The convergence of Krylov methods and Ritz values

Jan Brandts* and Henk van der Vorst †

April 23, 2003

Abstract

In this paper, we review convergence properties of the conjugate gradients method and of other related Krylov subspace methods. We also pay attention to the convergence of the approximations of eigenvalues that are derived from the Krylov subspace, and which are called *Ritz values*. All topics are mutually strongly connected.

1 Introduction

We start with recalling one of the many definitions of the conjugate gradients method for the approximation of the solution x of a linear system $Ax = b$ with symmetric and positive definite system matrix A. An advantage of this definition is that it is easy to state, and that it immediately shows the connection with the Lanczos method for approximation of eigenvalues of A. A disadvantage is that the actual *algorithms* for both the conjugate gradients and the Lanczos method do not follow too easily and require clever combination of several ingredients.

1.1 Basic relations and definitions

Given a non-singular $n \times n$ matrix A and a vector $v \in \mathbb{R}^n$, we denote by $\mathcal{K}^i(A, v)$ the i-th Krylov subspace for A with initial vector v, which is defined for $i \geq 1$ as

$$\mathcal{K}^i(A, v) = \text{span}\{v, Av, A^2v, \ldots, A^{i-1}v\},$$

and which consists of the nullvector for $i = 0$. Let x_0 be a given initial approximation for the solution x of $Ax = b$, and $r_0 = b - Ax_0$ the *residual* corresponding to

*Korteweg-de Vries Institute for Mathematics, Faculty of Science, University of Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, Netherlands. E-mail: brandts@science.uva.nl

†Mathematical Institute, Utrecht University, P.O. Box 80.010, 3508 TA Utrecht, Netherlands. E-mail: vorst@math.uu.nl

this approximation. The conjugate gradients method is an iterative procedure that results in a sequence $(x_i)_{i\geq 1}$ such that

$$x_0 - x_i \in \mathcal{K}^i(A, r_0) \quad \text{and} \quad b - Ax_i = r_i \perp \mathcal{K}^i(A, r_0).$$

Even though these two relations already determines $(x_i)_{i\geq 1}$ uniquely, we will give some more details on how it may be obtained. For ease of presentation, we will assume that all residuals r_i are non-zero, which corresponds to not considering the lucky situation that we find the exact solution.

Since $r_i \in \mathcal{K}^{i+1}(A, r_0)$ but also $r_i \perp \mathcal{K}^i(A, r_0)$, it is clear that $r_0, \ldots, r_{i-1}$ form an orthogonal basis for $\mathcal{K}^i(A, r_0)$. Defining the vectors $v_i = r_i/\|r_i\|$, we get an orthonormal basis. Let for all $i \geq 0$ the $n \times i$ matrix V_i be the defined by $V_i e_{j+1} = v_j$, for $j \in \{0, \ldots, i-1\}$, where e_j is the j-th standard basis vector of $\mathbb{R}^i$. So, the columns of V_i are $v_0, \ldots, v_{i-1}$, and consequently, any element from $\mathcal{K}^i(A, r_0)$ can be written as a linear combination of $v_0, \ldots, v_{i-1}$. In particular,

$$x_0 - x_i = V_i y_i \tag{1}$$

for some $y_i \in \mathbb{R}^i$. As a result,

$$r_i = b - Ax_i = b - Ax_0 + A(x_0 - x_i) = r_0 + AV_i y_i,$$

and since r_i is orthogonal to $\mathcal{K}^i(A, r_0)$, we have that $V_i^* r_i = 0$ and therefore y_i satisfies

$$V_i^* A V_i y_i = -V_i^* r_0 = -e_1 \|r_0\|. \tag{2}$$

Since r_0 is given, it is now clear (at least in theory) how to compute the approximations x_i for $i \geq 1$. First, we set $i = 1$, then we can compute $\|r_0\|$ and $V_1^* A V_1$, solve y_1 from (2), after which $x_1 = x_0 + V_1 y_1$ according to (1). The residual for this approximation can be computed and the process can be continued. As long as the residuals remain non-zero, the matrix V_i will be full rank and the fact that A is positive definite guarantees that y_i can always be solved from (2).

Because A maps $\mathcal{K}^i(A, r_0)$ into $\mathcal{K}^{i+1}(A, r_0)$ for all $i \geq 1$, it follows that Av_i is a linear combination of $v_0, \ldots, v_{i+1}$. This results in the observation that for all $i \geq 1$,

$$AV_i = V_{i+1} T_{i+1,i}, \tag{3}$$

and that the entries t_{ij} of the $(i+1) \times i$ matrix $T_{i+1,i}$ are zero for all $i > j + 1$. As a consequence, the system matrix from (2) can be found as

$$V_i^* A V_i = V_i^* V_{i+1} T_{i+1,i} = I_{i,i+1} T_{i+1,1} =: T_{ii}\,, \tag{4}$$

where $I_{i,i+1}$ represents the trivial embedding $\mathbb{R}^{i+1} \to \mathbb{R}^i$. Since the left-hand side in (4) is symmetric, also T_i must be symmetric. So, $t_{ij} = 0$ if $|i - j| > 1$, or, in other words, T_{ii} (and also $T_{i+1,i}$) is tridiagonal. This fact lies at the basis of the efficiency of the conjugate gradients method. The procedure described below (2) can be reduced to a three-term recursion, and using that T_i is positive definite, even be decoupled into two two-term recursions using Cholesky factorization. We will not go into detail how this is done, but refer to other papers in this volume for the actual algorithm.

The Lanczos method. As a suitable conclusion of this section with basic definitions and relations, we now define approximations of the eigenpairs of A according to the Lanczos method. For this, we associate to each pair $(z, \theta) \in \mathbb{C}^n \times \mathbb{C}$ with $\|z\| = 1$ an *eigenvalue residual* $r = Az - \theta z$. Given an arbitrary, so-called *initial vector* $v \in \mathbb{R}^n$, the Lanczos method aims to find $z \in \mathcal{K}^i(A, v)$ and $\theta \in \mathbb{C}$ such that

$$Az - \theta z = r \perp \mathcal{K}^i(A, v). \tag{5}$$

Let us now suppose that we apply the Lanczos method with the choice $v = r_0$, the initial residual of the conjugate gradients method described above. Since z is supposed to be in $\mathcal{K}^i(A, r_0)$, it can be written as $V_i y$ for some $y \in \mathbb{R}^i$. Then, demanding orthogonality of the residual to $\mathcal{K}^i(A, v)$ results in the quest for pairs $y_j^{(i)}, \theta_j^{(i)}$ such that

$$V_i^*(AV_i y_j^{(i)} - \theta_j^{(i)} V_i y_j^{(i)}) = 0, \quad \text{or equivalently,} \quad T_{ii} y_j^{(i)} = \theta_j^{(i)} y_j^{(i)},$$

which shows that the Lanczos approximations for eigenpairs of A, using r_0 as startvector are the pairs $V_i y_j^{(i)}, \theta_j^{(i)}$ derived from the matrix T_{ii} from the conjugate gradients method. The approximate eigenvalues $\theta_j^{(i)}$ are called *Ritz-values*, and the approximate eigenvectors $V_i y_j^{(i)}$ *Ritz vectors*. We will elaborate on the convergence of Ritz values in Section 3. Before that, we put the Lanczos method and conjugate gradients in the broader context of *subspace methods*.

1.2 CG and Lanczos as instances of more general methods

Suppose $\mathcal{V} \subset \mathbb{R}^n$ and $\dim \mathcal{V} = i < n$, and let $\mathcal{W} = A\mathcal{V}$. Let V be a full rank matrix with column span $\mathcal{V}$ and set $W = AV$. Then both the linear system $Ax = b$ and the eigenvalue problem $Az = \lambda z$ can be approximated in $\mathcal{V}$. We discuss here two general selection principles, by which we mean strategies to determine the i free variables that we have to our disposal. The first one is demanding orthogonality of the residual with respect to $\mathcal{V}$, whereas the second demands orthogonality with respect to $\mathcal{W}$. In both cases, this gives i equations for the i variables.

$\mathcal{V}$-orthogonal methods. The first principle is to find $v \in \mathcal{V}$ such that the residual r, where $r = b - Av$ in case of the linear system, and $r = Av - \theta v$ in case of the eigenvalue problem, is orthogonal to $\mathcal{V}$. As a result, one is interested in solving

$$V^*Wy = V^*b \quad \text{and} \quad V^*Wy = \theta V^*Vy \tag{6}$$

respectively. Indeed, in the first case, Vy can be interpreted as to approximate x, whereas in the second case, (Vy, θ) can be seen as an approximate eigenpair. Note that CG and Lanczos belong to this class of methods, with $\mathcal{V} = \mathcal{K}^i(A, b)$, whereas coarse grid corrections in MultiGrid [2] and the Jacobi-Davidson method [12] are instances of such methods in which $\mathcal{V}$ is *not* a Krylov subspace. The class is generally known as *Rayleigh-Ritz* methods, or $\mathcal{V}$-orthogonal methods.

$\mathcal{W}$-orthogonal methods. A second principle is given in by the following symmetry-argument. If A is invertible, then the eigenvalue problem $Az = \lambda z$ can be formulated equivalently as $A^{-1}z = \lambda^{-1}z$. Since we have subspaces $\mathcal{V}$ and $\mathcal{W}$ at our disposal together with the action of both A and A^{-1} between these spaces, there is no mathematical reason to prefer either formulation of the eigenvalue problem. Therefore, interchanging the role of $\mathcal{V}$ and $\mathcal{W}$, we may as well look for $w \in \mathcal{W}$ and scalars $\mu \in \mathbb{C}$ such that $A^{-1}w - \mu w \perp \mathcal{W}$. Of course, w can be written as Wy for some y, and using that $W = AV$ the orthogonality constraint results in

$$W^*Vy = \mu W^*Wy. \tag{7}$$

In case of the linear system $Ax = b$, the corresponding formulation becomes to find $w \in \mathcal{W}$ such that $w - b \perp \mathcal{W}$, which reduces to

$$W^*Wy = W^*b. \tag{8}$$

In this case, Wy can be seen as an approximation of b, and hence $A^{-1}Wy = Vy$ is an approximation of x. In correspondence to the $\mathcal{V}$-orthogonal methods above, this class of methods is sometimes referred to as $\mathcal{W}$-orthogonal methods, but also as *minimal residual* (in case of linear systems) methods or *harmonic Rayleigh-Ritz* (in case of eigenvalue approximation) methods.

Similarities in appearance. Notice that we have the freedom to choose $b \in \mathcal{V}$. For ease of presentation, we will assume that this is the case. Thus, there exists an $e \in \mathbb{R}^i$ such that $b = Ve$, and using this, a more clear relation between the two different linear system- and eigenvalue approximation methods becomes visible:

$$V^*Wy = V^*Ve, \;\; V^*Wy = V^*V\theta y, \;\; W^*W\mu y = W^*Vy, \;\; W^*Wy = W^*Ve. \tag{9}$$

If A is non-singular and V has full rank, then also W has full rank and W^*W, being a Gramm matrix, is non-singular. Therefore, only the first of the four problems in (9) needs extra assumptions to result in a well-defined approximation, of which symmetry and positive definiteness of A is sufficient.

Krylov subspace methods. Choose $\mathcal{V} = \mathcal{K}^k(A, b)$. Refraining from any extra assumptions, the first method in (9) leads to the full orthogonalization method (FOM) for general matrices A, whereas for symmetric A, we arrive at the Lanczos method for linear systems. Assuming that A is symmetric and positive definite, which is sufficient for a unique solution in $\mathcal{V}$, results in the conjugate gradients (CG) method. If A is merely symmetric, then the second method is the Lanczos method, whereas for general A it leads to the Arnoldi method. The third method in (9), corresponding to the second approach, leads to harmonic Lanczos and harmonic Arnoldi. Finally, the fourth method in (9) leads for general A to the generalized minimal residuals method (GMRES), and for symmetric A to minimal residuals (MinRes). See [5] for references in on Krylov subspace methods. We stress that we have only classified the methods according to their mathematical definition. We did not explain how from these definitions efficient algorithms can be derived. Generally, in case A is symmetric, algorithms can be derived that infer short recursions and low, constant memory costs, whereas for general matrices, the length of the recursions grows linearly with i and the memory costs quadratically in i. Finally, we emphasize that different implementations of the same method may lead to differently behaving algorithms in finite precision arithmetic.

1.3 Minimization properties for linear systems

Notice that for the linear system $Ax = b$, the two methods just described are linear mappings of the exact solution into the approximating space $\mathcal{V}$. Due to the left equation in (6), the corresponding operator, suggestively called $\mathcal{CG}$, can be written as

$$\mathcal{CG} : \mathbb{R}^n \to \mathcal{V} : x \mapsto V(V^*W)^{-1}V^*Ax \tag{10}$$

which can be checked to be idempotent and self-adjoint with respect to the bilinear form defined by $(\cdot,\cdot)_A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} : (x,y)_A \mapsto (Ax, y)$, hence, if A is symmetric positive definite, $\mathcal{CG}(x)$ is the best approximation of x within $\mathcal{V}$ with respect to the associated A-norm. If A is non-symmetric but coercive, the Lax-Milgram lemma gives quasi-optimality, but we will not go into detail here.

Secondly, for arbitrary non-singular A, orthogonalization of the residual to $\mathcal{W}$ as in (8), corresponds to an operator, suggestively called $\mathcal{MR}$, that can be written as

$$\mathcal{MR} : \mathbb{R}^n \to \mathcal{V} : x \mapsto V(W^*W)^{-1}W^*Ax, \tag{11}$$

which can be checked to be idempotent and self-adjoint with respect to the A^*A-inner product defined by $(\cdot,\cdot)_{A^*A} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} : (x,y)_{A^*A} \mapsto (A^*Ax, y)$, hence, $\mathcal{MR}(x)$ is the best approximation of x within $\mathcal{V}$ with respect to the associated A^*A-norm. Since the A^*A norm of the error $x - \mathcal{MR}(x)$ equals the L^2 norm of $A(x - \mathcal{MR}(x)) = b - A\mathcal{MR}(x)$, the name minimal residuals is explained.

A priori bounds. Similar as, for example, in finite element methods, the error of the method can now be bounded by the error of another element in the same subspace $\mathcal{V}$ that is easier to analyze. In finite element methods, one usually studies the error of some locally defined nodal interpolant. Also in Krylov methods, an interpolation problem can be formulated and solved in order to obtain upper bounds on the error. We will show how to do this in Section 2.

1.4 Circumventing the lack of symmetry

Before concluding the introduction, we would like to comment on several ways to avoid the long recursions and higher memory requirements that come with non-symmetric matrices A. The most obvious solution for linear systems is to consider the normal equations $A^*Ax = A^*b$ and to use a solver for symmetric positive definite system matrices. Unfortunately, the convergence of conjugate gradients, as we will see further on, depends heavily on the condition number of the system matrix, and the normal equations have a squared condition number, compared to the original ones. Another way to symmetrize the matrix would be to consider the system

$$\begin{pmatrix} M & A^* \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c \\ b \end{pmatrix}, \tag{12}$$

for some symmetric M. This corresponds to solving both the systems $Mx + A^*y = c$ and $Ax = b$ simultaneously. An immediate consequence is that something is computed that may have no use at all. The obvious choice $M = 0$ together with the use of conjugate gradients leads to the bi-conjugate gradients (BiCG) algorithm, which allow short recursions and a low use of memory. Since for all M, and hence in particular for $M = 0$ the resulting matrix is not positive definite, BiCG may break down.

An reasonable alternative seems to apply the approach from (8) to the larger symmetric system (12). This would result in short recurrences, and would minimize the residual over $\mathcal{V}$. Notice that for given $(v^*, w^*)^* \in \mathbb{R}^{2n}$, the squared residual norm equals

$$\left\| \begin{pmatrix} c \\ b \end{pmatrix} - \begin{pmatrix} M & A^* \\ A & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix} \right\|^2 = \|b - Av\|^2 + \|c - Mv - A^*w\|^2. \tag{13}$$

Not even for the choice $M = 0$, much can be concluded about the convergence behaviour of the single component $\|b - Av\|$. But although monotonicity cannot be guaranteed, clearly, any bound for the minimal residual method can be applied to the residual in (13). We will discuss this method briefly in Section 4.2.

A strategy motivated by mixed finite element methods. Another interesting choice in (12) is $M = I$ together with $c = 0$. Again, the resulting system is symmetric indefinite. Choosing subspaces $\mathcal{V} \subset \mathbb{R}^n$ and $\mathcal{W} \subset \mathbb{R}^n$, we can aim to solve $(\widehat{x}, \widehat{y}) \in \mathcal{V} \times \mathcal{W}$ from the following system

$$\begin{aligned} (\widehat{x}, v) + (\widehat{y}, Av) &= 0 \quad \text{for all } v \in \mathcal{V} \\ (A\widehat{x}, w) &= (b, w) \quad \text{for all } w \in \mathcal{W}. \end{aligned} \tag{14}$$

It is well-known from the theory of mixed finite element methods [3] for which choices there exists a unique solution of this discretized problem. For example, given $\mathcal{V}$, choosing $\mathcal{W} = A\mathcal{V}$ is a possibility. Note that we are solving the "second order" equation $AA^*y = b$ by splitting it up into two "first order" systems $x = A^*y$ and $Ax = b$, and "testing" the equations in $\mathcal{V}$ and $\mathcal{W}$ respectively, using "integration by parts" in the first. Instead of (12), and thinking back to the normal equations, we may also formulate $A^*Ax = A^*b$ in this way, by discretizing

$$\begin{pmatrix} I & A \\ A^* & 0 \end{pmatrix} \begin{pmatrix} y \\ x \end{pmatrix} = -\begin{pmatrix} 0 \\ A^*b \end{pmatrix}. \tag{15}$$

Here too, the link with mixed finite element methods is obvious, although at present, we know of no results in this direction.

Outline of the rest of this paper. In Section 2, we will concentrate on deriving a priori bounds for the conjugate gradients method, as a prototype of Krylov subspace methods in general. In Section 3 we examine the convergence of Ritz values in the Lanczos method. Both sections are strongly related. Finally, in Section 4 we reflect on the application of the techniques developed in the previous sections to more general Krylov subspace methods.

2 The convergence of Conjugate Gradients

The conjugate gradient method constructs in the i^{th} iteration step an approximation x_i of x such that $x_0 - x_i \in \mathcal{K}^i(A, r_0)$. Since each element from $\mathcal{K}^i(A, r_0)$ can be written as $Q_{i-1}(A)r_0$, where Q_{i-1} is a polynomial of degree $i-1$, we have

$$x_0 - x_i = Q_{i-1}(A)r_0 = Q_{i-1}(A)(Ax - Ax_0) = AQ_{i-1}(A)(x - x_0). \tag{16}$$

Adding $x - x_0$ to both sides, we see that

$$x_i - x = P_i(A)(x_0 - x), \tag{17}$$

for some polynomial P_i of degree i with $P_i(0) = 1$. By our observations in Section 1.3, P_i is such that $\|x_i - x\|_A$ is minimal over all polynomials of degree i for which $P_i(0) = 1$. If we denote the eigenvalues and the orthonormalized eigenvectors of A by λ_j, z_j, where $0 < \lambda_1 \leq \cdots \leq \lambda_n$, and write $r_0 = \sum_j \gamma_j z_j$, then we find that

$$r_i = b - Ax_i = A(x - x_i) = P_i(A)r_0 = \sum_j \gamma_j P_i(\lambda_j) z_j, \tag{18}$$

and hence,

$$\|x_i - x\|_A^2 = \sum_j \frac{\gamma_j^2}{\lambda_j} P_i^2(\lambda_j). \tag{19}$$

Notice that only those j for which $\gamma_j \neq 0$ give a positive contribution to this sum. In particular, if A would have been semidefinite with $\lambda_j = 0$ for all $j \leq J \leq n$ for some $J \leq n$, then the minimization problem is still well-defined as long as the corresponding coefficients $\gamma_j, j \leq J$ in the expansion of the initial residual are also zero. The situation in which these γ_j are small, for example due to rounding errors, is discussed in [6].

2.1 A priori bounds for the error

A priori bounds for the error in the A-norm are obtained by observing that the right-hand side of (19) can be bounded as follows. Using (19) itself with $i = 0$, we find that for each polynomial R_i of degree i with $R_i(0) = 1$,

$$\sum_j \frac{\gamma_j^2}{\lambda_j} P_i^2(\lambda_j) \leq \sum_j \frac{\gamma_j^2}{\lambda_j} R_i^2(\lambda_j) \leq \max_j R_i^2(\lambda_j) \sum_j \frac{\gamma_j^2}{\lambda_j} = \max_j R_i^2(\lambda_j) \|x_0 - x\|_A^2, \tag{20}$$

where the maximum, in principle, only needs to be taken over those j for which $\gamma_j \neq 0$. Since generally we do not know the location of the eigenvalues of A, but may have some information on its extremal eigenvalues, and hence on its condition number $\kappa \equiv \lambda_n/\lambda_1$, we will continue our analysis with the simple bound

$$\max_j R_i^2(\lambda_j) \leq \max_{x \in [\lambda_1, \lambda_n]} R_i^2(x). \tag{21}$$

A well-known and descriptive upper bound now results when we choose for R_i the i^{th} degree *Chebyshev polynomial* C_i transformed to the interval $[\lambda_1, \lambda_n]$ and scaled such that its value at 0 is equal to 1, or in other words,

$$R_i(x) = C_i\left(\frac{2x - (\lambda_1 + \lambda_n)}{\lambda_n - \lambda_1}\right) / C_i\left(\frac{-(\lambda_1 + \lambda_n)}{\lambda_n - \lambda_1}\right),$$

Uniformly, $R_i(x)$ is the best approximation of the zero function on $[\lambda_1, \lambda_n]$ under the restriction that it has value one at zero. The resulting upper bound for the A-norm of the error is (see [4, 5, 1]):

$$\|x_i - x\|_A \leq 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^i \|x_0 - x\|_A, \quad \text{for all } i \geq 1, \tag{22}$$

which shows that we have fast convergence for small condition numbers.

Upper bounds for special situations. Other, though similar bounds can be obtained by choosing tailor-made polynomials when extra information is available on the eigenvalue distribution. For instance Axelsson [1] considered the specific case in which λ_n is much larger than λ_{n-1}. Then the polynomial

$$R_i(x) = (1 - \frac{x}{\lambda_n})C_{i-1}\left(\frac{2x - (\lambda_1 + \lambda_{n-1})}{\lambda_{n-1} - \lambda_1}\right) / C_{i-1}\left(\frac{-(\lambda_1 + \lambda_{n-1})}{\lambda_{n-1} - \lambda_1}\right)$$

satisfies $R_i(0) = 1, R(\lambda_n) = 0$, and if λ_n is relatively much larger than λ_{n-1}, then on $[\lambda_1, \lambda_{n-1}]$, it is almost equal to the polynomial

$$C_{i-1}\left(\frac{2x - (\lambda_1 + \lambda_{n-1})}{\lambda_{n-1} - \lambda_1}\right) / C_{i-1}\left(\frac{-(\lambda_1 + \lambda_{n-1})}{\lambda_{n-1} - \lambda_1}\right)$$

This results in an improved upper bound, that approximately lags one step behind the upper bound (22) for a process with a condition number λ_{n-1}/λ_1.

These types of upper bounds show that it is important to have small condition numbers, or, in the case of larger condition numbers, to have eigenvalue distributions for which the eigenvalues that cause these large condition numbers are well separated. In that case we say that the (remaining) part of the spectrum is (relatively) clustered. The purpose of preconditioning is to reduce the condition number κ and/or clustering of the eigenvalues. In [1] the situation is analysed where the eigenvalues of $K^{-1}A$ are in disjoint a priori known intervals.

2.2 Local effects in the convergence behavior

Upper bounds as in (22) show that we have global convergence, but they do not help us to explain all sorts of local effects in the convergence behavior of CG. A very well-known effect is the so-called superlinear convergence: in many situations one observes that the average speed of convergence seems to increase as the iteration proceeds. As we have seen in Section 1.1, the conjugate gradients algorithm and the Lanczos algorithm are closely related. The eigenvalues of the implicitly generated tridiagonal matrix T_{ii} are the Ritz values of A with respect to the current Krylov subspace. It is known from Lanczos theory that these Ritz values converge to the eigenvalues of A, and that in general the extreme eigenvalues of A are well approximated already during early iterations [7, 9, 10]. Furthermore, the speed of convergence depends on how well these eigenvalues are separated from the others (gap ratio) [10]. This will help us to understand the superlinear convergence behavior of the conjugate gradient method (as well as other Krylov subspace methods). It can be shown that as soon as one of the extreme eigenvalues is modestly well approximated by a Ritz value, the procedure converges from then on as a process in which this eigenvalue is absent, i.e., as a process with a reduced condition number. Note that superlinear convergence behavior in this connection is used to indicate linear convergence with a factor that is gradually decreasing during the process, as more and more of the extreme eigenvalues are sufficiently well approximated.

Example. We consider a linear system of order $n = 100$ with one isolated eigenvalue 0.005 and the other 99 eigenvalues equally distributed over the interval $[0.08, 1.21]$. The right-hand side was chosen such that the initial residual has equal components in all eigenvector directions. The A-norms of the error $x - x_i$ are plotted on a logarithmic scale in Figure 1. We observe in this figure, that the convergence for the first 8 or so iteration steps is relatively slow, in agreement with the condition number of A, which is 242.0. After iteration 10 the convergence is noticeably faster: about 4 times faster than in the earlier phase. This corresponds to a condition number of about 15.1 and that is the condition number for a matrix with eigenvalue spectrum $[0.08, 1.21]$, that is, the smallest eigenvalue 0.005 seems to have lost its influence on the convergence behaviour after 10 iteration steps. The smallest Ritz value at iteration step 10 is $\theta_1^{(10)} \approx 0.0061$, at iteration step 11: $\theta_1^{(11)} \approx 0.0054$, and at step 12: $\theta_1^{(12)} \approx 0.0052$. Clearly, it is converging to $\lambda_1 = 0.005$, but we see that the faster convergence of CG already sets in when the smallest Ritz value is still a relatively crude approximation (at step 10 we have an error of about 20%). We will now analyze this behaviour in more detail.

2.3 Derivation of superlinear bounds

The local convergence behavior of CG, and especially the occurrence of superlinear convergence, was first explained in a qualitative sense in [4], and later in a quantitative sense in [13]. In both papers it was linked to the convergence of eigenvalues

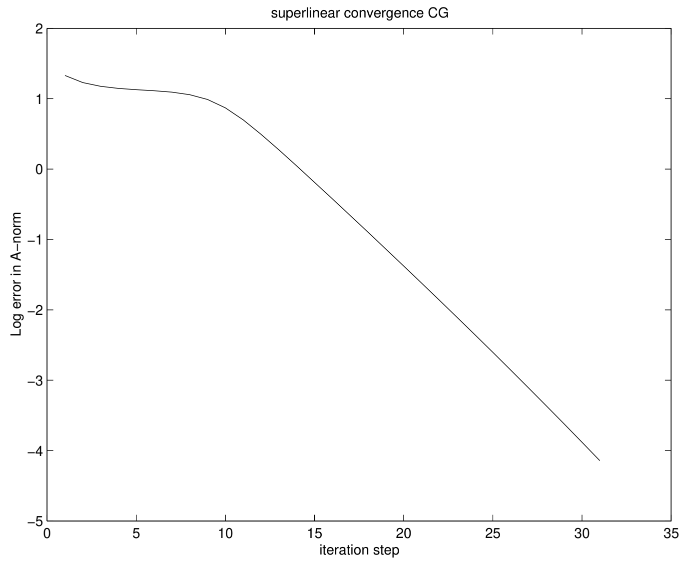

Figure 1: The convergence of CG for one isolated small eigenvalue

(Ritz values) of T_{ii} in (4) towards eigenvalues of A, for increasing i. We will follow here the exposition given in [13]. The main idea in that paper is to link the convergence of CG in a certain phase with a process in which one or more of the Ritz values are replaced by eigenvalues of the matrix. To that end we first have to make a link with the Ritz values.

As we have seen in (17), the residual r_i can be expressed as

$$r_i = P_i(A) r_0, \tag{23}$$

with $r_i \perp \mathcal{K}^i(A, r_0)$. It can be shown [13] that P_i is the characteristic polynomial of T_{ii} from (4), and hence,

$$\det(T_{ii}) P_i(t) = (\theta_1^{(i)} - t)(\theta_2^{(i)} - t) \cdots (\theta_i^{(i)} - t). \tag{24}$$

We will exploit this characterization to study the convergence of CG from a certain iteration i onwards. Writing the initial error $x - x_0$ as

$$x - x_0 = \sum_1^n \mu_j z_j, \tag{25}$$

on account of (17) we have that

$$x - x_i = \sum_1^n \mu_j P_i(\lambda_j) z_j.$$

Now, define $\bar{x}_0 = x_i + \mu_1 P_i(\lambda_1) z_1$, and use $\bar{x}_0$ as the starting vector of an auxilliary CG process (with the same A and b), of which the initial error obviously satisfies

$$x - \bar{x}_0 = \sum_2^n \mu_j P_i(\lambda_j) z_j.$$

This new process generates a sequence $(\bar{x}_k)_{k\geq 0}$, for which the errors are characterized by polynomials $\bar{P}_k$ of degree k as follows,

$$x - \bar{x}_k = \bar{P}_k(A)(x - \bar{x}_0) = \sum_2^n \mu_j \bar{P}_k(\lambda_j) P_i(\lambda_j) z_j. \tag{26}$$

For given k, the product of $\bar{P}_k$ and the polynomial q_i defined by

$$q_i(t) = \frac{(\lambda_1 - t)(\theta_2^{(i)} - t)\cdots(\theta_i^{(i)} - t)}{\lambda_1 \theta_2^{(i)} \cdots \theta_i^{(i)}} = \frac{\theta_1^{(i)}(\lambda_1 - t)}{\lambda_1(\theta_1^{(i)} - t)} P_i(t) \tag{27}$$

is of degree $k+i$ and has value one at $t=0$. Therefore, it may be used to bound the error in the A-norm of the $(k+i)$-th iterate of the original CG process, hence

$$\|x - x_{i+k}\|_A^2 \leq \|\bar{P}_k(A) q_i(A)(x - x_0)\|_A^2 = \sum_2^n \lambda_j \bar{P}_k(\lambda_j)^2 q_i(\lambda_j)^2 \mu_j^2. \tag{28}$$

Consider the rational factor in (27) representing the quotient q_i/P_i. Assuming that $\theta_1^{(i)}$ does not equal any of the eigenvalues $\lambda_2, \ldots, \lambda_n$, the following maximum is well-defined,

$$F_i \equiv \frac{\theta_1^{(i)}}{\lambda_1} \max_{j\geq 2} \left| \frac{\lambda_j - \lambda_1}{\lambda_j - \theta_1^{(i)}} \right|,$$

with as a consequence that $|q_i(\lambda_j)| \leq F_i |P_i(\lambda_j)|$ for all $j \geq 2$. We can now further simplify the upper bound in (28) as

$$\begin{aligned} \|x - x_{i+k}\|_A^2 &\leq F_i^2 \sum_2^n \lambda_j \bar{P}_k(\lambda_j)^2 P_i(\lambda_j)^2 \mu_j^2 \\ &= F_i^2 \|x - \bar{x}_k\|_A^2 \\ &\leq F_i^2 \frac{\|x - \bar{x}_k\|_A^2}{\|x - \bar{x}_0\|_A^2} \|x - x_i\|_A^2. \end{aligned} \tag{29}$$

We have now proved a result stated in Theorem 3.1 from [13]. Its interpretation is as follows. First note that when θ_1^i is close to λ_1 for some i then $F_i \approx 1$. This is illustrated in Figure 2, where we have sketched the rational factor from (27). In fact, if $\lambda_1 \leq \theta_1^i < \lambda_2$, then F_i equals the product of two factors with a simple interpretation: θ_1^i/λ_1, which shows how good the approximation θ_1^i is, and $\frac{\lambda_2-\lambda_1}{\lambda_2-\theta_1^i}$, which measures the relative position of θ_1^i in the interval $[\lambda_1, \lambda_2]$. We conclude that

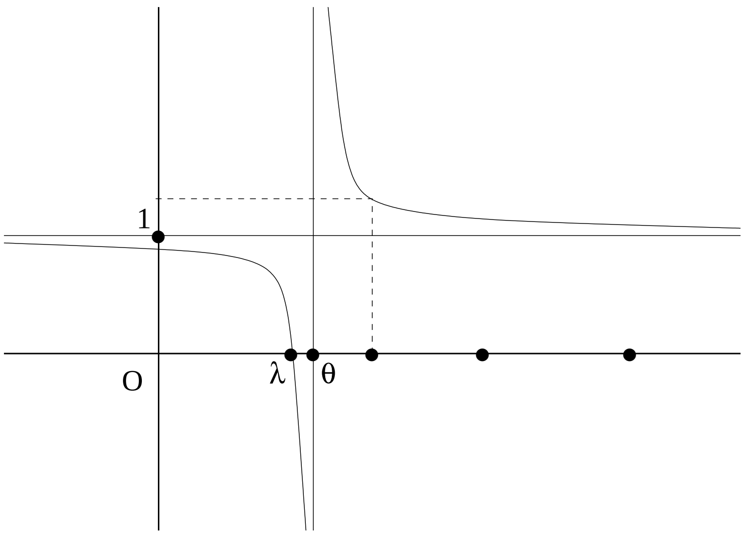

Figure 2: Qualitative sketch of $t \mapsto \frac{\theta(\lambda-t)}{\lambda(\theta-t)}$.

as soon as λ_1 is sufficiently well approximated, we may expect a reduction in the next steps that, apart from a small multiplicative factor, is bounded by the reduction that we obtain for a CG process for $Ax = b$ in which λ_1 is missing. If we define the effective condition number for that process as $\kappa_2 = \lambda_n/\lambda_2$, we have that

$$\|x - x_{i+k}\|_A \leq 2F_i \left(\frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1}\right)^k \|x - x_i\|_A.$$

Hence, if $\theta_1^{(i)} \approx \lambda_1$, then after the i-th step we have a reduction that is bounded by an expression in which we see the condition number for the remaining part of the spectrum. In fact, (29) reveils that the approximation of λ_1 by the smallest Ritz value does not need to be extremely accurate for the superlinear convergence to become apparent. For instance, if

$$\frac{\theta_1^{(i)} - \lambda_1}{\lambda_1} < 0.1 \quad \text{and} \quad \frac{\theta_1^{(i)} - \lambda_1}{\lambda_2 - \lambda_1} < 0.1,$$

then we already have that $F_i < 1.25$, which shows that we have from then on a reduction as for the reduced process except for a very modest factor. Things cannot be worse from then on, because it can be proved (see also Section 3) that $\theta_1^{(i)}$ is strictly decreasing towards λ_1 for increasing i. This means that also F_i becomes smaller in further iterations.

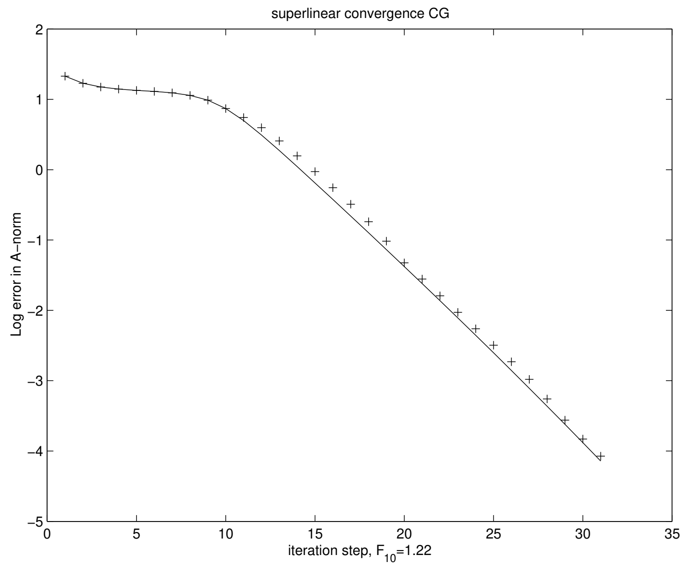

Figure 3: CG (–) and the comparison process(+++)

Example (continued). We now return to our earlier example of Figure 1. At iteration step 10 we have a value $F_{10} \approx 1.22$. In Figure 3 we have plotted the errors in A-norm for the CG process (the drawn line), and we have also marked with + the results for the comparison process. In this comparison process, we have removed at iteration step 10 the error component in the eigenvector direction corresponding to $\lambda_1 = 0.005$, and from then on we have multiplied the error-norms for the modified process by F_{10}. We see that this comparison process, which has only eigenvalues in the interval $[0.08, 1.21]$, describes the observed convergence behaviour for the original CG process (with all eigenvalues present) rather well.

In [13] other results are presented that shed more light on all sorts of special situations, such as almost double eigenvalues in the spectrum of A, or situations in which more Ritz values have converged to a number of successive eigenvalues. Not surprisingly, these results show that after (modest) approximations of Ritz values for the smallest m eigenvalues, we will see convergence of the CG process from then on, governed by a condition number $\kappa_m \equiv \lambda_n/\lambda_{m+1}$. Of course, similar results can be derived for convergence of Ritz values towards the largest eigenvalues, but usually removal of small eigenvalues leads to much smaller condition numbers.

3 The convergence of Ritz values

We have introduced the Lanczos method as an algorithm based on finding pairs $(z, \theta) \in \mathcal{K}^i(A, v) \times \mathbb{C}$ such that the corresponding residual $r = Az - z\theta$ is orthogonal to the Krylov subspace $\mathcal{K}^i(A, v)$, which was spanned by the columns of the matrix V_i. The defining *Rayleigh-Ritz* orthogonality relation

$$V_i^*(Az - \theta z) = 0, \tag{30}$$

led, since each $z \in \mathcal{K}^i(A, v)$ can be written as $z = V_i y$ for some $y \in \mathbb{R}^i$, to the smaller eigenvalue problem

$$T_{ii} y = \theta y, \tag{31}$$

for the tridiagonal matrix T_{ii} from (4). The eigenpairs of T_{ii} are usually referred to as *Ritz values* and *Ritz vectors*, and are considered as approximations of eigenpairs of A. Strictly speaking, the first basisvector v of $\mathcal{K}^i(A, v)$ may be chosen arbitrarily and need not be an initial residual of an explicit conjugate gradient process, but note that it can always be interpreted as one, simply for solving $Ay = v$ with initial approximation $y_0 = 0$. This is what we will do in the sequel in order to analyze the convergence of the Ritz values.

3.1 Monotonicity through the Minimax characterization

Notice that for given non-zero $z \in \mathcal{K}^i(A, r_0)$ we may write the *Rayleigh quotient* $\mathcal{R}(A, z)$ for z as

$$\mathcal{R}(A, z) = \frac{z^* A z}{z^* z} = \frac{y^* V_i^* A V_i y}{y^* V_i^* V_i y} = \frac{y^* T_{ii} y}{y^* y} = \mathcal{R}(T_{ii}, y), \tag{32}$$

where $y \in \mathbb{R}^i$ is the vector for which $z = V_i y$. This gives some insight into how the Ritz values are related to the eigenvalues of A by means of their *minimax characterization*. Indeed, according to the Minimax Theorem, see, e.g.,[10], the eigenvalues λ_j of A are equal to

$$\lambda_j = \min_{S \subset \mathbb{R}^n, \dim(S) = j} \left(\max_{0 \neq z \in S} \mathcal{R}(A, z) \right), \tag{33}$$

and the maximum is taken at the corresponding eigenvector(s) z_j of A. Similarly, the eigenvalues of T_{ii} (the Ritz values) are equal to

$$\theta_j^{(i)} = \min_{U \subset \mathbb{R}^i, \dim(U)=j} \left(\max_{0 \neq y \in U} \mathcal{R}(T_{ii}, y) \right), \tag{34}$$

and the maximum is taken at the corresponding eigenvector(s) $y_j^{(i)}$ of T_{ii}. Combining (32) with (33) and (34), leads to the interesting observation that the ordered Ritz values $\theta_1^{(i)} \leq \theta_2^{(i)} \leq \cdots \leq \theta_i^{(i)}$ are monotonically decreasing to the ordered eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ of A. To be more explicit, with $j \leq i$ we have that

$$\lambda_j = \min_{S \subset \mathbb{R}^n, \dim(S)=j} \left(\max_{0 \neq z \in S} \mathcal{R}(A, z) \right) \leq \min_{S \subset \mathcal{K}^i, \dim(S)=j} \left(\max_{0 \neq z \in S} \mathcal{R}(A, z) \right) = \theta_j^{(i)}, \tag{35}$$

where the latter equality results from the application of the change in coordinates (32) to (34). This leads, for increasing i, to

$$\lambda_1 \leq \theta_1^{(1)}, \quad \lambda_1 \leq \theta_1^{(2)} \text{ and } \lambda_2 \leq \theta_2^{(2)}, \quad \lambda_1 \leq \theta_1^{(3)},\ \lambda_2 \leq \theta_2^{(3)}, \text{ and } \lambda_3 \leq \theta_3^{(3)}, \quad \ldots$$

This shows that the $\theta_j^{(i)}$ with j fixed and $i \geq j$ running, are bounded from below by λ_j, whereas similar observations can be made for the other extremal end of the spectrum by considering the maximin characterization (we will not go into detail). The right-hand side of (35) also shows that for fixed j we have

$$\lambda_j = \theta_j^{(n)} \leq \theta_j^{(n-1)} \leq \cdots \leq \theta_j^{(j+1)} \leq \theta_j^{(j)}. \tag{36}$$

Since, in exact arithmetic, the Lanczos process is a finite process, we can strictly speaking, not regard the sequence of smallest Ritz values as a converging process. However, in finite precision, the process will usually deliver accurate approximations that cannot be essentially improved, long before the finiteness of the process would play a role. In this sense, it behaves as a converging process that is terminated when finite precision effects inhibit further convergence. We will therefore, in the context of Ritz values and Ritz vectors speak of convergence, and we will consider an eigenpair as (practically) converged when its residual is in norm smaller than some given tolerance.

3.2 The polynomial characterization

We have seen that the basis vectors v_i for the Krylov subpace are scalar multiples of the conjugate gradient residuals r_i. This means that they can be expressed in terms of the initial vector $v_0 = r_0/\|r_0\|$ as

$$v_i = c_i P_i(A) v_0,$$

where P_i is the polynomial from (17),(23) and (24) and $c_i = \|r_0\|/\|r_i\|$ to assure that $\|v_i\|_2 = 1$. This polynomial characterization permits nice error bounds. In order to show this, we define $R_{i,1}$ as the unique polynomial for which $P_i(t) = (t - \theta_1^{(i)})R_{i,1}$ and write $P_i(A)v_0$ accordingly as

$$P_i(A)v_0 = \frac{1}{\|r_0\|}P_i(A)r_0 = \frac{1}{\|r_0\|}\sum_j(\lambda_j - \theta_1^{(i)})R_{i,1}(\lambda_j)\gamma_j z_j,$$

where we have written $r_0 = \sum_j \gamma_j z_j$ as before. Because $P_i(A)v_0$ is orthogonal to all lower degree polynomials in A applied to v_0, we have that

$$0 = (P_i(A)v_0)^T R_{i,1}(A)v_0 = \sum_j(\lambda_j - \theta_1^{(i)})R_{i,1}^2(\lambda_j)\gamma_j^2.$$

This gives the following relation for $\theta_1^{(i)}$:

$$\theta_1^{(i)} = \frac{\sum_j \lambda_j R_{i,1}^2(\lambda_j)\gamma_j^2}{\sum_j R_{i,1}^2(\lambda_j)\gamma_j^2} = \frac{z^*Az}{z^*z} \quad \text{where} \quad z = R_{i,1}(A)v_0. \tag{37}$$

By combining this with our knowledge that the Ritz values are the (local) extrema of the Rayleigh quotient, we conclude that this gives us two nice results:

1. The Ritz vector corresponding to $\theta_1^{(i)}$ is (apart from a scalar factor) given by $z = R_{i,1}(A)v_0$.
2. Because of the minimax property, the above characterization gives us an excellent means to find *a priori* upper bounds for $\theta_1^{(i)}$, namely, any polynomial of degree $i-1$ may replace $R_{i,k}$ and this will give an upper bound.

3.3 Construction of a related CG-process

From now on we will assume that $\gamma_1 \neq 0$, which is no loss of generality. Note that expression (37) is scaling invariant with respect to the polynomial $R_{i,1}$, and we scale it such that $\widetilde{R}_{i,1}(\lambda_1) = 1$, that is we replace $R_{i,1}$ by

$$\widetilde{R}_{i,1}(\lambda) = R_{i,1}(\lambda)/R_{i,1}(\lambda_1).$$

Then we obtain the following expression for the distance of the smallest Ritz value $\theta_1^{(i)}$ to the smallest eigenvalue λ_1 (with $\widetilde{\gamma_j} = \gamma_j/\gamma_1$):

$$0 \le \theta_1^{(i)} - \lambda_1 = \frac{\sum_j(\lambda_j - \lambda_1)R_{i,1}^2(\lambda_j)\gamma_j^2}{\sum_j R_{i,1}^2(\lambda_j)\gamma_j^2} = \frac{\sum_{j>1}(\lambda_j - \lambda_1)\widetilde{R}_{i,1}^2(\lambda_j)\widetilde{\gamma}_j^2}{1 + \sum_{j>1}\widetilde{R}_{i,1}^2(\lambda_j)\widetilde{\gamma}_j^2}. \tag{38}$$

Since the right-hand side of (38) is, minus a fixed number λ_1, a Rayleigh quotient of which we know from (35) that it takes its minimum (over the Krylov subspace) at precisely the Ritz vector belonging to the smallest Ritz value, we know that $\widetilde{R}_{i,1}$ is the polynomial that minimizes the right-hand side of (38) over all polynomials of degree $i-1$ having value one at λ_1.

As we will show now, the error in the Ritz value can be interpreted as the square of the norm of the error of some associated conjugate gradient process for a system $By = d$, where d remains to be specified, and with starting vector $y_0 = 0$. For this process, in analogy to (17), we have a relation of the form

$$y - y_i = (I - BQ_{i-1}(B))(y - y_0) \equiv G_i(B)(y - y_0) = G_i(B)y, \tag{39}$$

where G_i is such, that

$$\|y - y_i\|_B = \|G_i(B)y\|_B$$

is minimal over all polynomials of degree i that have value one at zero. Hence, comparing this with (38), we see that the numerator in the right-hand of (38) can be interpreted as the error, in the B norm, of a CG-process for an $(n-1)\times(n-1)$ matrix B with eigenvalues $\lambda_1 - \lambda_j$ $(j > 1)$ and with eigenvectors w_j that are unknown, but this is irrelevant since they do not appear explicitly in the expressions. The solution y can be written as

$$y = \sum_{j\geq 2} \widetilde{\gamma}_j w_j.$$

With Q_{i-1} the polynomial from (39) that minimizes the B-norm of this CG-process in step $i-1$, this yields an upper bound for the approximation error in the smallest Ritz value simply by neglecting the denominator, which is larger than one. In fact, we then have that

$$\theta_1^{(i)} - \lambda_1 \leq \sum_{j>1} (\lambda_j - \lambda_1) Q_{i-1}^2(\lambda_j)\widetilde{\gamma}_j^2 = \|y_{i-1} - y\|_B^2. \tag{40}$$

This shows that $\theta_1^{(i)} - \lambda_1$ is bounded from above by the square of the B-norm of the error in a related conjugate gradient process. As soon as $\theta_1^{(i)}$ has converged to some modest degree, then (38) implies that

$$1 + \sum_{j>1} \widetilde{R}_{i,1}^2(\lambda_j)\widetilde{\gamma}_j^2 \approx 1,$$

and then the approximation error almost coincides with the CG-error. Or said in a slightly different way, the same kind (only marginally better) of convergence behavior that we observe for conjugate gradients (with a reduced and shifted spectrum) can be observed for the convergence behavior of the smallest Ritz value.

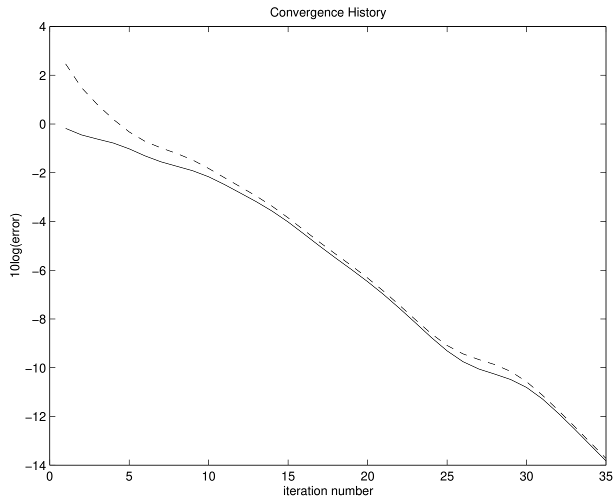

Figure 4: Ritz error (-line-) and CG-error (-dashed-)

Example. We consider the diagonal matrix of order 400 with eigenvalues $\lambda_1 = -0.034$, $\lambda_2 = -0.081$, $\lambda_3 = -0.0811$, $\lambda_4 = -0.127$, $\lambda_5 = -0.155$, and $\lambda_6 = -0.190$. The other 394 eigenvalues of this matrix are uniformly distributed in $[-1.2, -0.2]$. We select a v_0 that has equal components in all eigenvector directions: $\gamma_j = 1$. In Fig. 4 we see the values of the error $\theta_1^{(i)} - \lambda_1$, and the (squares of the) errors in B-norm of the associated CG-process. We see that the Ritz errors are nicely below the CG-errors, and ultimately they almost coincide.

We also see that temporary slow-downs in the convergence behaviour of CG is reflected by a similar slowdown of the Ritz value, although not at the same iteration numbers and at the same level. We would like to stress that the associated CG process is not the CG process that defines the same Ritz values as the Lanczos process that we have used.

In Section 2.3 we have argued that the conjugate gradient process converges superlinearly, or more precisely, that conjugate gradients converges as a process from which an extremal eigenvalue has been removed, as soon as this eigenvalue is modestly well approximated by the Ritz value of the corresponding Lanczos process (that is indeed the Lanczos process defined by the iteration coefficients of CG).

In a similar way, we have superlinear convergence for the smallest Ritz value, as soon as the largest eigenvalue of the shifted reduced problem is modestly well approximated by a Ritz value. Since the Ritz values are shift invariant, this implies

that we may expect faster convergence in average, for the smallest Ritz value, as soon as the next Ritz value gets modestly close to the next eigenvalue.
Proceeding in this way, we see that we may expect an increase of the average speed of convergence, each time when the next smallest eigenvalue gets a Ritz value in its neighbourhood. This explains nicely what we see in practice.
This interpretation also shows that the error can be bounded as an error in a CG-process, and for this upper bounds have been derived that decay with a linear factor that depends on the condition number of B, or, in our case $(\lambda_n - \lambda_1)/(\lambda_2 - \lambda_1)$.

The similarity between CG errors and Ritz errors had already been noticed and analyzed in [11, Section 6]. There it has been shown that the CG-error for a process with matrix A, characterized by the eigenvalues λ_j and weights γ_j, can be interpreted as a Ritz value for a related process that is characterized by eigenvalues $\bar{\lambda}_j$, augmented with $\bar{\lambda}_0 = 0$, and a set of weights $\bar{\gamma}_j$. They give explicit formulas for these eigenvalues and weights, and also lower bounds for the smallest Ritz value of this scheme. The approach in [11] is essentially the same as ours, but gives slightly more detailed results (at the cost of a more complicated analysis). We see that this correspondence between Ritz value convergence and CG-convergence allows us to transfer all the insights that result from analysis that has been carried out for convergence aspects of Ritz values (that is, for instance, misconvergence or stagnation, see [14]) to CG-behavior.

4 The convergence of Minimal Residuals

We introduced the minimal residual method in Section 1.2 as a specific $\mathcal{W}$-orthogonal method, i.e., the one in which $\mathcal{V} = \mathcal{K}^i(A, r_0)$. In Section 1.3 we explained its name by showing that $\mathcal{W}$-orthogonality coincides with minimizing the A^*A-norm of the error, or equivalently, the L^2-norm of the residual. Here we will briefly discuss its convergence behavior, following an exposition by Wathen et al.[15].

4.1 Bounds for symmetric indefinite problems

Considering the problem $Ax = b$ with A symmetric indefinite. The minimal residual method constructs in the i-th iteration step an approximation x_i of x such that $x_0 - x_i \in \mathcal{K}^i(A, r_0)$, and $r_i = b - Ax_i$ has minimal Euclidean norm. As in (16), we have $x_i = x_0 + Q_{i-1}(A)r_0$ for some polynomial Q_{i-1} of degree $i-1$ and hence,

$$r_i = b - Ax_i = b - Ax_0 - AQ_{i-1}(A)r_0 = (I - AQ_{i-1}(A))r_0 = P_i(A)r_0, \qquad (41)$$

where $P_i(A)$ is, in this case, the polynomial of degree i with $P_i(0) = 1$ such that $\|P_i(A)r_0\|$ is minimal over all such polynomials. Following the same approach as in

Section 2, and writing $r_0 = \sum_j \gamma_j z_j$, we get

$$\|r_i\|^2 = \sum_{j=1}^{n} \gamma_j^2 P_i(\lambda_j)^2 \leq \max_j |P_i(\lambda_j)|^2 \|r_0\|^2. \tag{42}$$

The location of the eigenvalues is unknown, but it may be able to show, in specific situations, that they lie in a set of the form

$$E = [-a, -b] \cup [c, d]$$

with a, b, c, d positive real numbers. In that case, it will be interested in finding a polynomial R_i of degree i with $R_i(0) = 1$ which is uniformly small on E. Similar as in conjugate gradients, this would lead to upper bounds of the error, though this time in the A^2-norm instead of the A-norm (which, for indefinite A, is not a norm anymore).

Comparison with CG applied to $A^2x = Ab$. It is reasonable to aim for convergence bounds of the minimal residual method that are better than the bounds one would obtain by solving the squared system $A^2x = Ab$ by the conjugate gradients method. For this reason, Wathen et al. [15] remark that a bound based on *symmetric* Chebychev polynomials (see [8]) for the special case that the two intervals containing the eigenvalues have equal length, i.e., $a - b = d - c$, gives no advantage compared to conjugate gradients applied to the squared system. At the same time, a construction is presented in [15] showing that whenever the intervals have *different* lengths, minimal residuals is to be preferred. This construction is however not within the scope of this paper.

4.2 Non-symmetric systems.

With the above observations in mind, we will briefly return to the methods proposed in Section 1.4 to circumvent the lack of symmetry. In case A is not symmetric, we noticed that it may be worthwhile to apply the minimal residual method to the double-sized symmetric indefinite linear system

$$\begin{pmatrix} M & A^* \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c \\ b \end{pmatrix}. \tag{43}$$

Denote the double-sized matrix by $\mathcal{C}(M)$, then it is easily seen that $\mathcal{C}(M)$ and $\mathcal{C}(0)$ are congruent,

$$\mathcal{C}(M) = \begin{pmatrix} I & \frac{1}{2}MA^{-1} \\ 0 & I \end{pmatrix} \mathcal{C}(0) \begin{pmatrix} I & 0 \\ \frac{1}{2}A^{-*}M & I \end{pmatrix}, \tag{44}$$

and hence, by Sylvester's law, the inertia of the double-sized matrix is independent of the choice of M. Now, consider $M = \alpha I$ for some real number α, possibly zero, then

$$\mathcal{C}(\alpha I) \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix} \quad \Leftrightarrow \quad \mathcal{C}(\alpha I) \begin{pmatrix} -y \\ x \end{pmatrix} = (\alpha - \lambda) \begin{pmatrix} -y \\ x \end{pmatrix}, \tag{45}$$

which can be easily checked by writing out the equations. Therefore, λ is an eigenvalue of $\mathcal{C}(\alpha I)$ if and only if $\alpha - \lambda$ is. Hence, the eigenvalues of $\mathcal{C}(\alpha I)$ are symmetrically distributed around $\frac{1}{2}\alpha$. Since this is the case also for $\alpha = 0$, and taking into account the constant inertia in α, this means that there cannot be eigenvalues between zero and α. Or, in other words, λ and $\alpha - \lambda$ must be on different sides of the origin.

With respect to the convergence of minimal residuals, the eigenvalues of $\mathcal{C}(0)$ are distributed as badly as possible, i.e., symmetrically around the origin. According to [15], application of minimal residuals to solve the system would, in fact, be equally expensive as applying conjugate gradients to the squared system with system matrix $\mathcal{C}(0)^2$. This squared system involves the solution of both the systems $A^*Ax = A^*b$ and $AA^*y = Ac$. This, in turn, is more expensive than solving only $A^*Ax = A^*b$ by conjugate gradients.

For nonzero choices of α, the two intervals containing the eigenvalues are also of equal length, although it is not clear a priori of which length. It may be that there exists an optimal value for it, but this issue needs further research.

Acknowledgments

Jan Brandts acknowledges the support of a Research Fellowship (1999-2004) of the *Royal Netherlands Academy of Arts and Sciences*, whereas both authors thank Michal Křížek of the Academy of Science of the Czech Republic and the anonymous referee for commenting on earlier versions of this paper.

References

[1] O. Axelsson. Solution of linear systems of equations: Iterative methods. In V. A. Barker, editor, *Sparse Matrix Techniques*, pages 1–51, Berlin, 1977. Copenhagen 1976, Springer Verlag.

[2] D. Braess. *Finite elements. Theory, fast solvers, and applications in solid mechanics.* Cambridge University Press, Second Edition, 2001.

[3] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *RAIRO Anal. Numér.*, 8:129–151, 1974.

[4] P. Concus, G. H. Golub, and D. P. O'Leary. A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations. In J. R. Bunch and D. J. Rose, editors, *Sparse Matrix Computations*. Academic Press, New York, 1976.

[5] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 1996.

[6] E. F. Kaasschieter. A practical termination criterion for the Conjugate Gradient method. *BIT*, 28:308–322, 1988.

[7] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Natl. Bur. Standard*, 45:225–280, 1950.

[8] V.I. Lebedev. An iteration method for the solution of operator equations with their spectrum lying on several intervals. *USSR Comput. Math. and Math. Phys.*, 9:17-24, 1969.

[9] C. C. Paige. Computational variants of the Lanczos method for the eigenproblem. *J. Inst. Math. Appl.*, 10:373–381, 1972.

[10] B. N. Parlett. *The Symmetric Eigenvalue Problem*. Prentice-Hall, Englewood Cliffs, N.J., 1980.

[11] G.L.G. Sleijpen and A. (van der Sluis). Further results on the convergence of conjugate gradients and Ritz values. *Linear Algebra Appl.*, 246:233-278, 1996.

[12] G.L.G. Sleijpen and H. A. (van der Vorst). A Jacobi Davidson method for the eigenvalue problem. *SIAM J. Sci. Comp.*, xx:yy-zz,1996.

[13] A. van der Sluis and H. A. van der Vorst. The rate of convergence of conjugate gradients. *Numer. Math.*, 48:543–560, 1986.

[14] A. van der Sluis and H. A. van der Vorst. The convergence behavior of Ritz values in the presence of close eigenvalues. *Lin. Algebra and Appl.*, 88/89:651–694, 1987.

[15] A.J. Wathen, B. Fischer, and D.J. Silvester. The convergence of iterative solution methods for symmetric and indefinite linear systems. In 'Numerical Analysis

1997', D.F. Griffiths and G.A. Watson, editors, Pitman Research Notes in Mathematics Series, Addison Wesley Longman, Harlow, England, pp. 230-243, 1997.