sign}(\mathbf{Q}))\mathbf{x} = \mathbf{b}, \quad \|\mathbf{b}\| = 1 \quad (r \geq 1), \tag{3.1}$$

where $\mathbf{Q}$ and $\mathbf{\Gamma_5}$ are sparse Hermitian indefinite matrices. The matrix $\mathrm{sign}(\mathbf{Q})$ is the so-called *matrix sign function*, see, e.g., [10, p. 372]. This matrix is dense and is only known implicitly since we are only given the matrices $\mathbf{Q}$ and $\mathbf{\Gamma_5}$. Realistic simulations require in the order of one to ten million unknowns. Usually, Equation (3.1) 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 vector iteration 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 with a Lanczos approximation. For an overview and comparison of methods used in this context we refer to [29].

In this paper we consider the more general problem of solving (2.1) where we assume that we are given some approximation $\mathcal{M}_\eta : \mathbb{C}^n \to \mathbb{C}^n$ with the property that

$$\mathcal{M}_\eta(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{g} \quad \text{with} \quad \|\mathbf{g}\|_2 \leq \eta\|\mathbf{A}\|_2\,\|\mathbf{y}\|_2. \tag{3.2}$$

It is, furthermore, assumed that the smaller η is chosen, the more time consuming this approximation becomes to construct.

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 $\mathbf{y}$. If this approximation is accomplished by replacing the matrix-vector product with the approximation $\mathcal{M}_\eta$, we will refer to the resulting method as an *inexact* Krylov subspace method. It is equivalent to view an inexact Krylov subspace method as a variant of an exact Krylov subspace method where a perturbation $\mathbf{g}_{j-1}$ is added to the matrix-vector product in step j with $\mathbf{g}_{j-1}$ such that $\|\mathbf{g}_{j-1}\|_2 \leq \eta_{j-1}\|\mathbf{A}\|_2\,\|\mathbf{y}\|_2$.

Due to the existence of the errors, $\mathbf{g}_{j-1}$, 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 in this paper 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.

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

$$\eta_j = \max\left\{\frac{\varepsilon}{\|\mathbf{b} - \mathbf{A}\mathbf{x}_j\|_2}, \varepsilon\right\}. \tag{3.3}$$

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 as soon as the method starts to converge, that is, 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 (3.3) 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 approximate residual as computed in the inexact Krylov subspace method (cf. §3.2) can serve as an alternative. We conclude with the remark that this condition was derived empirically in [2] based on the experience of the authors with a large number of experiments.

3.2. The analysis of inexact Krylov subspace methods. In the remainder of this paper we will see that, for the methods that we consider, the approximate residuals, $\mathbf{r}_j$, computed in the inexact Krylov subspace method now satisfy the perturbed relation

$$\mathbf{A}\mathbf{R}_k + \mathbf{F}_k = \mathbf{R}_{k+1}\underline{S}_k, \quad \mathbf{R}_k e_1 = \mathbf{b}, \quad \text{with} \quad \vec{1}^*\underline{S}_k = \vec{0}^*. \tag{3.4}$$

The columns of the matrix $\mathbf{F}_k$ are a function of the errors in the matrix-vector products. Furthermore, $\mathbf{x}_j$ still satisfies (2.5) (or equivalently (2.6)) because of the assumption of exact arithmetic. For the moment we assume these two properties but we stress that their validity must be checked for every inexact Krylov subspace method which is obtained by replacing in a particular method the exact matrix-vector product with some approximation.

As a consequence of the perturbation term $\mathbf{F}_k$, the vector $\mathbf{r}_k$ is usually not a residual anymore for the approximate solution $\mathbf{x}_k$. Therefore, we will refer to the vector $\mathbf{r}_k$ as the *computed residual* in contrast to the *true residual* defined by $\mathbf{b}-\mathbf{A}\mathbf{x}_k$. In the analysis of inexact Krylov methods, the true residuals are the quantities of interest and we have

$$\|\mathbf{b}-\mathbf{A}\mathbf{x}_k\|_2 \leq \|\mathbf{r}_k\|_2 + \|\mathbf{r}_k - (\mathbf{b}-\mathbf{A}\mathbf{x}_k)\|_2. \tag{3.5}$$

This inequality forms the basis of our analysis. If the computed residuals, for sufficiently large k, become small compared to the residual gap, then it follows from (3.5) that the stagnation level of the inexact Krylov subspace method is determined by the *residual gap*, the difference between the computed residual and the true residual. Furthermore, in the early iterations the norm of the computed residuals is large compared to the size of the residual gap. This shows that the initial convergence of the true residuals is determined by the residuals computed in the inexact Krylov subspace method.

In the coming sections, we will analyze the effect of inexact matrix-vector products and relaxation strategies as in (3.3) for different Krylov subspace methods by writing the residual relation into the form (3.4) and by bounding the residual gap. 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. It turns out that, under our assumptions, a general expression can be given for the residual gap. We give this expression in the next section and exploit it in the remainder of this paper.

For the analysis in this paper, we assume the use of exact arithmetic operations: here, we are interested in the effect of errors in the matrix-vector multiplication, but it is also 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 (3.1) mentioned in the beginning of Section 3, 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.

3.3. A general expression for the residual gap. The goal is to get an expression for the residual gap. Assuming that $\mathbf{x}_k$ is of the form (2.6) and the computed residuals satisfy (3.4), then we find, using again (2.4), the following expression

$$\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k) = \mathbf{r}_k - \mathbf{r}_0 + \mathbf{A}\mathbf{R}_k S_k^{-1} e_1 = -\mathbf{F}_k S_k^{-1} e_1 = -\sum_{j=1}^{k} \mathbf{f}_{j-1} e_j^* S_k^{-1} e_1. \tag{3.6}$$

This shows that the expression for the gap is a linear combination of the columns of $\mathbf{F}_k$, i.e., the vectors $\mathbf{f}_{j-1}$, with the coefficients $-(e_j^* S_k^{-1} e_1)$ which determine the propagation of the perturbations through the recurrences. Our approach for bounding the gap 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 frequently be used.

LEMMA 3.1. *Let $\underline{T}_k$ be upper Hessenberg and of full rank. For $j \leq k$, we have*

$$|e_j^* \underline{T}_k^\dagger e_1| \leq \|\underline{T}_k^\dagger\|_2 \frac{1}{\|\vec{\gamma}_{j-1}\|_2}, \quad |e_j^* T_k^{-1} e_1| \leq \|\underline{T}_k^\dagger\|_2 \left(\frac{1}{\|\vec{\gamma}_{j-1}\|_2} + \frac{1}{|\gamma_k|} \right). \tag{3.7}$$

Proof. To prove (3.7), 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

$$e_j^* \underline{T}_k^\dagger e_1 = e_j^* \underline{T}_k^\dagger \frac{\vec{\gamma}_{j-1}}{\|\vec{\gamma}_{j-1}\|_2^2} = e_j^* \underline{T}_k^\dagger \frac{e_j}{\gamma_{j-1}} \text{ and } e_j^* T_k^{-1} e_1 = e_j^* \underline{T}_k^\dagger e_1 - e_j^* \underline{T}_k^\dagger \frac{e_{k+1}}{\gamma_k}. \tag{3.8}$$

and (3.7) 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 $\|T_m^{-1}\|_2 \geq \|\underline{T}_k{}^\dagger\|_2$ 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.

4. Inexact Richardson iteration. One of the simplest iterative method for linear systems is *Richardson iteration*, e.g., [15]. This method allows a straightforward analysis, however, it already demonstrates some important aspects of our analysis.

Therefore, Richardson iteration is a useful starting point. With a perturbed matrix-vector product, this method is described by the following recurrences for $j = 1, \ldots, k$ (with $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{r}_0 = \mathbf{b}$)

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

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

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:

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

where $\lambda_{\min}$ and $\lambda_{\max}$ 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 (3.4) 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 (3.6) and, using $e_j^* S_k^{-1} e_1 = \alpha$, we get the following bound on the norm of the residual gap

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

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 $\mathcal{C}(\mathbf{A}) \equiv \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2$. We stress that the residual gap for this simple iteration method can be obtained by comparing the recursions for $\mathbf{r}_j$ and $\mathbf{b} - \mathbf{A}\mathbf{x}_j$ directly. We have used here a slightly more involved approach to demonstrate the use of our general formula (3.6) which becomes more important when studying more advanced methods.

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.

THEOREM 4.1. *Let* $\overline{\mathbf{r}}_k$ *satisfy* (4.1) *with* $\eta_j = 0$, *and let* $\mathbf{r}_k$ *satisfy* (4.1) *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

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

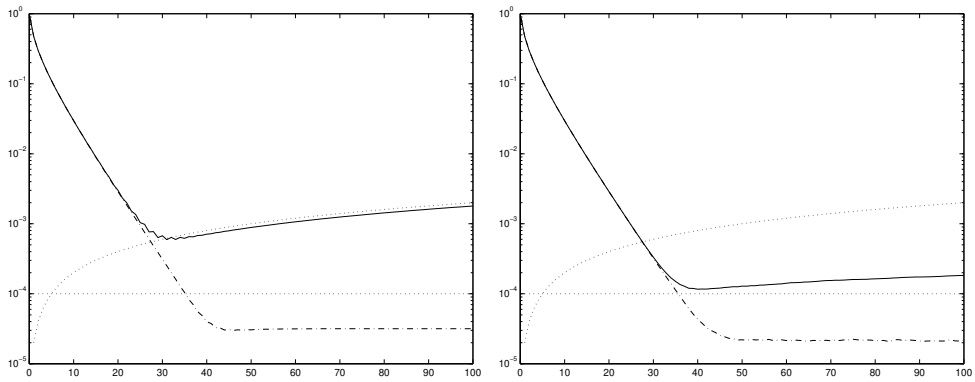

FIG. 4.1: Richardson iteration with $\eta_j = 10^{-5}/\|\mathbf{r}_j\|_2$, true residuals (—), norm computed residual (- ·) and the quantities $10^{-5}\mathcal{C}(\mathbf{A})$, $2j10^{-5}$ (both dotted) as a function of j. The matrix $\mathbf{A}$ has dimension 1000 and $\mathcal{C}(\mathbf{A}) = 10$. Left picture: errors have all components equal. Right picture: random errors.

Since $\overline{\mathbf{r}}_k$ 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 (3.3).

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 (3.6)

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

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 $[\lambda_{\min}, \lambda_{\max}]$ to the interval $[-1, 1]$, so (for example)

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

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 $\gamma_j = c_j(\phi(0))$. A recursion for the approximate solutions $\mathbf{x}_j$ is given by (2.6). For convenience of the reader, we give the resulting recurrence relations: for $j = 2, \ldots, k$, we have

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

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

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 computed residuals in the inexact Chebyshev method satisfy (3.4) 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 (3.6), we have to bound $e_j^* S_k^{-1} e_1$, this is accomplished in the following lemma.

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

$$|e_j^* S_k^{-1} e_1| = |e_j^* T_k^{-1} e_j| \leq \frac{2\alpha}{\sqrt{\beta^2 - 1}} = \frac{2}{\sqrt{\lambda_{\max}\lambda_{\min}}} = 2\frac{\sqrt{\mathcal{C}(\mathbf{A})}}{\|\mathbf{A}\|_2}. \tag{5.5}$$

Proof. Using (2.4) we see that

$$e_j^* S_k^{-1} e_1 = e_j^* S_k^{-1}(e_1 - S_k(S_{j-1}^{-1} e_1)) = e_j^* S_k^{-1}(e_1 - \underline{S}_{j-1}(S_{j-1}^{-1} e_1)) = e_j^* S_k^{-1} e_j.$$

The first equality now follows from the relation $S_k = \Gamma_k T_k \Gamma_k^{-1}$.

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 $i = 1, 2, \ldots$, see Lemma A.1 in Appendix A. 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 and (3.6) 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 theorem 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.

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

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

Proof. If we subtract (2.3) from (3.4), 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 $\lambda_{\min}$. We will show that $\|\overline{\mathbf{r}}_k - \mathbf{r}_k\|_2$ is maximal when for all perturbations we have $\mathbf{f}_j = \varepsilon\|\mathbf{A}\|_2\mathbf{v}_{\min}$ (or $\mathbf{F}_k = \varepsilon\|\mathbf{A}\|_2\mathbf{v}_{\min}\vec{1}^*$). 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 $\mathbf{d}_j \equiv (\mathbf{r}_j - \overline{\mathbf{r}}_j)\gamma_j$. 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

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

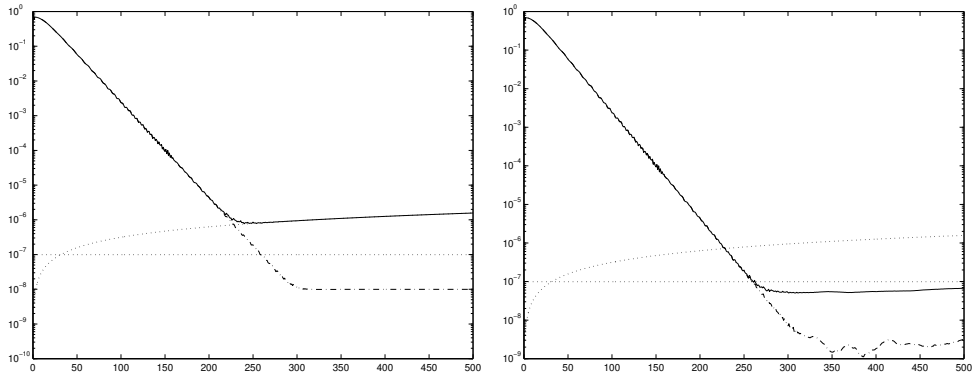

FIG. 5.1: Chebyshev iteration with $\eta_j = 10^{-10}/\|\mathbf{r}_j\|$, true residuals (—), norm computed residual (- ·) and the quantities $10^{-10}\mathcal{C}(\mathbf{A})$, $2j10^{-10}\sqrt{\mathcal{C}(\mathbf{A})}$ (both dotted) as a function of j. The matrix $\mathbf{A}$ has dimension 100 and $\mathcal{C}(\mathbf{A}) = 1000$. Left picture: errors have all components equal. Right picture: random errors.

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

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

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

In order to solve (5.7) with $\mathbf{F}_k = \varepsilon\|\mathbf{A}\|_2\mathbf{v}_{\min}\vec{1}^*$, 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 $\varepsilon\|\mathbf{A}\|_2\mathbf{v}_{\min}$, which is

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

By noting that $0 \leq c_k(-1)/\gamma_k \leq 1$ 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 [32], 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 forward 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 [27]). 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 (3.6) 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.

For the Chebyshev method with inexact preconditioning, called *flexible preconditioning* in [9], convergence results have been established by Golub and Overton [9] for $\eta_j = \eta$ but where η can be modest. Moreover, under certain assumptions for the cost of the flexible preconditioner, it is shown in [8] that a fixed threshold strategy is optimal. It is not difficult to see that, if one sets the preconditioner to $\mathbf{M} = \mathbf{I}$, the residuals of this flexible process satisfy the perturbed residual relation given in (3.4). However, since the perturbation is the consequence of inexact preconditioning, instead of inexact matrix-vector products, we still have that $\mathbf{r}_j = \mathbf{b} - \mathbf{A}\mathbf{x}_j$. This shows that, although there are common elements, flexible preconditioning is different from the case of inexact matrix-vector products. Since, for the latter case there is also an accuracy issue.

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 gradients that are conjugate. The most popular formulation of the CG method is due to Hestenes and Stiefel [20, Section 3] and consists of three coupled two-term recurrences. For $j = 1, \ldots, k$ this method, with inexact matrix-vector product, is defined by the recurrences

$$\mathbf{c} = \mathbf{A}\mathbf{p}_{j-1} + \mathbf{g}_{j-1} \tag{6.1}$$

$$\mathbf{r}_j = \mathbf{r}_{j-1} - \alpha_{j-1}\mathbf{c} \tag{6.2}$$

$$\mathbf{x}_j = \mathbf{x}_{j-1} + \alpha_{j-1}\mathbf{p}_{j-1} \tag{6.3}$$

$$\mathbf{p}_j = \mathbf{r}_j + \beta_{j-1}\mathbf{p}_{j-1}, \tag{6.4}$$

with

$$\alpha_{j-1} \equiv \frac{\|\mathbf{r}_{j-1}\|_2^2}{\mathbf{p}_{j-1}^*\mathbf{c}} \quad \text{and} \quad \beta_{j-1} \equiv \frac{\|\mathbf{r}_j\|_2^2}{\|\mathbf{r}_{j-1}\|_2^2}, \tag{6.5}$$

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 (6.2) 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 about 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 in the end as for Chebyshev iteration in the previous section.

We define

$$\widetilde{U}_k \equiv \begin{bmatrix} 1 & -\beta_0 & & & \\ & 1 & -\beta_1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & -\beta_{k-2} \\ & & & & 1 \end{bmatrix}, \Delta_k \equiv \begin{bmatrix} \alpha_0 & & & & \\ & \alpha_1 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \alpha_{k-1} \end{bmatrix}.$$

This gives us the following equivalent matrix formulations of the recurrences of the inexact CG method

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

Combining these relations shows that

$$\mathbf{A}\mathbf{R}_k + (\mathbf{G}_k\widetilde{U}_k) = \mathbf{R}_{k+1}(\underline{J}_k\Delta_k^{-1}\widetilde{U}_k) \text{ and } -\mathbf{R}_k = \mathbf{X}_{k+1}(\underline{J}_k\Delta_k^{-1}\widetilde{U}_k). \tag{6.6}$$

We see that (3.4) and (2.6) are satisfied for this method with $\underline{S}_k \equiv \underline{J}_k\Delta_k^{-1}\widetilde{U}_k$ and $\mathbf{F}_k \equiv \mathbf{G}_k\widetilde{U}_k$. Therefore, we can use our familiar formula (3.6) to get an expression for the residual gap:

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

This expression can also be obtained by an inductive combination of (6.2) and (6.3). This simpler argument, that avoids the matrix formulation, was used in [26, 14]. However, the present argument explains how CG fits in the general framework of this paper. Moreover, for the conclusions below we need the matrix formulation anyway.

From $\|\mathbf{g}_j\|_2 \leq \eta_j\|\mathbf{A}\|_2\|\mathbf{p}_j\|_2$, we get the following bound on the norm of the residual gap:

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 \leq \sum_{j=0}^{k-1}\eta_j|\alpha_j|\|\mathbf{A}\|_2\|\mathbf{p}_j\|_2. \tag{6.7}$$

Thus, the problem of deriving relaxation strategies for the Conjugate Gradient method amounts to bounding $|\alpha_j|\|\mathbf{p}_j\|_2$. We do this in the remainder of this section.

The Conjugate Gradient method (CG) is intimately connected with the Lanczos method, e.g., [10, Chapter 9]. In order to continue, we introduce for theoretical purposes the following *inexact* Lanczos process

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

where $\underline{T}_k \equiv \Gamma_{k+1}^{-1}\underline{S}_k\Gamma_k$, $\Gamma_k \equiv \mathrm{diag}(\vec{\gamma}_{k-1})$, $\gamma_j \equiv (-1)^j\|\mathbf{r}_j\|_2^{-1}$, $\mathbf{V}_k \equiv \mathbf{R}_k\Gamma_k$, and $\widetilde{\mathbf{F}}_k \equiv \mathbf{F}_k\Gamma_k$. From (6.6) and Section 2 it follows that $\mathbf{x}_j = \mathbf{R}_j S_j^{-1}e_1 = \mathbf{V}_j T_j^{-1}e_1$ and combining this with (6.3) shows that

$$\alpha_j\mathbf{p}_j = \mathbf{V}_k(T_{j+1}^{-1}e_1 - T_j^{-1}e_1). \tag{6.9}$$

We will use this relation to bound $|\alpha_j|\|\mathbf{p}_j\|_2$.

6.1. The case of T_k positive definite. First we assume that T_k is positive definite. In the previous section we reduced the problem of bounding the gap to bounding $|\alpha_j|\|\mathbf{p}_j\|_2$. We will do this using (6.9) and the following result.

LEMMA 6.1. *Let $j < k$. Then,*

$$T_{j+1}^{-1}e_1 - T_j^{-1}e_1 = T_{j+1}^{-1}\frac{e_{j+1}}{\gamma_j} = \frac{\vec{\gamma}_j}{\vec{\gamma}_j^* T_{j+1}\vec{\gamma}_j}. \tag{6.10}$$

Proof. First observe that

$$T_{j+1}^{-1}e_1 - T_j^{-1}e_1 = T_{j+1}^{-1}(e_1 - T_{j+1}T_j^{-1}e_1) = T_{j+1}^{-1}(e_1 - \underline{T}_j T_j^{-1}e_1).$$

Now, the first identity in (6.10) follows from Lemma 2.2.

Since $\vec{\gamma}_{j+1}^*\underline{T}_{j+1} = \vec{0}^*$, we see that $\vec{\gamma}_j^* T_{j+1} = \delta e_{j+1}^*$ for some scalar δ. Multiplication from the right with $\vec{\gamma}_j$, shows that $\delta = \vec{\gamma}_j^* T_{j+1}\vec{\gamma}_j/\gamma_j$. Since T_{j+1} is symmetric, we find $\vec{\gamma}_j = \delta T_{j+1}^{-1}e_{j+1}$, which leads to second identity. ∎

We combine this lemma with (6.9) and arrive at the estimate

$$|\alpha_j|\|\mathbf{p}_j\|_2 \le \|\mathbf{V}_k\|_2\|T_k^{-1}\|_2\rho_j, \text{ with } \rho_j \equiv \frac{1}{\|\vec{\gamma}_j\|_2} = \left(\sum_{i=0}^{j}\|\mathbf{r}_i\|_2^{-2}\right)^{-1/2}. \tag{6.11}$$

We substitute this estimate in (6.7) and find the following bound on the norm of the residual gap

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 \le \|\mathbf{V}_k\|_2\|\mathbf{A}\|_2\|T_k^{-1}\|_2\sum_{j=0}^{k-1}\eta_j\rho_j. \tag{6.12}$$

This estimate can be further bounded by using $\|\mathbf{V}_k\|_2 \le \|\mathbf{V}_k\|_F \le \sqrt{k}$. In practice, this turns out to be crude since $\|\mathbf{V}_k\|_2$ is close to one or only a modest multiple of one. If $\mathbf{A}$ is symmetric positive definite, then, in the exact case, $\|T_k^{-1}\|_2 \le \|\mathbf{A}^{-1}\|_2$. In the inexact case, $\|\mathbf{A}\|_2\|T_k^{-1}\|_2$ can be viewed as an approximation to $\mathcal{C}(\mathbf{A})$. It is tempting to refer to the results of Paige [22] for perturbed Lanczos processes to bound this quantity. However, the perturbations in our context are not assumed to be uniformly bounded. In fact, they are allowed to grow during the process. Therefore, we cannot make use of his results. Of course, we can monitor this quantity during the inexact process and, possibly, incorporate this estimate into our tolerance η_j.

Bouras, Frayssé and Giraud proposed in [3], following their work for inexact GMRES and (3.3), a relaxation strategy for the CG method where they take

$$\eta_j = \max\left\{\frac{\varepsilon}{\|\mathbf{r}_j\|_2}, \varepsilon\right\}. \tag{6.13}$$

If we take the larger tolerance $\eta_j = \frac{\varepsilon}{\rho_j}$ (since $\rho_j \le \|\mathbf{r}_j\|_2$), we have from (6.12) that

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 \le \varepsilon k\|\mathbf{V}_k\|_2\|\mathbf{A}\|_2\|T_k^{-1}\|_2. \tag{6.14}$$

We saw that our analysis of the residual gap helps to provide insight into the practical success of the Bouras-Frayssé-Giraud condition (6.13) and even suggests that we can relax more aggressively than previously proposed. Indeed, numerical experiments with symmetric positive definite matrices $\mathbf{A}$ confirm this.

An alternative for bounding $|\alpha_j|\|\mathbf{p}_j\|_2$ follows from noticing in (6.10) that, for fixed values of i, the quantities $e_i^*(T_j^{-1}e_1 - T_{j-1}^{-1}e_1)$ have a constant sign for all j (or are zero). Therefore, we have that

$$\|T_j^{-1}e_1 - T_{j-1}^{-1}e_1\|_2 \leq \|T_i^{-1}e_1\|_2 \quad \text{for} \quad i \geq j.$$

This provides a similar bound on $|\alpha_j|\|\mathbf{p}_j\|_2$ as derived by Greenbaum in [14] for the residual gap of CG in order to study the attainable accuracy of the CG method in finite precision computations. She uses that the errors of the CG method are monotonically decreasing in 2-norm in order to bound $\|\alpha_j\mathbf{p}_j\|_2$. In our context this approach is too crude since it does not lead to a relaxation strategy.

6.2. The case of T_k indefinite. 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]. If $\mathbf{A}$ is symmetric indefinite, but nonsingular, then, even in the exact case, T_k will not be definite and we cannot uniformly bound $\vec{\gamma}_j^* T_k \vec{\gamma}_j$ away from zero. We may not expect that Lemma 6.1 leads to useful results for bounding $|\alpha_j|\|\mathbf{p}_j\|_2$ using (6.9). As an alternative, we use the following lemma.

LEMMA 6.2. *Let $j < k$. Then,*

$$T_{j+1}^{-1}e_1 - T_j^{-1}e_1 = \underline{T}_{j+1}^{\dagger}\left(\frac{e_{j+1}}{\gamma_j} - \frac{e_{j+2}}{\gamma_{j+1}}\right). \tag{6.15}$$

Proof. We observe that $\underline{T}_{j+1}^{\dagger}\underline{T}_{j+1}$ is the identity on $j+1$-vectors and conclude that

$$T_{j+1}^{-1}e_1 - T_j^{-1}e_1 = \underline{T}_{j+1}^{\dagger}\left((e_1 - \underline{T}_{j+1}T_j^{-1}e_1) - (e_1 - \underline{T}_{j+1}T_{j+1}^{-1}e_1)\right).$$

The proof can be concluded by rewriting the expressions on the right using Lemma 2.2. ∎

If we use that $\|\underline{T}_{j+1}^{\dagger}\|_2 \leq \|\underline{T}_k^{\dagger}\|_2$ for $k > j$ and, from (6.5), that $\beta_j = \gamma_j^2/\gamma_{j+1}^2$, then we can bound the norm of the residual gap as

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 \leq \|\mathbf{V}_k\|_2\|\mathbf{A}\|_2\|\underline{T}_k^{\dagger}\|_2 \sum_{j=0}^{k-1} \eta_j\|\mathbf{r}_j\|_2\sqrt{1+\beta_j}. \tag{6.16}$$

A similar expression can be found in [26, 14], where the perturbations are assumed to be small and second order terms have been neglected (then it can be proven that $\|\mathbf{A}\|_2\|\underline{T}_k^{\dagger}\|_2 \lesssim \mathcal{C}(\mathbf{A})$). For the choice $\eta_j = \frac{\varepsilon}{\|\mathbf{r}_j\|_2}$, we get, using (6.16),

$$\|\mathbf{r}_k - (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|_2 \leq \varepsilon k\|\mathbf{V}_k\|_2\|\mathbf{A}\|_2\|\underline{T}_k^{\dagger}\|_2 \max_{0\leq j<k}\sqrt{1+\beta_j}. \tag{6.17}$$

We see that, as long as the β_j are bounded, this strategy can work very well. However, practical problems often lead to a matrix $\mathbf{A}$ that is indefinite, for instance in the QCD example discussed in Section 3. In this case there can be very large intermediate residuals caused by an eigenvalue of T_k being 'accidentally' close to zero. 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 (6.17) that the proposed strategy in (6.13) may fail in achieving the required residual precision.

From (6.16) 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 (6.7) and pick

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

If the approximation of the matrix-vector product is computed with an iterative method, then the inner product of $\mathbf{p}_j$ with the 'current' approximation to the matrix-vector 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 6.4.

6.3. The behavior of the computed residuals. Studying the convergence and stagnation level of the computed residuals is a much more difficult topic. Greenbaum [13] showed that the convergence of a slightly perturbed CG process is equal to that of the exact method applied to a 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. Numerical experiments indeed suggest that a relaxation strategy for the accuracy of the matrix-vector products does not spoil the convergence of the computed residuals and they seem to stagnate at a level in the order of ε.

However, a near breakdown of the Conjugate Gradient method can severely alter the behavior of the computed residuals. In this case, $\widetilde{\mathbf{F}}_k$ in (6.8) has some relatively very large columns. We know that for the j-th column of $\widetilde{\mathbf{F}}_k$ we have $\|\widetilde{\mathbf{f}}_{j-1}\| = \|\mathbf{g}_{j-1} - \beta_{j-2}\mathbf{g}_{j-2}\|_2/\|\mathbf{r}_{j-1}\|_2$. 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 ρ_j is as defined in (6.11). Notice that ρ_j can be viewed as the norm of a smoothed residual e.g., [20, Section 7]. We have the following upper bound for the norm of the j-th 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 then we have that $\rho_{j-1} \ll \|\mathbf{r}_{j-1}\|_2$. We see that, when there is a near breakdown, there can be a relatively very large perturbation of the Lanczos relation. One consequence is a large residual gap (as discussed). Another effect is a potential delay in the convergence (or even worse). A simple numerical example of this is given in the next section.

6.4. Variants of the Conjugate Gradient method. Mathematically equivalent variants of the CG method can be derived from the Lanczos method. 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 for the residuals instead of the coupled two-term recurrences of the Hestenes and Stiefel implementation discussed in the previous sections. We start with a short derivation of these alternatives.

Since the CG residuals are multiples of the Lanczos vectors, we can derive the coefficients for the recurrence (2.3) from the Lanczos relation by virtue of (2.9). To see this, we write

$$\underline{T}_k \equiv \begin{bmatrix} \alpha_0 & \beta_0 & & & \\ \beta_0 & \alpha_1 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{k-2} \\ & & & \beta_{k-2} & \alpha_{k-1} \\ & & & & \beta_{k-1} \end{bmatrix}, \quad \underline{S}_k \equiv \begin{bmatrix} \mu_0 & \delta_0 & & & \\ \tau_0 & \mu_1 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \delta_{k-2} \\ & & & \tau_{k-2} & \mu_{k-1} \\ & & & & \tau_{k-1} \end{bmatrix}.$$

The matrix $\underline{T}_k$ is computed using the Lanczos method and we want expressions for the elements of the matrix $\underline{S}_k$. We can do this similar to our derivation of Chebyshev iteration in Section 5. From the necessary property that $\vec{1}^*\underline{S}_k = \vec{0}$, it immediately follows that $\tau_j = -(\mu_j + \delta_{j-1})$ (with $\delta_{-1} = 0$). Using (2.9) we see that $\mu_j = \alpha_j$, $\delta_j = \beta_j(\gamma_j/\gamma_{j+1})$ 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),

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

Computing the residuals and iterates with these coefficients using the recurrences 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

$$\begin{aligned} \mathbf{R}_{k+1}\underline{J}_k &= \Delta\mathbf{R}_k, \quad \Delta\mathbf{R}_k\, U_k = \mathbf{A}\mathbf{R}_k \quad \text{and} \\ \mathbf{X}_{k+1}\underline{J}_k &= \Delta\mathbf{X}_k, \quad \Delta\mathbf{X}_k\, U_k = -\mathbf{R}_k. \end{aligned} \tag{6.18}$$

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

$$\mathbf{R}_k + \mathbf{F}_k = \mathbf{R}_{k+1}\underline{S}_k, \quad \mathbf{R}_k e_1 = \mathbf{b}, \quad \text{with} \quad \vec{1}^*\underline{S}_k = \vec{0}^*,$$

where $\mathbf{F}_k = \mathbf{G}_k$ and, therefore, $\|\mathbf{f}_j\|_2 \le \eta_j \|\mathbf{A}\|_2 \|\mathbf{r}_j\|_2$. 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 notice 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 $\underline{T}_k = \Gamma_{k+1}^{-1}\underline{S}_k\Gamma_k$, $\Gamma_k = \mathrm{diag}(\vec{\gamma}_{k-1})$ and $\gamma_j = (-1)^j\|\mathbf{r}_j\|_2^{-1}$. Combining this with Lemma 3.1 shows that

$$|e_j^* S_k^{-1} e_1| \le \|\underline{T}_k^\dagger\|_2 \frac{1}{\|\mathbf{r}_{j-1}\|_2} \left(\rho_{j-1} + \|\mathbf{r}_k\|_2\right), \tag{6.19}$$

with ρ_{j-1} as defined in (6.11). The general expression for the residual gap (3.6), now leads to the following bound

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

Recall that we assume that the computed residuals ultimately become small enough. Now, assume that we terminate the iterative process for $\|\mathbf{r}_k\|_2 \leq \varepsilon$. In this case we see that the size of the gap is essentially determined by the values of the ρ_j, the η_j, and $\|\underline{T}_k^\dagger\|_2$. Unfortunately, we have no a priori knowledge about the size of $\|\underline{T}_k^\dagger\|_2$. We hope that this quantity is in the order of $\|\mathbf{A}^{-1}\|_2$. For inexact Orthores (and Rutishauser's variant) we propose the following relaxation strategy

$$\eta_j = \frac{\varepsilon}{\rho_j}, \tag{6.20}$$

where ρ_j is given in (6.19). Notice that ρ_j can be computed at little additional cost. With this relaxation strategy we get for the residual gap

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

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 $\|\mathbf{r}_k\|_2 \leq \varepsilon$, the gap is hopefully $\mathcal{O}(\varepsilon)$. An alternative is to pick $\eta_j = \varepsilon/(\varepsilon + \rho_j)$ which somewhat simplifies the resulting expression that bounds the gap.

Let us summarize our findings. 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 (6.13). 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 $\mathbf{A} = \mathrm{diag}(1:100) - 5.2025\,\mathbf{I}$. 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.

6.5. Discussion. For positive definite systems, the standard CG method seems appropriate in the inexact setting. 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 in situations where the matrix $\mathbf{A}$ is not too ill-conditioned and convergence is irregular. Numerical experiments are given in Section 8.

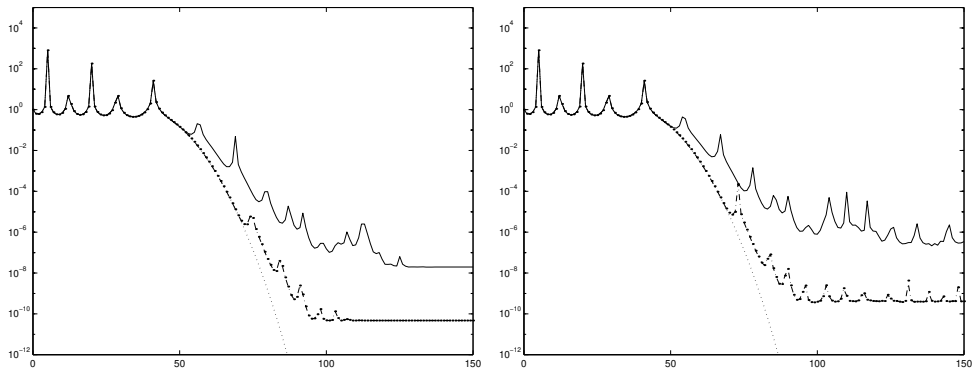

FIG. 6.1: True residuals exact FOM (dotted), CG (–), Orthores (- ·), Rutishauser's variant (dots) as a function of j. In both pictures $\varepsilon = 10^{-10}$. Left picture: $\eta_j = \varepsilon$. Right picture: $\eta_j = \varepsilon/\rho_j$.

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 [19], 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 design choices in the construction of a 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. It is also often observed that the initial convergence speed is comparable to the convergence speed of the exact method. Nevertheless, in a few cases, small perturbations of the matrix-vector product can 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 and further discussion.

As a final remark we notice 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 chosen optimality properties of the residuals. 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 of iterative methods for Hermitian matrices $\mathbf{A}$. For non-Hermitian systems, the *Arnoldi* method, see, for instance, [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. Recall that we assume that $\mathbf{b}$ is assumed to have unit length.

If in step j of the Arnoldi method the matrix-vector product is computed 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 an *inexact* Arnoldi method. This latter method satisfies the following perturbed Arnoldi relation

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

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 and construct their iterates as

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

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].

Notice 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 $\widehat{\mathbf{A}}_n \mathbf{x} = \mathbf{b}$. Therefore, these methods, after at most n steps, terminate with $\mathbf{x}_n^{\mathrm{F}} = \mathbf{x}_n^{\mathrm{G}} = (\mathbf{A} + \widetilde{\mathbf{F}}_n \mathbf{V}_n^*)^{-1} \mathbf{b}$ and in the inexact GMRES method, the computed residuals are monotonically decreasing. In the remainder of this section, we will drop the superscripts F or G in expressions that are valid for both methods.

In order to bound the residual gap in step k, we use an expression for the gap that is equivalent to (3.6) 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 $\widehat{\mathbf{A}}_n^{-1}\mathbf{b}$, and thus $y_k^{\mathrm{F}} \approx y_k^{\mathrm{G}}$, it is evident from (7.4) that an appropriate relaxation strategy for inexact GMRES is also suitable for inexact FOM, and vice versa. This will be confirmed by the analysis below.

If we plug (3.7) into (7.5), then we get the following bound for the residual gap of inexact FOM,

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

and for inexact GMRES we get

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

We follow the same approach as for Orthores in Section 6.4 and assume that we terminate the inexact FOM/GMRES method in step k when $\|\mathbf{r}_k\|_2 \leq \varepsilon$, where ε is again in the order of the required residual precision. We see that in step k the residual gap is essentially determined by the tolerances 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 and incorporate this quantity in our choice for η). We, again, see that relaxation is possible with $\eta_j = \varepsilon/\rho_j$. This results for inexact FOM in the bound

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

and for inexact GMRES we get

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

We see that the relaxation strategy derived from the bounds on the residual gap confirms the empirical choice of Bouras et al. in (3.3) for GMRES and can explain the success of this approach. See also the numerical experiments in [2]. Furthermore, we note that the expression for the residual gap of the inexact FOM method and inexact Orthores from the previous section coincide which can be explained by the fact that, for both methods, the matrix-vector products in the exact counterparts are applied to an orthogonal basis. Of course, the behavior of the computed residuals and the values of $\|\underline{T}_k^{\dagger}\|_2$ differ.

7.1. The behavior of the computed residuals. For inexact GMRES we know that the size of the computed residuals monotonically decrease and $\mathbf{r}_n = \mathbf{0}$. 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. The many numerical experiments in [2] suggest that the convergence of the inexact method with the proposed relaxation strategy is comparable to the convergence speed of the exact method. It is, however, very difficult to give a rigorous analysis of this observation. In some cases it can be proven 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 for inexact processes where the perturbation is of the special form:

$$\widetilde{\mathbf{F}}_k = \mathbf{V}_{k+1}\underline{E}_k, \tag{7.10}$$

with $\underline{E}_k$ some upper Hessenberg matrix. In this case we have

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\underline{T}_k - \widetilde{\mathbf{F}}_k = \mathbf{V}_{k+1}\underline{\overline{T}}_k, \quad \text{with} \quad \underline{\overline{T}}_k \equiv \underline{T}_k - \underline{E}_k.$$

This shows that only the Hessenberg matrix $\underline{T}_k$ differs from the Hessenberg matrix of the unperturbed process $\underline{\overline{T}}_k$ and the perturbation does not change the Krylov subspace, or its basis given by $\mathbf{V}_{k+1}$.

To understand the convergence of the inexact process, we compare the norm of the computed residual for the inexact process, with perturbations of the form (7.10), to that of the exact method. We denote the computed residuals of both methods with, respectively, $\mathbf{r}_j$ and $\overline{\mathbf{r}}_j$. For the GMRES method these 'residuals' are given by the following expressions:

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

Since we have that $\underline{\overline{T}}_k = \underline{T}_k - \underline{E}_k$, we can apply standard perturbation theory for the least squares problem. For example, with Theorem 19.1 in [21] we can show that

$$\left|\|\overline{\mathbf{r}_k^{\mathrm{G}}}\|_2 - \|\mathbf{r}_k^{\mathrm{G}}\|_2\right| \le \|\overline{\mathbf{r}_k^{\mathrm{G}}} - \mathbf{r}_k^{\mathrm{G}}\|_2 \le (1 + 2\|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2)\|\underline{E}_k\|_2.$$

This shows that if $\|\overline{\mathbf{r}_k^{\mathrm{G}}}\|_2 = \mathcal{O}(\varepsilon)$ all the η_j should be about ε in order to retain the speed of convergence of the exact method. This simple argument is not sufficient for explaining the fast convergence of the inexact method with the relaxation strategy (3.3) and some more work is necessary. By generalizing Theorem 19.1 in [21], we get the following theorem.

THEOREM 7.1. *Let* $\mathbf{W}_k \equiv \widehat{\mathbf{A}}_k\mathbf{V}_k$ *and let* $\mathbf{P}_k$ *be the skew projection along* $\mathbf{V}_k$ *on* $\mathrm{span}(\mathbf{W}_k)$*:*

$$\mathbf{P}_k = \mathbf{W}_k(\mathbf{V}_k^*\mathbf{W}_k)^{-1}\mathbf{V}_k^* = \mathbf{V}_{k+1}\underline{\overline{T}}_k\overline{T}_k^{-1}\mathbf{V}_k^*.$$

Then, we have for the inexact FOM method

$$\|\overline{\mathbf{r}_k^{\mathrm{F}}} - \mathbf{r}_k^{\mathrm{F}}\|_2 \le \|\mathbf{I} - \mathbf{P}_k\|_2\|\mathbf{A}\|_2\|\underline{T}_k^{\dagger}\|_2 \sum_{j=0}^{k-1} \eta_j(\|\mathbf{r}_j^{\mathrm{G}}\|_2 + \|\mathbf{r}_k^{\mathrm{F}}\|_2).$$

For the inexact GMRES method we have that

$$\|\overline{\mathbf{r}_k^{\mathrm{G}}} - \mathbf{r}_k^{\mathrm{G}}\|_2 \le \|\mathbf{A}\|_2\|\underline{T}_k^{\dagger}\|_2 \sum_{j=0}^{k-1} \eta_j(\|\mathbf{r}_j^{\mathrm{G}}\|_2 + \|\mathbf{r}_k^{\mathrm{G}}\|_2).$$

Proof. We proof the first statement.

$$\begin{aligned}
\|\overline{\mathbf{r}_k^{\mathrm{F}}} - \mathbf{r}_k^{\mathrm{F}}\|_2 &= \|\underline{\overline{T}}_k\overline{T}_k^{-1}e_1 - \underline{T}_kT_k^{-1}e_1\|_2 \\
&= \|[(\underline{\overline{T}}_k - \underline{T}_k) - \underline{\overline{T}}_k\overline{T}_k^{-1}(\overline{T}_k - T_k)]T_k^{-1}e_1\|_2 \\
&= \|(\underline{E}_k - \underline{\overline{T}}_k\overline{T}_k^{-1}\underline{E}_k)T_k^{-1}e_1\|_2 = \|(\mathbf{I} - \mathbf{P}_k)\widetilde{\mathbf{F}}_kT_k^{-1}e_1\|_2 \\
&\le \sum_{j=0}^{k-1} \|(\mathbf{I} - \mathbf{P}_k)\widetilde{\mathbf{f}}_j\|_2\,|e_{j+1}^*T_k^{-1}e_1| \\
&\le \|\mathbf{I} - \mathbf{P}_k\|_2\|\mathbf{A}\|_2\,\|\underline{T}_k^{\dagger}\|_2 \sum_{j=0}^{k-1} \eta_j(\|\mathbf{r}_j^{\mathrm{G}}\|_2 + \|\mathbf{r}_k^{\mathrm{F}}\|_2).
\end{aligned}$$

Where, in the last line we have used Lemma 3.1. This proves the first statement.

For the proof for inexact GMRES, we define $\mathbf{Q}_k$ as the orthogonal projection onto $\mathrm{span}(\mathbf{W}_k)$, then

$$\mathbf{Q}_k = \mathbf{W}_k(\mathbf{W}_k^*\mathbf{W}_k)^{-1}\mathbf{W}_k^* = \mathbf{V}_{k+1}\underline{\overline{T}}_k\underline{\overline{T}}_k^\dagger\mathbf{V}_{k+1}^*.$$

We have that

$$\begin{aligned}
\|\overline{\mathbf{r}_k^{\mathrm{G}}} - \mathbf{r}_k^{\mathrm{G}}\|_2 &= \|\underline{\overline{T}}_k\underline{\overline{T}}_k^\dagger e_1 - \underline{T}_k\underline{T}_k^\dagger e_1\|_2 \\
&= \|\underline{\overline{T}}_k\underline{\overline{T}}_k^\dagger(I - \underline{T}_k\underline{T}_k^\dagger)e_1 - (I - \underline{\overline{T}}_k\underline{\overline{T}}_k^\dagger)\underline{T}_k\underline{T}_k^\dagger e_1\|_2 \\
&\leq \|\underline{\overline{T}}_k\underline{\overline{T}}_k^\dagger(I - \underline{T}_k\underline{T}_k^\dagger)\|_2\,\|(I - \underline{T}_k\underline{T}_k^\dagger)e_1\|_2 + \|(I - \underline{\overline{T}}_k\underline{\overline{T}}_k^\dagger)\underline{E}_k\,\underline{T}_k^\dagger e_1\|_2 \\
&\leq \|(I - \underline{\overline{T}}_k\underline{\overline{T}}_k^\dagger)\underline{E}_k\,\underline{T}_k^\dagger\|_2\,\|\mathbf{r}_k^{\mathrm{G}}\|_2 + \|(I - \underline{\overline{T}}_k\underline{\overline{T}}_k^\dagger)\underline{E}_k\,\underline{T}_k^\dagger e_1\|_2 \\
&= \|(\mathbf{I} - \mathbf{Q}_k)\widetilde{\mathbf{F}}_k\,\underline{T}_k^\dagger\|_2\,\|\mathbf{r}_k^{\mathrm{G}}\|_2 + \|(\mathbf{I} - \mathbf{Q}_k)\widetilde{\mathbf{F}}_k\,\underline{T}_k^\dagger e_1\|_2 \\
&\leq \|\widetilde{\mathbf{F}}_k\|_2\|\underline{T}_k^\dagger\|_2\,\|\mathbf{r}_k^{\mathrm{G}}\|_2 + \sum_{j=0}^{k}\|\widetilde{\mathbf{f}}_j\|_2\,|e_{j+1}^*\underline{T}_k^\dagger e_1| \\
&\leq \|\mathbf{A}\|_2\,\|\underline{T}_k^\dagger\|_2\sum_{j=0}^{k-1}\eta_j(\|\mathbf{r}_j^{\mathrm{G}}\|_2 + \|\mathbf{r}_k^{\mathrm{G}}\|_2).
\end{aligned}$$

Here we used Lemma 3.1 and the identities $(I - \underline{\overline{T}}_k\underline{\overline{T}}_k^\dagger)\underline{T}_k = -(I - \underline{\overline{T}}_k\underline{\overline{T}}_k^\dagger)\underline{E}_k$ and $\|\underline{\overline{T}}_k\underline{\overline{T}}_k^\dagger(I - \underline{T}_k\underline{T}_k^\dagger)\|_2 = \|(I - \underline{\overline{T}}_k\underline{\overline{T}}_k^\dagger)\underline{T}_k\underline{T}_k^\dagger\|_2$, which, for example, can be found in [28]. □

This theorem shows that for special perturbations, the relaxation strategy also preserves the convergence speed of the exact method until the norm of the residuals becomes in the order of the required residual precision. Of course, this does not explain the often good results with relaxed GMRES that is observed, for example, in the experiments in [2]. However, this theorem is difficult to extend to more general perturbations, since the Hessenberg reduction is not forward stable, see [31]. This means that small perturbations in the matrix-vector product can drastically change the resulting Hessenberg matrix. We emphasize that this does not necessarily imply a severe loss of convergence speed for general perturbations but only that the usefulness of the analytical approach taken here is limited. Nevertheless, small perturbations of the matrix-vector product can indeed delay convergence (but they seem to have not a big impact on the stagnation level). We illustrate this by the following experiment with inexact FOM. (Notice that the convergence of the computed residuals of inexact FOM and GMRES are related, see (7.3).)

The matrix $\mathbf{A} \in \mathbb{R}^{100\times 100}$ is lower bidiagonal with diagonal elements $(\mathbf{A})_{j,j} = j$ and has ones on its lower bidiagonal. For the the right-hand side we have taken $\mathbf{b} = e_1$. 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 mainly oriented in the direction of a few eigenvectors of $\mathbf{A}$ and the errors in the matrix-vector product introduce components in directions for which convergence is slow. We mention that convergence of GMRES for this system for general right-hand sides is much slower

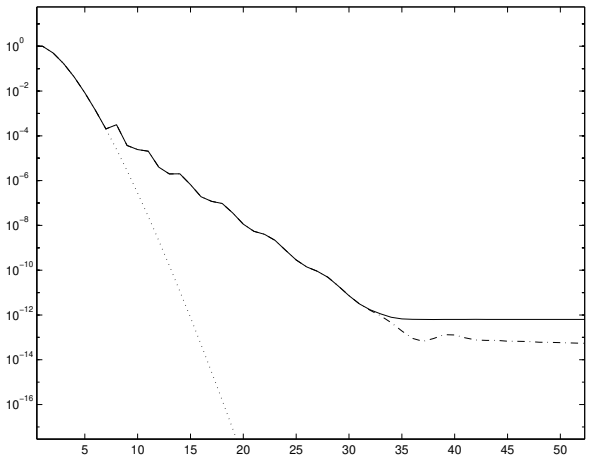

FIG. 7.1: Convergence inexact FOM with $\eta_j = \varepsilon = 10^{-12}$: true residual (—), computed residual (- ·) and $1/j!$ (dotted) as a function of j.

than for the right-hand side taken in this example. We must, however, emphasize that this example is academic since also in finite precision computations the convergence can be much slower than the exact expression for the residuals suggests.

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. 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 6.4 ($\eta_j = \varepsilon/\rho_j$).

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 6.4 (although we lose some additional digits compared to the constant precision case).

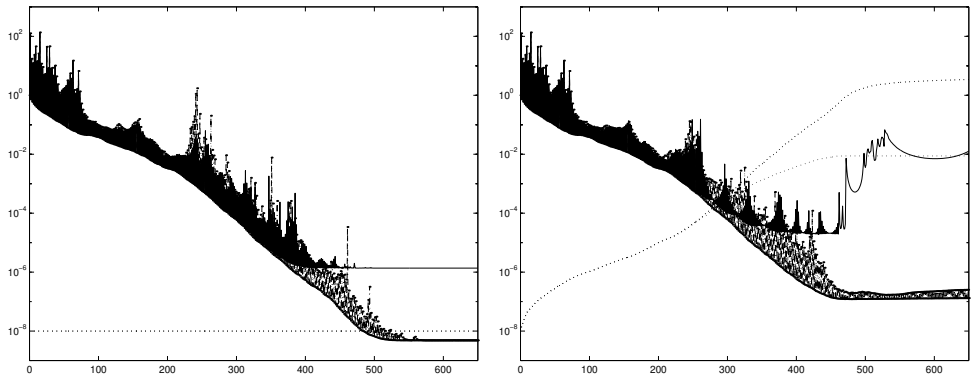

FIG. 8.1: True residuals CG (solid), Orthores (- ·), Rutishauser's variant (dots), η_j (dotted) as a function of j. In both pictures $\varepsilon = 10^{-8}$. Left picture: $\eta_j = \varepsilon$. Right picture: $\eta_j = \varepsilon/\rho_j$.

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. Our results provide insights into the mechanisms behind the successful results with the relaxation strategies 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 in order for a relaxation strategy to be possible. 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. For the GMRES method we have given some results in case the perturbations are of a special form. In future work we plan to further study the effect of inexact matrix-vector products on optimal Krylov subspace methods. And, in particular, the effect of increasing the error during the process.

As a side product of our work, we have shown that using the matrix formulations of the Krylov subspace methods in some cases can simplify the analysis of the residual gap, which is a problem that frequently occurs in analyses of the attainable accuracy of subspace methods. In particular, for three-term recurrences insightful expressions can be easily obtained like for the Chebyshev method and Orthores.

In future work we want to apply the observations in this paper to the simulation of overlap fermions (as mentioned in the beginning of Section 3) and combine this

with the work in [29] for the computation of the matrix sign function acting on a vector. 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.

Postscript. After the submission of this paper the presentation in [25] resulted in the paper [24]. We discuss some differences with this work. The analysis in [25] mainly focuses on the inexact GMRES and inexact FOM method and is based on showing that the true residuals satisfy a *quasi-orthogonality condition* of the form $\|\mathbf{U}_k^*(\mathbf{b}-\mathbf{A}\mathbf{x}_k)\| \leq \mathcal{O}(\varepsilon)$ for some matrix $\mathbf{U}_k$. It is interesting to notice that the quasi-orthogonality is equal to a projection of the residual gap. Therefore, in their presentation, the authors in the end presented a result similar to our Lemma 3.1 to bound this quasi-orthogonality. Paper [24] considers a large number of practical applications. Moreover, the approach taken in the analysis is very different. In the present paper, we are interested in the convergence and stagnation level of the true residuals which are indicators for the quality of the iterates. The basis of our analysis is the splitting into a study of the residual gap, which is connected to the stagnation level, and the convergence and stagnation of the computed residuals. In [24], the authors consider two aspects of inexact Krylov subspace methods: the already discussed quasi-orthogonality of the true residuals and the variational properties of inexact GMRES and inexact FOM method. (This is equivalent to the observation in Section 7 that the computed residuals in inexact GMRES and FOM are residuals of an exact GMRES/FOM process applied to a 'nearby' matrix). There seems to be no discussion in [24] about the direct consequence of quasi-orthogonality and the conserved variational properties of the Krylov subspace method on the stagnation level and convergence speed of the inexact method.

Acknowledgment. The authors are thankful to Valeria Simoncini for providing them with a copy of the slides from [25]. We are thankful to the referees for their constructive comments. Their remarks have helped us to improve the presentation of this paper.

Appendix A. A technical result.

LEMMA A.1. *Let Δ_k be the k by k matrix with zeros entries everywhere except at the positions $(j-1,j)$ and $(j,j-1)$, where it has the value one and the $(2,1)$ element is 2. Then*

$$|e_j^*\Delta_k^{2i}e_j| \leq 2\frac{(2i)!}{(i!)^2} \quad \textit{for all } i,j \geq 1,\ j \leq k.$$

Proof. Let $\mathbb{R}^{\mathbb{N}}$ and $\mathbb{R}^{\mathbb{Z}}$ be the space of vectors with indices in $\mathbb{N}$ and $\mathbb{Z}$, respectively. Consider the map $\widetilde{\Delta}$ on $\mathbb{R}^{\mathbb{Z}}$ given by $\widetilde{\Delta}e_j \equiv e_{j-1}+e_{j+1}$ for all $j \in \mathbb{Z}$. Extend the map Δ_k on $\mathbb{R}^k$ to the map Δ on $\mathbb{R}^{\mathbb{N}}$ given by $\Delta e_j \equiv e_{j-1}+e_{j+1}$ for $j>1$ and $\Delta e_1 \equiv 2e_2$. Note that $0 \leq e_i^*\Delta_k e_j \leq e_i^*\Delta e_j$ for all $i,j \in \mathbb{N}$: here we follow the convention that $\Delta_k e_j = \mathbf{0}$ if $j>k$. Consider the linear map $\mathbf{P}:\mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{N}}$ defined by $\mathbf{P}e_{j+1} = e_{|j|+1}$. One can easily check that $\mathbf{P}\widetilde{\Delta}e_j = \Delta\,\mathbf{P}e_j$ for all $j \in \mathbb{Z}$. Therefore, $\mathbf{P}\widetilde{\Delta} = \Delta\,\mathbf{P}$, and, for $j \geq 0$, we have that

$$\Delta^{2i}e_{j+1} = \Delta^{2i}\mathbf{P}e_{j+1} = \mathbf{P}(\widetilde{\Delta}^{2i}e_{j+1}) = \mathbf{P}\left(\sum_{\ell=0}^{2i}\frac{(2i)!}{\ell!(2i-\ell)!}e_{j-2i+2\ell+1}\right).$$

If $i < j$ then $|j - 2i + 2\ell| + 1 = j + 1$ only if $\ell = i$. Hence, if $i < j$ we find that $e_{j+1}^* \Delta_k^{2i} e_{j+1} \leq e_{j+1}^* \Delta^{2i} e_{j+1} = \frac{2i!}{(i!)^2}$. If $\ell \equiv i - j \geq 0$ then $|j - 2i + 2\ell| + 1 = j + 1$ and $e_{j+1}^* \Delta_k^{2i} e_{j+1} \leq e_{j+1}^* \Delta^{2i} e_{j+1} = \frac{2i!}{(i!)^2} + \frac{2i!}{(i-j)!(i+j)!} \leq 2\frac{2i!}{(i!)^2}$. ∎

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.

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

[5] B. Fischer, *Polynomial Based Iteration Methods for Symmetric Linear Systems*, John Wiley & Sons Ltd., Chichester, 1996.

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

[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.

[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).

[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.

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

[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.

[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*, Linear Algebra Appl., 309 (2000), pp. 289–306.

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

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

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

[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.

[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.

[18] M. H. Gutknecht and S. Röllin, *The Chebyshev iteration revisited*, Parallel Computing, 28 (2002), pp. 263–283.

[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).

[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).

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

[22] C. C. Paige, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Math. Appl., 18 (1976), pp. 341–349.

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

[24] V. Simoncini and D. 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 January 2003.

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

[26] 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.

[27] 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).

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

[29] 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.

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

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

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