The Convergence of Krylov Subspace Methods for Large Unsymmetric Linear Systems*

Zhongxiao Jia†

Abstract

The convergence problem of Krylov subspace methods, e.g. FOM, GMRES, GCR and many others, for solving large unsymmetric linear systems has been intensively investigated. There are many results in the literature, mainly for the case where the coefficient matrix A is diagonalizable and its spectrum lies in the open right (left) half plane. In this paper, we focus on a convergence analysis of those Krylov subspace methods which give rise to a similar minimax problem in the case where the coefficient matrix A is defective. When the spectrum of A either lies in the open right (left) half plane or is on the real axis, we establish the related theoretical error bounds and reveal some intrinsic relationships between the convergence speed and the spectrum of A. The results show that these methods are likely to converge slowly whenever one of three cases occurs: A is defective, the distribution of its spectrum is not favorable, or the Jordan basis of A is ill-conditioned.

Keywords. unsymmetric linear systems, Krylov subspace, the Chebyshev polynomials, defective, derivatives.

AMS(MOS) subject classifications. 65F10, 49K35

1 Introduction

In recent years, considerable efforts have been devoted to solving large unsymmetric linear system

$$Ax = b, \tag{1}$$

*This work was supported by the Graduiertenkolleg at the University of Bielefeld

†Fakultät für Mathematik, Universität Bielefeld, email: jia@mathematik.uni-bielefeld.de, Postfach 10 03 01, 33501 Bielefeld, Germany

where $A \in C^{N \times N}$ is large, unsymmetric (non-Hermitian) and nonsingular, and b an $N \times 1$ column vector. A major class of methods for solving (1) is Krylov subspace type or conjugate gradient type methods, and so far many researchers have made contributions to them [2, 9, 8, 6, 11, 22, 23, 24, 25, 31, 32]; see the survey papers [26, 12]. There are some very successful schemes among them, e.g. GCR or ORTHOMIN [8, 9, 31], FOM or Arnoldi's method [22], GMRES [25] and QMR [11], to name only a few.

We can observe that as far as the convergence analysis is concerned, many of the Krylov subspace methods including those named above and some others, e.g. the method presented in [6], all give rise to a similar minimax problem. The only difference is that some of the methods require that the spectrum of the coefficient matrix A is restricted to the open right (left) half plane, e.g. GCR, while some of them do not, e.g. FOM and GMRES. Until now, several authors, e.g. Manteuffel [18, 19], Eisenstat et al. [8], Saad et al. [22, 25] and Freund et al. [11], have derived a number of results, mainly when A is diagonalizable and its spectrum lies in the open right (left) half plane.

This paper focuses on the case where A is defective and its spectrum either lies in the open right (left) half plane or is on the real axis, and studies the associated convergence problem of these methods. We establish the related theoretical error bounds and reveal some intrinsic relationships between the convergence speed and the spectrum of A. Here I want to point out that Trefethen [28, 20] has introduced some concepts of *ϵ-pseudo-spectra* to deal with non-normal matrices, which are very useful in making a convergence analysis. Indeed, in the presence of round errors, an original defective matrix will become a diagonalizable one with simple eigenvalues. However, the eigenvector matrix will be very ill-conditioned in that case. Numerical experiments have shown that one could not see the difference in convergence behaviour for matrices with simple eigenvalues, but ill-conditioned eigenvectors for some almost multiple eigenvalues, and matrices with those almost multiple eigenvalues replaced by Jordan blocks of appropriate size, e.g. [30]. Therefore, from both mathematical and numerical points of view, it is worthwhile to consider the Jordan block case in a convergence analysis, and our analysis may help to understand phenomena that are observed in actual floating-point computation.

In Section 2 we introduce the notation used; in Section 3, we briefly review FOM and GMRES from a theoretical point of view; taking FOM and GMRES as examples, we devote Section 4 to a convergence analysis of these methods.

2 Notation

This part is mainly taken from [13, 19].

Assume that $A \in C^{N \times N}, N \gg 1$, is unsymmetric (non-Hermitian). Let Q_m denote the set of all polynomials of degree not exceeding m and the Krylov subspace $\mathcal{K}_m(r^{(0)}, A) = \text{span}\{r^{(0)}, Ar^{(0)}, \ldots, A^{m-1}r^{(0)}\}$, and let π_m be the orthogonal projector on $\mathcal{K}_m(r^{(0)}, A)$. Let $\|\cdot\|$ denote the 2-norm, the superscript "H" the conjugate transpose, $\kappa(S) = \|S\|\|S^{-1}\|$ and R the set of real numbers.

According to the above assumptions, there exists a nonsingular matrix S such that $A = S^{-1}JS$, where $J = \mathrm{diag}(J_1, J_2, \ldots, J_M)$ with

$$J_i = \begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda_i \end{pmatrix}$$

being the $d_i \times d_i$ Jordan block. In this case, for any $p \in Q_m$ we have

$$p(A) = p(S^{-1}JS) = S^{-1}p(J)S = S^{-1}\mathrm{diag}(p(J_1), p(J_2), \ldots, p(J_M))S$$

with

$$p(J_i) = \begin{pmatrix} p(\lambda_i) & p'(\lambda_i) & & \cdots & \frac{p^{(d_i-1)}(\lambda_i)}{(d_i-1)!} \\ & p(\lambda_i) & \ddots & & \vdots \\ & & \ddots & \ddots & \\ & & & \ddots & p'(\lambda_i) \\ & & & & p(\lambda_i) \end{pmatrix}.$$

3 FOM and GMRES

In order to give the convergence problem and the notation to be used in the convergence analysis, in this section we briefly review two typical and well known Krylov subspace methods, FOM and GMRES.

Let x^* be the solution of (1), $x^{(0)}$ an initial guess, and $r^{(0)} = b - Ax^{(0)}$ the initial residual, and let $\dim(\mathcal{K}_m(r^{(0)}, A)) = m$. Then setting $x = x^{(0)} + z$, we have the following linear system equivalent to (1):

$$Az = r^{(0)}. \tag{2}$$

The Full Orthogonalization Method (FOM) seeks an approximate solution of the form $x_F^{(m)} = x^{(0)} + z_F^{(m)}$ satisfying the orthogonal projection

$$\begin{cases} z_F^{(m)} & \in & \mathcal{K}_m(r^{(0)}, A), \\ r_F^{(m)} & \perp & \mathcal{K}_m(r^{(0)}, A), \end{cases} \tag{3}$$

where $r_F^{(m)} = b - Ax_F^{(m)} = r^{(0)} - Az_F^{(m)}$, while GMRES consists in finding an approximate solution of the form $x_G^{(m)} = x^{(0)} + z_G^{(m)}$ satisfying the orthogonal projection

$$\begin{cases} z_G^{(m)} & \in & \mathcal{K}_m(r^{(0)}, A), \\ r_G^{(m)} & \perp & A\mathcal{K}_m(r^{(0)}, A), \end{cases} \tag{4}$$

where $r_G^{(m)} = b - Ax_G^{(m)} = r^{(0)} - Az_G^{(m)}$, or equivalently finding $z_G^{(m)} \in \mathcal{K}_m(r^{(0)}, A)$ such that

$$\|r_G^{(m)}\| = \|r^{(0)} - Az_G^{(m)}\| = \min_{z \in \mathcal{K}_m(r^{(0)},A)} \|r^{(0)} - Az\|. \tag{5}$$

Let $A_m = \pi_m A \pi_m$ and note that $\pi_m r^{(0)} = r^{(0)}$. Then (3) is equivalent to solving the following linear system in $\mathcal{K}_m(r^{(0)}, A)$:

$$A_m z = r^{(0)}. \tag{6}$$

Let $z^* = A^{-1} r^{(0)}$ be the exact solution of $Az = r^{(0)}$. Then we have $x^* = x^{(0)} + z^*$. In terms of the distance $\|(I - \pi_m)z^*\|$ between z^* and $\mathcal{K}_m(r^{(0)}, A)$, Saad [22, 23] has established the following results.

Proposition 1 *Let $\gamma_m = \|\pi_m A(I - \pi_m)\|$. Then the residual of z^* for (6) satisfies*

$$\|r^{(0)} - A_m z^*\| \le \gamma_m \|(I - \pi_m)z^*\|. \tag{7}$$

Corollary 1 *Let γ_m be defined as above and $\kappa_m = \|(A_m|_{\mathcal{K}_m(r^{(0)},A)})^{-1}\|$. Then*

$$\|z^* - z_F^{(m)}\| \le \sqrt{1 + \gamma_m^2 \kappa_m^2}\,\|(I - \pi_m)z^*\|, \tag{8}$$

where $A_m|_{\mathcal{K}_m(r^{(0)},A)}$ is the restriction of A_m to $\mathcal{K}_m(r^{(0)}, A)$.

The corollary shows that if (6) is not ill-conditioned, the error $\|z^* - z_F^{(m)}\|$ will be of the same order as $\|(I - \pi_m)z^*\|$.

Brown [4] has studied the intimate connections between FOM and GMRES. Here we only point out that in the context of general unsymmetric matrices, unlike symmetric positive matrices, we cannot guarantee that $A_m|_{\mathcal{K}_m(r^{(0)},A)}$ is invertible for any A. In what follows, we always assume that $(A_m|_{\mathcal{K}_m(r^{(0)},A)})^{-1}$ exists for all $m \le N$.

We can present some algorithms based on Krylov subspace methods described above. For more details, refer to [22, 25].

4 Convergence Analysis

Taking FOM and GMRES as examples, in this section we consider the difficult problem of analyzing how these methods converge as m increases. Saad [22, 25] has given the following

Proposition 2 *The distance $\|(I - \pi_m)z^*\|$ between z^* and $\mathcal{K}_m(r^{(0)}, A)$ and $\|r_G^{(m)}\|$ satisfy*

$$\|(I - \pi_m)z^*\| = \min_{p \in Q_m, p(0)=1} \|p(A)z^*\|, \tag{9}$$

$$\|r_G^{(m)}\| = \min_{p \in Q_m, p(0)=1} \|p(A)r^{(0)}\|, \tag{10}$$

respectively.

In terms of Proposition 2, it is easy to prove that these methods give rise to the following similar minimax problem.

Theorem 1 *With the notation in Section 2, we have*

$$\begin{aligned} \|(I-\pi_m)z^*\| &\le \kappa(S)\epsilon^{(m)}\|z^*\|, & (11)\\ \|r_G^{(m)}\| &\le \kappa(S)\epsilon^{(m)}\|r^{(0)}\|, & (12)\end{aligned}$$

where

$$\epsilon^{(m)} = \min_{p\in Q_m, p(0)=1} \max_{i=1,2,\ldots,M} \|p(J_i)\|. \qquad (13)$$

Proof. It is enough for us to look at $\min_{p\in Q_m,p(0)=1}\|p(A)z^*\|$. By Proposition 2, we have

$$\begin{aligned} \|(I-\pi_m)z^*\| &\le \min_{p\in Q_m,p(0)=1} \|p(A)\|\|z^*\| \\ &= \min_{p\in Q_m,p(0)=1} \|S^{-1}p(J)S\|\|z^*\| \\ &\le \kappa(S) \min_{p\in Q_m,p(0)=1} \|p(J)\|\|z^*\| \\ &= \kappa(S) \min_{p\in Q_m,p(0)=1} \max_{i=1,2,\ldots,M} \|p(J_i)\|\|z^*\|. \end{aligned}$$

Thus, the assertions hold. □

Therefore, the convergence problem of these methods amounts to estimating $\epsilon^{(m)}$, provided that the Jordan basis of A is not too ill-conditioned.

Remark.

- If $\kappa(S)$ is very large, namely, the Jordan basis of A is very ill-conditioned, these methods may converge very slowly even though $\epsilon^{(m)}$ tends to zero quite rapidly. We can thus expect that all their restarted versions may not converge at all unless the steps m per restarting are taken sufficiently large.

- When A is diagonalizable, there are many upper bounds for $\epsilon^{(m)}$, e.g. [14, 17, 1, 18, 19, 32, 22, 7, 3].

- For large unsymmetric eigenproblems, a much more complicated analog of $\epsilon^{(m)}$ has been established in [16].

From Theorem 1, we now want to find a polynomial $p \in Q_m$ such that it satisfies the condition $p(0) = 1$, and its values of function and the jth order derivatives, $j < d_i, i = 1, 2, \ldots, M$, are as small as possible on the spectrum of A. To this end, we need a lemma.

Lemma 1 *Let $T_m(z)$ be the first kind Chebyshev polynomial of degree m in the complex plane [21]. Then*

1. *If $z \in [-1,1]$ and $0 \leq j \leq m$, then*

$$| T_m^{(j)}(z) | \leq | T_m^{(j)}(\pm 1) | = \frac{m^2(m^2-1)\cdots(m^2-(j-1)^2)}{1\cdot 3\cdot 5\cdots(2j-1)} = m^{2j}C(m,j). \quad (14)$$

Here for $j \geq 1$, $C(m,j) = \frac{(1-\frac{1}{m^2})(1-\frac{2^2}{m^2})\cdots(1-\frac{(j-1)^2}{m^2})}{1\cdot 3\cdot 5\cdots(2j-1)}$ is decreasing in j, and $C(m,0) = 1$.

2. *If $z \neq \pm 1$, $0 \leq j \leq m$ and for $Re(z) \geq 0$ $B_m(z,j)$ is defined by*

$$T_m^{(j)}(z) = m^j \frac{(z+\sqrt{z^2-1})^m}{(z^2-1)^j} B_m(z,j), \quad (15)$$

then $B_m(z,j)$ is uniformly bounded for m, and is at most of order $O(z^j)$ if $z \notin [-1,1]$.

If $z > 1$ is fixed and $0 \leq j \leq m$, $T_m^{(j)}(z)$ is increasing in j.

3. *Assume $E(0,1,a)$ to be an ellipse with the center at the origin, the foci distance 1 and the main semiaxis a. Then*

$$\begin{aligned} m^{2j}C(m,j) &< \max_{z\in E(0,1,a)} | T_m^{(j)}(z) | \\ &= m^j \frac{(a+\sqrt{a^2-1})^m}{(a^2-1)^j} B_m(a,j). \end{aligned} \quad (16)$$

Proof. Part 1. See Rivlin [21, p. 33].

Part 2. According to one of the definitions of $T_m(z)$, for $Re(z) \geq 0$ we can write $T_m(z)$ in the form

$$\begin{aligned} T_m(z) &= \frac{1}{2}[(z+\sqrt{z^2-1})^m + (z+\sqrt{z^2-1})^{-m}] \quad (17) \\ &= \frac{1}{2}(z+\sqrt{z^2-1})^m Q_m(z), \quad (18) \end{aligned}$$

where $Q_m(z) = 1 + (z+\sqrt{z^2-1})^{-2m}$ (If $Re(z) < 0$, $T_m(z) = \frac{1}{2}(z-\sqrt{z^2-1})^m Q_m(z)$, where $Q_m(z) = 1 + (z-\sqrt{z^2-1})^{-2m}$). Then we have

$$\begin{aligned} \lim_{m\to\infty} Q_m(z) &= 1, \text{ for } z \notin [-1,1], \\ | Q_m(z) | &\leq 2, \text{ for } z \in [-1,1]. \end{aligned}$$

It is clear that both $(z+\sqrt{z^2-1})^m$ and $Q_m(z)$ are analytic in the complex plane excluding the points ± 1 though $T_m(z)$ is analytic in the whole complex plane.

Now from (18), we get

$$\begin{aligned} T_m'(z) &= \frac{1}{2}m\frac{(z+\sqrt{z^2-1})^m}{\sqrt{z^2-1}}Q_m(z)+\frac{1}{2}(z+\sqrt{z^2-1})^m Q_m'(z) \\ &= m\frac{(z+\sqrt{z^2-1})^m}{z^2-1}(\frac{1}{2}\sqrt{z^2-1}Q_m(z)+\frac{1}{2m}(z^2-1)Q_m'(z)) \\ &= m\frac{(z+\sqrt{z^2-1})^m}{(z^2-1)}B_m(z,1). \end{aligned}$$

Obviously, from the above analysis, for $z \notin [-1,1]$ we can see that $B_m(z,0)=\frac{1}{2}Q_m(z)$ contains the factor $(z+\sqrt{z^2-1})^{-2m}$ and is of order $O(1)$. It follows that $B_m(z,1)$ also contains this factor. Therefore, $B_m(z,1)$ is uniformly bounded for all m, and $B_m(z,1)=O(z)O(1)+O(z^2)O(z^{-1})=O(z)$. So the assertion holds for $j=1$.

By induction, suppose for $k=j$ that the assertion is valid, and $B_m(z,j)$ contains the factor $(z+\sqrt{z^2-1})^{-2m}$. Then

$$\begin{aligned} T_m^{(j+1)}(z) &= m^j\frac{m(z+\sqrt{z^2-1})^m(\sqrt{z^2-1})^{2j-1}-2j(z+\sqrt{z^2-1})^m z(z^2-1)^{j-1}}{(z^2-1)^{2j}}B_m(z,j) \\ &+m^j\frac{(z+\sqrt{z^2-1})^m}{(z^2-1)^j}B_m'(z,j) \\ &= m^{j+1}\frac{(z+\sqrt{z^2-1})^m}{(z^2-1)^{j+1}}(\sqrt{z^2-1}B_m(z,j)-\frac{2j}{m}zB_m(z,j)+\frac{1}{m}(z^2-1)B_m'(z,j)) \\ &= m^{j+1}\frac{(z+\sqrt{z^2-1})^m}{(z^2-1)^{j+1}}B_m(z,j+1). \end{aligned}$$

Since $B_m(z,j)$ contains the factor $(z+\sqrt{z^2-1})^{-2m}$, it is easily seen that $B_m(z,j+1)$ is uniformly bounded for all m, and for $z \notin [-1,1]$

$$B_m(z,j+1)=O(z)(z^j)+O(z)O(z^j)+O(z^2)O(z^{j-1})=O(z^{j+1}).$$

In terms of a result of [21, p. 51]

$$T_m^{(j)}(z)=\sum_{l=0}^{m-j}A_{lj}T_l(z), 1\le j\le m, \tag{19}$$

where $A_{lj}\ge 0, 0\le j\le m$, we obtain

$$T_m^{(j+1)}(z)-T_m^{(j)}(z)=\sum_{l=0}^{m-j}A_{lj}(T_l'(z)-T_l(z)), 1\le j+1\le m.$$

It is known that the left side of last relation is nonnegative since it is easy to show that $T_l'(z)-T_l(z)>0$ for $z>1$, so the assertion is proved.

Part 3. Assume ∂E to be the boundary of $E(0,1,a)$. Then, by the maximum modulus principle and (19), we can get

$$\begin{aligned}\max_{z\in E(0,1,a)} \mid T_m^{(j)}(z) \mid &= \max_{z\in\partial E} \mid T_m^{(j)}(z) \mid \\ &= \max_{z\in\partial E} \mid \sum_{l=0}^{m-j} A_{lj} T_l(z) \mid .\end{aligned}$$

Since $A_{lj} \geq 0$ and $T_l(z)$, $0 \leq l \leq m-j$, achieves the maximum at the point a, it follows immediately that

$$\begin{aligned}\max_{z\in E(0,1,a)} \mid T_m^{(j)}(z) \mid &= T_m^{(j)}(a) \\ &= m^j \frac{(a+\sqrt{a^2-1})^m}{(a^2-1)^j} B_m(a,j).\end{aligned}$$

On the other hand, since the points ± 1 are the interior points of $E(0,1,a)$, we have

$$\begin{aligned}m^{2j} C(m,j) &= \mid T_m^{(j)}(\pm 1) \mid < \max_{z\in E(0,1,a)} \mid T_m^{(j)}(z) \mid \\ &= m^j \frac{(a+\sqrt{a^2-1})^m}{(a^2-1)^j} B_m(a,j).\end{aligned}$$

Thus, the assertion (16) holds. □

Remark.

- By a continuity argument and comparing Part 2 with Part 1 of Lemma 1, for z belonging to a neighborhood of the point 1 or -1, and $0 \leq j \leq m$, we then have

$$\frac{B_m(z,j)}{(z^2-1)^j} = O(m^j) \text{ for } m \to \infty. \tag{20}$$

This result corrects an error in [18], and Lemma 1 also appears in [16].

4.1 The case of matrices with real spectrum

Assume that the eigenvalues of A are ordered as follows:

$$0 <\mid \lambda_1 \mid \leq \mid \lambda_2 \mid \leq \cdots \leq \mid \lambda_M \mid . \tag{21}$$

First we consider the case where the spectrum of A is real positive.

Theorem 2 *Assume the spectrum of A to be real positive and the ordering rule as (21), and let*

$$d = \max_{i=1,2,\ldots,M}\{d_i\}, \gamma = \frac{\lambda_M+\lambda_1}{\lambda_M-\lambda_1}, \beta = \frac{2}{\lambda_M-\lambda_1}.$$

Then
$$\epsilon^{(m)} \le \begin{cases} dm^{2(d-1)}/T_m(\gamma), & \text{if } \beta \le 1, \\ d(\beta m)^{2(d-1)}/T_m(\gamma), & \text{if } \beta > 1, \end{cases} \tag{22}$$
where $T_m(x)$ is the first kind Chebyshev polynomial of degree m.

Proof. Write

$$J_i(x) = \begin{pmatrix} x & 1 & & & \\ & x & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & x \end{pmatrix}_{d_i \times d_i}, \quad i = 1, 2, \ldots, M,$$

$$I_A = [\lambda_1, \lambda_M].$$

Then by Theorem 1, we have

$$\epsilon^{(m)} \le \min_{p \in Q_m, p(0)=1} \max_{i=1,2,\ldots,M, x \in I_A} \|p(J_i(x))\|. \tag{23}$$

From approximation theory [5], we know that the optimal polynomial $p_o \in Q_m$ satisfying (23) exists. Furthermore it is well known [21] that for the minimax problem

$$\min_{p \in Q_m, p(0)=1} \max_{x \in I_A} | \, p(x) \, |,$$

the optimal polynomial is

$$p_o(x) = T_m\left(\frac{\lambda_M + \lambda_1 - 2x}{\lambda_M - \lambda_1}\right)/T_m(\gamma).$$

Therefore, for (23) we can take an approximate optimal polynomial

$$p_a(x) = p_o(x).$$

In view of [15, p. 547], we get

$$\begin{aligned} \epsilon^{(m)} &\le \max_{i=1,2,\ldots,M, x \in I_A} \|p_a(J_i(x))\| \\ &\le \max_{i=1,2,\ldots,M} d_i \max_{0 \le j \le d_i - 1, x \in I_A} \left| \frac{p_a^{(j)}(x)}{j!} \right| \\ &\le d \max_{0 \le j \le d-1, x \in I_A} \left| \frac{p_a^{(j)}(x)}{j!} \right| . \end{aligned}$$

Now due to

$$\frac{\lambda_M + \lambda_1 - 2x}{\lambda_M - \lambda_1} \in [-1, 1],$$

so by Part 1 of Lemma 1 we have

$$\epsilon^{(m)} \leq d \max_{0\leq j\leq d-1} \frac{1}{j!}\beta^{2j}m^{2j}C(m,j)/T_m(\gamma).$$

Since, from Part 1 of Lemma 1, $C(m,j)$ is a decreasing function in j and $C(m,0)=1$, we obtain

$$\epsilon^{(m)} \leq \begin{cases} dm^{2(d-1)}/T_m(\gamma), & \text{if } \beta\leq 1, \\ d\beta^{2(d-1)}m^{2(d-1)}/T_m(\gamma), & \text{if } \beta > 1, \end{cases}$$

which shows that (22) holds. □

From mathematical analysis, the above estimates converge to zero as m increases since $T_m^{-1}(\gamma)\approx 2(\gamma+\sqrt{\gamma^2-1})^{-m}$ for large m. Moreover, they usually converge to zero more rapidly whenever λ_1 is not well isolated from λ_M since $\gamma+\sqrt{\gamma^2-1}$ is not close to 1; otherwise, they might converge very slowly when λ_1 is well isolated from λ_M since $\gamma+\sqrt{\gamma^2-1}$ is close to 1 at this point.

It can be seen from Theorem 2 and its proof that the defectiveness of A affects the convergence speed of these methods. In general, the higher the degree of nonlinear elementary divisors is, the more slowly these methods are likely to converge.

Now, we consider the case where the spectrum of A is on the real axis .

Theorem 3 *Assume the spectrum of A to be real and the ordering rule as (21). Let d as before and*

$$\sigma = \frac{\mid\lambda_M\mid}{\mid\lambda_1\mid}, \eta = \frac{2}{\lambda_M^2-\lambda_1^2}.$$

Then

$$\epsilon^{(m)} \leq \begin{cases} d[\frac{m}{2}]^{2(d-1)}/T_{[\frac{m}{2}]}(\frac{\sigma^2+1}{\sigma^2-1}), & \text{if } \eta\leq 1, \\ d(\eta[\frac{m}{2}])^{2(d-1)}/T_{[\frac{m}{2}]}(\frac{\sigma^2+1}{\sigma^2-1}), & \text{if } \eta > 1, \end{cases} \tag{24}$$

where $[\frac{m}{2}]$ denotes the integer part of $\frac{m}{2}$, and $T_{[\frac{m}{2}]}(x)$ is the first kind Chebyshev polynomial of degree $[\frac{m}{2}]$.

Proof. Let the notation $J_i(x), i=1,2,\ldots,M$ as before, and write $D_A=[\lambda_1^2,\lambda_M^2]$. Then according to Theorem 1, we have

$$\epsilon^{(m)} \leq \min_{p\in Q_m, p(0)=1} \max_{i=1,2,\ldots,M, |\lambda_1|\leq|x|\leq|\lambda_M|} \|p(J_i(x))\|.$$

From well known results in approximation theory [5], the optimal polynomial $p_o\in Q_m$ satisfying the above minimax problem exists; furthermore, it can be verified that p_o is an even function using the symmetry of variable x, i.e. $p_o(x)=q_o(x^2)$,$q_o\in Q_{[\frac{m}{2}]}$. It thus follows that

$$\epsilon^{(m)} \leq \min_{q\in Q_{[\frac{m}{2}]}, q(0)=1} \max_{i=1,2,\ldots,M, x\in D_A} \|q(J_i(x))\|. \tag{25}$$

It is well known [21] that for the minimax problem

$$\min_{q\in Q_{\frac{m}{2}},q(0)=1}\max_{x\in D_A}\mid q(x)\mid,$$

the optimal polynomial is

$$q^*(x)=T_{[\frac{m}{2}]}(\frac{\lambda_M^2+\lambda_1^2-2x}{\lambda_M^2-\lambda_1^2})/T_{[\frac{m}{2}]}(\frac{\sigma^2+1}{\sigma^2-1}).$$

We can thus choose an approximate optimal polynomial $q_a(x)=q^*(x)$ for (25). With this choice, we then get

$$\begin{aligned}\epsilon^{(m)} &\leq \max_{i=1,2,\ldots,M,x\in D_A}\|q_a(J_i(x))\| \\ &\leq \max_{i=1,2,\ldots,M} d_i \max_{0\leq j\leq d_i-1,x\in D_A}\mid \frac{q_a^{(j)}(x)}{j!}\mid \\ &\leq d\max_{0\leq j\leq d-1,x\in D_A}\mid \frac{q_a^{(j)}(x)}{j!}\mid .\end{aligned}$$

Since

$$\frac{\lambda_M^2+\lambda_1^2-2x}{\lambda_M^2-\lambda_1^2}\in[-1,1]$$

and due to Part 1 of Lemma 1, we obtain

$$\epsilon^{(m)}\leq d\max_{0\leq j\leq d-1}\frac{1}{j!}\eta^{2j}[\frac{m}{2}]^{2j}C([\frac{m}{2}],j)/T_{[\frac{m}{2}]}(\frac{\sigma^2+1}{\sigma^2-1}),$$

from which it follows that our assertion holds. □

As in Theorem 2, the defectiveness of A might make the convergence speed of these methods become slow. Usually, the higher the degree of nonlinear elementary divisors is, the more slowly these methods are likely to converge. Another important point is that the distribution of the spectrum of A has a strong effect on the convergence speed. Generally speaking, the bigger the difference between $\mid\lambda_1\mid$ and $\mid\lambda_M\mid$, the more slowly these methods may converge; otherwise, they converge relatively rapidly.

4.2 The case of matrices whose spectrum A lies in the open right (left) half plane

We consider upper bounds for $\epsilon^{(m)}$ in which the spectrum of A is contained in an ellipse E which is symmetric with respect to the real axis and whose major axis is either the real axis or parallel to the imaginary axis.

Theorem 4 *Assume that the spectrum of A lies in an ellipse E with the center c, the foci $c+e, c-e$ and the major semiaxis a. Let E be symmetric with respect to the real axis and its major axis either the real axis or parallel to the imaginary axis, and let E exclude the origin. Define d as before, and*

$$\alpha = a/e.$$

Then, if $\mid e \mid \leq 1$,

$$\epsilon^{(m)} \leq d \mid e \mid^{-d+1} m^{d-1} \frac{B_m(\alpha, d-1)}{(\alpha^2-1)^{d-1}} \frac{(\alpha+\sqrt{\alpha^2-1})^m}{\mid T_m(c/e) \mid}; \tag{26}$$

if $\mid e \mid > 1$,

$$\epsilon^{(m)} \leq d m^{d-1} \frac{B_m(\alpha, d-1)}{(\alpha^2-1)^{d-1}} \frac{(\alpha+\sqrt{\alpha^2-1})^m}{\mid T_m(c/e) \mid}, \tag{27}$$

where $T_m(z)$ is the first kind Chebyshev polynomial of degree m.

Proof. Let us define $J_i(z), i = 1, 2, \ldots, M$ analogously. Then

$$\epsilon^{(m)} \leq \min_{p \in Q_m, p(0)=1} \max_{i=1,2,\ldots,M, z \in E} \|p(J_i(z))\|. \tag{28}$$

It is difficult for us to find the optimal polynomial satisfying the above minimax problem, so we can only seek an approximate optimal one. For the minimax problem

$$\min_{p \in Q_m, p(0)=1} \max_{z \in E} \mid p(z) \mid,$$

$$T_m(\frac{c-z}{e})/T_m(c/e)$$

often turns out to be optimal though it is not always the case [10]. Based on this assertion, we take for (28)

$$p_a(z) = T_m(\frac{c-z}{e})/T_m(c/e).$$

Since $(c-z)/e$ belongs to an ellipse with the center at zero, the foci distance one and the major semiaxis $\alpha = a/e$, we can obtain in terms of Theorem 1 and Part 3 of Lemma 1

$$\begin{aligned}
\epsilon^{(m)} &\leq \max_{i=1,2,\ldots,M, z \in E} \|p_a(J_i(z))\| \\
&\leq \max_{i=1,2,\ldots,M} d_j \max_{z \in E, 0 \leq j \leq d_j - 1} \mid \frac{p_a^{(j)}(z)}{j!} \mid \\
&\leq d \max_{z \in E, 0 \leq j \leq d-1} \mid \frac{p_a^{(j)}(z)}{j!} \mid \\
&= d \max_{0 \leq j \leq d-1} \frac{1}{j!} \mid e \mid^{-j} m^j \frac{B_m(\alpha, j)}{(\alpha^2-1)^j} \frac{(\alpha+\sqrt{\alpha^2-1})^m}{\mid T_m(c/e) \mid},
\end{aligned}$$

from which and Part 2 of Lemma 1 it follows that our assertions hold. □

Obviously, it follows from (26) and (27) that $\epsilon^{(m)}$ converges to zero as m goes to infinity. It is easy to see how rapidly these methods converge depends strongly on the factors

$$m^{d-1}\frac{B_m(\alpha, d-1)}{(\alpha^2-1)^{d-1}}, \ \tau_1 = \frac{\mid c/e \mid + \sqrt{\mid c/e \mid^2 - 1}}{\alpha + \sqrt{\alpha^2 - 1}} = \frac{\mid c \mid + \sqrt{\mid c \mid^2 - e^2}}{a + \sqrt{a^2 - e^2}}$$

if the major axis of E is the real axis, or

$$m^{d-1}\frac{B_m(\alpha, d-1)}{(\alpha^2-1)^{d-1}}, \ \tau_2 = \frac{\mid c/e \mid + \sqrt{\mid c/e \mid^2 + 1}}{\alpha + \sqrt{\alpha^2 - 1}} = \frac{\mid c \mid + \sqrt{\mid c \mid^2 + \mid e \mid^2}}{\mid a \mid + \sqrt{\mid a \mid^2 - \mid e \mid^2}}$$

if the major axis of E is parallel to the imaginary axis.

Generally speaking, by Part 3 of Lemma 1, we may claim that these methods are likely to converge more slowly in the presence of complex eigenvalues. If the major axis of E is the real axis, we may assert that the convergence speed is likely to be more rapid, provided that the spectrum is close to the real line since τ_1 will be larger in this case; if E has nearly a circular shape, the convergence speed is likely to be slower since both e and τ_1 are small; if the major axis of E is parallel to the imaginary axis, we may assert that the convergence speed is likely to be slower whenever the spectrum is almost purely imaginary since τ_2 will be close to one at this moment. Also, the higher the degree of nonlinear elementary divisors is, the slower the convergence speed is likely to be.

It can be seen from Theorem 4 that, if $e = 0$, our results (26) and (27) have no meaning. In fact, if $e = 0$, E will degenerate to a disk D with the center c and the radius a. In this case, we can establish the following result.

Theorem 5 *Assume that the spectrum of A lies in the open right (left) half plane, and D contains the spectrum of A but not the origin. Let d as before. Then once m is bigger than certain value depending on a,*

$$\epsilon^{(m)} \leq dc(m,d)m^{d-1}a^{m-d+1}/ \mid c \mid^m, \tag{29}$$

where $c(m,d)$ is a constant less than unity, depending on m and d.

Proof. As before, we define $J_i(z), i = 1, 2, \ldots, M$. Then, in terms of Theorem 1, we get

$$\epsilon^{(m)} \leq \min_{p \in Q_m, p(0)=1} \max_{i=1,2,\ldots,M, z \in D} \|p(J_i(z))\|. \tag{30}$$

For the minimax problem

$$\min_{p \in Q_m, p(0)=1} \max_{z \in D} \mid p(z) \mid,$$

it is known from [29] that the optimal polynomial is

$$(\frac{c-z}{c})^m$$

whose maximum over the disk D is equal to $(\frac{a}{|c|})^m$ and tends to zero as m increases since $\frac{a}{|c|} < 1$. Based on this assertion, we choose for (30)

$$p_a(z) = (\frac{c-z}{c})^m.$$

Therefore,

$$\begin{aligned}
\epsilon^{(m)} &\leq \max_{i=1,2,\ldots,M, z\in D} \|p_a(J_i(z))\| \\
&\leq \max_{i=1,2,\ldots,M} d_i \max_{0\leq j\leq d_i-1, z\in D} \mid \frac{p_a^{(j)}(z)}{j!} \mid \\
&\leq d \max_{0\leq j\leq d-1, z\in D} \mid \frac{p_a^{(j)}(z)}{j!} \mid .
\end{aligned}$$

It is easy to prove that for $0 \leq i \leq d-1$

$$\max_{z\in D} \mid p_a^{(i)}(z) \mid = m(m-1)\cdots(m-i+1)a^{m-i} / \mid c \mid^m .$$

Thus, for $0 \leq j < i$, we have

$$\begin{aligned}
\frac{\max_{z\in D} \mid p_a^{(i)}(z) \mid}{\max_{z\in D} \mid p_a^{(j)}(z) \mid} &= \frac{m(m-1)\cdots(m-i+1)a^{m-i}}{m(m-1)\cdots(m-j+1)a^{m-j}} \\
&= (m-j)(m-j-1)\cdots(m-i+1)a^{j-i}. \qquad (31)
\end{aligned}$$

It is then obvious that (31) is bigger than unity for any $m \geq d-1$ if $a \leq 1$. Therefore, in this case, we have

$$\begin{aligned}
\max_{0\leq j\leq d-1, z\in D} \mid p_a^{(j)}(z) \mid &= m(m-1)\cdots(m-d+2)a^{m-d+1} / \mid c \mid^m \\
&= c(m,d)m^{d-1}a^{m-d+1} / \mid c \mid^m,
\end{aligned}$$

where $c(m,d) = (1-\frac{1}{m})\cdots(1-\frac{d-2}{m})$. It then follows that

$$\epsilon^{(m)} \leq dc(m,d)m^{d-1}a^{m-d+1} / \mid c \mid^m,$$

which is just (29).

If $a > 1$, it is easy to verify that (31) is bigger than unity once m achieves certain value depending a. Hence, (31) still holds if m is bigger than some value depending on a. $\square$

It is easily seen from (29) that $\epsilon^{(m)}$ converges to zero as m increases. Further analysis shows that the defectiveness of A decreases the estimated convergence speed, and the better the spectrum of A is isolated from the origin, the more rapidly these methods may converge.

5 Conclusion

The goal of this paper is to prove how the Krylov subspace methods, which give rise to a similar minimax problem concerning the convergence analysis, behave as m increases. In the paper, we have established the error bounds when A is defective and the spectrum of A either lies in the open right (left) half plane or is on the real axis. The results show that these methods are likely to converge slowly whenever one of three cases occurs: The matrix in question is defective, the distribution of its spectrum not favorable and the Jordan basis of the matrix in question ill-conditioned, such that solving (1) may become difficult.

Finally, we point out that although the upper bounds for $\epsilon^{(m)}$ are likely to be overestimates, they are adequate for the purpose of convergence analysis.

Acknowledgement. I thank Professor L. Elsner for helpful discussions on Lemma 1 and other suggestions and comments, which made me improve the presentation of this paper.

References

[1] B. Atlestam, *Tschebyscheff-polynomials for sets consisting of two disjoint intervals with applications to convergence estimates for the conjugate gradient method*, Technical Report 77.06 R (1977).

[2] O. Axelsson, *Conjugate gradient type methods for unsymmetric and inconsistent systems of linear equations*, Linear Algebra Appl., 29 (1980) 1–16.

[3] O. Axelsson and G. Lindskog, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 48 (1986) 499–523.

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

[5] W. E. Cheney, *Introduction to Approximation Theory*, McGraw-Hill, New York (1966).

[6] A. T. Chronopoulos, *s-Step iterative methods for (non)symmetric (in)definite linear systems*, SIAM J. Numer. Anal., 28 (1991) 1776–1789.

[7] C. de Boor and J. R. Rice, *Extremal polynomials with application to Richardson iteration for indefinite linear systems*, SIAM J. Sci. Statist. Comput., 3 (1982) 47–57.

[8] S. C. Eisenstat, H. C. Elman and M. H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 20 (1983) 345–357.

[9] H. C. Elman, *Iterative methods for large sparse nonsymmetric systems of linear equations*, Ph. D. Thesis, Yale University (1982).

[10] B. Fischer and R. W. Freund, *Chebyshev polynomials are not always optimal*, J. Approx. Theory, 65 (1991) 261–272.

[11] R. W. Freund and N. M. Nachtigal, *QMR: a Quasi-minimal method for non-Hermitian linear systems*, Numer. Math., 60 (1991) 291–313.

[12] R. W. Freund, G. H. Golub and N. M. Nachtigal, *Iterative solution of linear systems*, Acta Numerica (1992) 1–44.

[13] F. R. Gantmacher, *The Theory of Matrices*, Vol. 2, Chelsea Publishing Company, New York (1959).

[14] G. H. Golub and R. S. Varga, *Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods*, Numer. Math., 3 (1961) 147–168.

[15] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The John Hopkins University Press, 2nd Edition, Baltimore (1989).

[16] Z. Jia, *The convergence of generalized Lanczos methods for large unsymmetric eigenproblems*, submitted for publication, 1993.

[17] V. I. Lebedev, *Iterative methods for solving operator equations with a spectrum contained in several intervals*, USSR Comput. Math. Math. Phys., 9 (1969) 17–24.

[18] T. A. Manteuffel, *An iterative method for solving nonsymmetric linear systems with dynamic estimation of parameters*, Digital Computer Laboratory Reports, Rep. UIUCDS-R-75-7758, University of Ill., (1975).

[19] T. A. Manteuffel, *The Tchebychev iteration for nonsymmetric linear systems*, Numer. Math., 28 (1977) 307–327.

[20] N. W. Nachtigal, S. C. Reddy and L. N. Trefethen, *How fast are nonsymmetric matrix iterations*? SIAM J. Matrix Anal. Appl., 13 (1992), pp. 778–795.

[21] T. R. Rivlin, *The Chebyshev Polynomials*, J. Wiley and Sons Inc., New York (1974).

[22] Y. Saad, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comput., 37 (1981) 105–126.

[23] Y. Saad, *The Lanczos biorthogonalization algorithm and other oblique projection methods for solving unsymmetric systems*, SIAM J. Numer. Anal., 19 (1982) 485–506.

[24] Y. Saad, *Practical use of some Krylov subspace methods for solving indefinite and nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 5 (1984) 203–228.

[25] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear system*, SIAM J. Sci. Stat. Comput., 7 (1986) 856–869.

[26] Y. Saad, *Krylov subspace methods on supercomputers*, SIAM J. Sci. Statist. Comput., 10 (1989) 1200–1232.

[27] J. Stoer and R. W. Freund, *On the solution of large indefinite systems of linear equations by conjugate gradient algorithms*, in Computing Methods in Applied Science and Engineering V, eds. R. Glowinski and J. L. Lions, North-Holland, Amsterdam (1982) 35–53.

[28] L. N. Trefethen, *Pseudospectra of Matrices*, in D. F. Griffiths and G. A. Watson, Numerical Analysis 1991, Longman, 1992.

[29] R. S. Varga, *A comparison of the successive overrelaxation method and semi-iterative method using Chebyshev polynomials*, J. Soc. Indust. Appl. Math., 5 (1957) 39–46.

[30] H. A. Van der Vorst and C. Vuik, *The superlinear convergence behaviour of GMRES*, J. Comput. Appl. Math., 48 (1993), pp. 327–341.

[31] P. K. W. Vinsome, *Orthomin, an iterative method for solving sparse sets of simultaneous linear equations*, in Proc. Fourth SPE Symposium on Reservoir Simulation, Los Angeles (1976) 149–160.

[32] O. Widlund, *A Lanczos method for a class of nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 15 (1978) 801–812.