# ON THE CONVERGENCE OF THE ARNOLDI PROCESS FOR EIGENVALUE PROBLEMS

M. BELLALIJ[*], Y. SAAD[†], AND H. SADOK[‡]

**Abstract.** This paper takes another look at the convergence of the Arnoldi procedure for solving nonsymmetric eigenvalue problems. Three different viewpoints are considered. The first uses a bound on the distance from the eigenvector to the Krylov subspace from the smallest singular value of matrix consisting of the Krylov basis. A second approach relies on the Schur factorization. Finally, a third viewpoint, uses expansions in the eigenvector basis for the diagonalizable case.

## 1. Introduction.

Projection techniques on Krylov subspaces are currently among the most important tools used to compute eigenvalues and eigenvectors of large sparse non-Hermitian matrices. The convergence of these methods has been analyzed in detail in the Hermitian case, but the analysis becomes much more difficult in the non-Hermitian or non-normal case and so there are few results available in the literature. This is in contrast with the convergence analysis of Krylov methods for solving linear systems which received far more attention. In this paper we will examine several approaches to the problem.

Each of these approaches utilizes a different 'parameter', or set of parameters, which is (are) singled out as the main value (s) on which the analysis depends. Often this parameter is difficult to estimate. This approach is similar to standard analyses where there is a core expression used as a measure against which an error bound is developed. For example, for linear systems, there has been analyses which exploit the polynomial representation of a vector in the Krylov subspace. Thus, when solving a linear system $Ax = b$, in the case when $A$ is diagonalizable, with a matrix $X$ of eigenvectors, the standard bound for GMRES [13]

$$\|b - Ax_m\|_2 \leq \kappa_2(X) \min_{p \,\in\, \mathbb{P}_m} \; \max_{\lambda \,\in\, \Lambda(A)} |p(\lambda)| \; \|b - Ax_0\|_2 \ ,$$

uses the min-max quantity on the right as a parameter which is then estimated in certain ways. Here and throughout the paper $\|v\|_2$ denotes the 2-norm of a vector $v$ and for a matrix $A$, $\|A\|_2$ denotes its 2-norm and $\kappa_2(A)$ denotes its 2-norm condition number of $A$. We will also denote by $I$ the identity matrix and by $e_k$ its $k$th column and assume exact arithmetic. In general, the above residual bound for GMRES is not satisfactory because it involves the condition number of $X$, the matrix that diagonalizes $A$, which is not known and which can be very large. An alternative, exploited in [14] (see also Ipsen [7]), uses as a primary indicator the (1,1) entry of the inverse of $K_m^T K_m$ where $K_m$ is the matrix whose columns are the vectors of the canonical basis of Krylov subspace.

Other ways to analyze convergence of Krylov methods have been explored. For GMRES, one such type of analysis assumes that the field of values does not contain the origin. With this, quite a few bounds can be found; we refer for example to [5, 3, 4, 12]. Underlying the difficulty is the fact that the norm $\|p(A)\|_2$ is not always

[*] Université de Valenciennes, Le Mont Houy, 59313 Valenciennes Cedex, France. E-mail: Mohammed.Bellalij@univ-valenciennes.fr.

[†]University of Minnesota, Dept. of Computer Science and Engineering E-mail: saad@cs.umn.edu. Work supported by NSF under grant ACI-0305120, and by the Minnesota Supercomputing Institute.

[‡]L.M.P.A, Université du Littoral, 50 rue F. Buisson BP699, F-62228 Calais Cedex, France. E-mail: sadok@lmpa.univ-littoral.fr.

easy to estimate in the non-normal case. Yet, most bounds will rely in an implicit way, in an estimate of the minimum of $\|p(A)v\|_2$ over some normalized polynomials.

In the sequel we will consider three distinct approaches for analyzing the convergence of the Arnoldi process. It is important to recall at the outset that, unlike the situation for Hermitian matrices, there are no easy optimality results to be exploited. Most of our analysis will be based on estimating the distance from an exact eigenvector from the Krylov subspace.

**2. Background.** Consider the eigenvalue problem: find $u$ belonging to $\mathbb{C}^N$ and $\lambda$ belonging to $\mathbb{C}$ such that

$$A\,u = \lambda u, \tag{2.1}$$

where the matrix $A$ is of order $N$. For a given vector $v \in \mathbb{C}^N$, the Krylov subspace $\mathbb{K}_m(A, v)$ is defined by

$$\mathbb{K}_m(A, v) = \text{span}\{v, A\,v, \dots, A^{m-1}v\}. \tag{2.2}$$

**2.1. The Arnoldi process.** The Arnoldi method computes approximate eigenpairs $\widetilde{\lambda}^{(m)}, \widetilde{u}^{(m)}$ by enforcing the standard Petrov-Galerkin condition

$$\widetilde{u}^{(m)} \in \mathbb{K}_m(A, v), \tag{2.3}$$

and

$$(A\,\widetilde{u}^{(m)} - \widetilde{\lambda}^{(m)}\,\widetilde{u}^{(m)}\,, A^i v) = 0 \qquad \text{for} \quad i = 0, \dots, m-1\ . \tag{2.4}$$

The standard way of extracting the approximate eigenpairs from the above conditions is to resort to the Arnoldi algorithm which generates an orthonormal basis $v_1, \dots, v_m$ of $\mathbb{K}_m$ in which the conditions (2.4) are expressed.

Algorithm 2.1. *Arnoldi*

```
Input: Initial vector $v$, and $m$.
Set $v_1 = \frac{v}{\|v\|_2}$
For $j = 1, ..., m$ do
   Compute $w := Av_j$
   For $i = 1, \dots, j,$ do   $\left\{ \begin{array}{l} h_{i,j} := (w, v_i) \\ w := w - h_{i,j} v_i \end{array} \right.$
   $h_{j+1,j} = \|w\|_2;$        $v_{j+1} = w/h_{j+1,j}$
End
```

The outputs of the algorithm are an othonormal basis $V_m = [v_1, v_2, ..., v_m]$ and a Hessenberg matrix $H_m$ whose entries are the scalars $h_{ij}$ generated by the procedure. In addition, the following relations are satisfied:

1. $AV_m = V_m H_m + h_{m+1,m} v_{m+1} {e_m}^T$
2. $V_m^T A V_m = H_m$

The approximate eigenvalue problem can now be written as

$$V_m^H\,AV_m\,y^{(m)} = \widetilde{\lambda}^{(m)}\,y^{(m)}, \tag{2.5}$$

where $y^{(m)} = R_m\,z^{(m)}$. Which is equivalent to

$$V_m^H(A - \tilde{\lambda}^{(m)} I)V_m y^{(m)} = 0.$$

Approximate eigenvalues are eigenvalues of $H_m$ and are obtained by solving the preceding eigenvalue problem, the associated approximate eigenvectors are $\tilde{u}^{(m)} = V_m y^{(m)}$. Typically, a few of the outermost eigenvalues will converge first.

**2.2. Convergence.** The convergence of projection methods for eigenvalue problems, such as the Arnoldi algorithm, is difficult to analyze in the non-normal case. The simplest analysis, though somewhat incomplete, uses the distance of a given eigenvector from the Krylov subspace, see [11]. Let $\mathcal{P}_m$ be the orthogonal projector onto $\mathbb{K}_m$. Then, the approximate problem amounts to solving

$$\mathcal{P}_m(Ax - \lambda x) = 0, \quad x \ \in \ \mathbb{K}_m,$$

or in operator form

$$\mathcal{P}_m A \mathcal{P}_m x = \lambda x.$$

Define, $A_m \equiv \mathcal{P}_m A \mathcal{P}_m$. Then the following theorem is easy to prove (see [11]).

THEOREM 2.1. *Let $\gamma_m = \|\mathcal{P}_m A(I - \mathcal{P}_m)\|_2$. Then the residual norms of the pairs $\lambda, \mathcal{P}_m u$ and $\lambda, u$ for the linear operator $A_m$ satisfy, respectively*

$$\|(A_m - \lambda I)\mathcal{P}_m u\|_2 \le \gamma_m \|(I - \mathcal{P}_m)u\|_2,$$

$$\|(A_m - \lambda I)u\|_2 \le \sqrt{|\lambda|^2 + \gamma_m^2} \ \|(I - \mathcal{P}_m)u\|_2 \ .$$

Note that the second bound of the theorem gives an unusual result in that it states how accurate the *exact eigenpair* is with respect to the *approximate problem.* This is stated in terms of the distance of the exact eigenvector $u$ from the Krylov subspace. The remaining issue is how to estimate $\|(I - \mathcal{P}_m)u\|_2$.

## 3. Projection-based analysis.

In the following we analyze the distance

$$d(w, X) \equiv \min_{x \ \in \ X} \|w - x\|_2$$

in general terms where $X$ is an arbitrary subspace of some dimension $m$. We begin by showing a number of simple results. First observe that given *any* basis $V$ of the subspace $X$, $x$ can be written as $Vy$, where $y \ \in \ \mathbb{C}^m$, so that

$$\|w - x\|_2^2 = \|w - Vy\|_2^2 = w^H w - 2w^H V y + y^H V^H V y.$$

The above expression is in fact of the form

$$\|w - x\|_2^2 = \begin{pmatrix} 1 \\ -y \end{pmatrix}^H \underbrace{\begin{pmatrix} w^H w & w^H V \\ V^H w & V^H V \end{pmatrix}}_{\equiv C} \begin{pmatrix} 1 \\ -y \end{pmatrix} \ . \tag{3.1}$$

Note that minimizing $\|w - x\|_2$ over $X$ is equivalent to minimizing $\|w + x\|_2$ over the same subspace, so the signs of the $y$'s in the above expression can be changed when seeking the minimum distance. In the end,

$$\min_{x \ \in \ X} \|w - x\|_2^2 = \min_{y \ \in \ \mathbb{C}^m} \|w - Vy\|_2^2 = \min_{y \ \in \ \mathbb{C}^m} \begin{pmatrix} 1 \\ y \end{pmatrix}^H \ C \ \begin{pmatrix} 1 \\ y \end{pmatrix} \tag{3.2}$$

where $C$ was defined in (3.1). It is interesting to note the above minimization can be converted into a trivial generalized eigenvalue problem:

$$\min_{y \ \in \ \mathbb{C}^m} \begin{pmatrix} 1 \\ y \end{pmatrix}^H \ C \ \begin{pmatrix} 1 \\ y \end{pmatrix} = \min_{z \ \in \ \mathbb{C}^{m+1}, \ e_1^H z = 1} z^H C z = \min_{z \ \in \ \mathbb{C}^{m+1}} \frac{z^H C z}{z^H e_1 e_1^H z} \ . \tag{3.3}$$

Therefore, the smallest squared distance achieved between the vector $w$ and vectors of the subspace $X$, is the smallest eigenvalue of the generalized eigenvalue problem $Cz = \mu(e_1 e_1^H)z$. This problem has only one finite eigenvalue as can be seen from converting it with the help of the Cholesky factorization $C = LL^H$:

$$LL^H z = \mu(e_1 e_1^H)z \to L^H z = \mu(L^{-1}e_1 e_1^H L^{-H})L^H z \to \frac{1}{\mu}u = (L^{-1}e_1 e_1^H L^{-H})u,$$

where we have set $u = L^H z$. The only nonzero eigenvalue of the rank-one matrix $L^{-1}e_1 e_1^H L^{-H}$ is $e_1^H L^{-H} L^{-1} e_1$. So

$$\mu_{min} = \frac{1}{e_1^H L^{-H} L^{-1} e_1} = \frac{1}{e_1^H C^{-1} e_1}.$$

Therefore, we have proved the following result:

LEMMA 3.1. *Let $X$ be an arbitrary subspace with a basis $V = [v_1, \cdots, v_m]$ and let $w \notin X$. Let $\mathcal{P}$ be the orthogonal projector onto $X$. Then, we have*

$$\|(I - \mathcal{P})w\|_2^2 = \frac{1}{e_1^H C^{-1} e_1}$$

*where*

$$C = [w, V]^H [w, V] = \begin{pmatrix} w^H w & w^H V \\ V^H w & V^H V \end{pmatrix} .$$

*Proof.* The result was proved above. An alternative proof which will help establish some relations is as follows. Given an arbitrary vector $w \in \mathbb{C}^N$, observe that

$$\|(I - \mathcal{P})w\|_2^2 = w^H (I - \mathcal{P})(I - \mathcal{P})w = w^H (I - \mathcal{P})w = w^H w - w^H \mathcal{P} w .$$

with $\mathcal{P} = V(V^H V)^{-1} V^H$. From this it follows that

$$\|(I - \mathcal{P})w\|_2^2 = w^H w - w^H \mathcal{P} w = w^H w - w^H V (V^H V)^{-1} V^H w . \qquad (3.4)$$

The right-hand side of (3.4) is simply the Schur complement of the (1,1) entry of $C$, which as is well-known is the inverse of the (1,1) entry of $C^{-1}$. □

Let $\sigma_{min}[w, V]$ and $\sigma_{max}[w, V]$ be the smallest and largest singular values of $[w, V]$. Then a consequence of (3.2) and (3.3) is that

$$\sigma_{min}[w, V] \le \|(I - \mathcal{P})w\|_2 \le \sigma_{max}[w, V] . \qquad (3.5)$$

This is because $e_1^H C^{-1} e_1$ is a Rayleigh quotient of $C^{-1}$, and so

$$\frac{1}{\lambda_{max}(C)} \equiv \lambda_{min}(C^{-1}) \le e_1^H C^{-1} e_1 \le \lambda_{max}(C^{-1}) \equiv \frac{1}{\lambda_{min}(C)} ,$$

and the result follows by inverting the above inequalities.

Clearly, the right part of the bound (3.5) is too pessimistic. We expect $\|(I - \mathcal{P})w\|_2$ to be closer to the smallest singular value. A sharper result can be obtained by exploiting an appropriate singular vector in (3.3).

LEMMA 3.2. *Let $\sigma_{min}[w, V]$ be the smallest singular value of $[w, V]$ and $w_{min}$ the associated right singular vector and assume that $e_1^H w_{min} \neq 0$. Then,*

$$\sigma_{min}[w, V] \le \|(I - \mathcal{P})w\|_2 \le \frac{\sigma_{min}[w, V]}{|e_1^H w_{min}|} . \qquad (3.6)$$

*Proof.* To prove the right inequality, we use (3.2) and (3.3), and select as particular vector $z$ the right singular vector $w_{min}$. This results in,

$$\min_{x \,\in\, X} \|w - x\|_2^2 \le \frac{w_{min}^H C w_{min}}{w_{min}^H e_1 e_1^H w_{min}} = \frac{w_{min}^H [w, V][w, V]^H w_{min}}{w_{min}^H e_1 e_1^H w_{min}} = \frac{\sigma_{min}^2 [w, V]}{|e_1^H w_{min}|^2} \ .$$

The left inequality was established above. It also follows from (3.3) and the observation that $|e_1^H z| < \|z\|_2$. □

Consider now using this result for the situation of interest, i.e., when $X$ is a Krylov subspace $\mathbb{K}_m$ and $w$ is an eigenvector $u_i$ of $A$. The left side of (3.6) indicates that we cannot have a good approximation if $[u_i, V]$ is well conditioned. Linear dependence of the set $[u_i, V]$ can take place in two ways. As expected, the first is when $u_i$ is close to the subspace $\mathbb{K}_m$. The second is when the basis $V$ is ill-conditioned. However, in this situation we would also need $e_1^H w_{min}$ to be not too small.

Note that the above bound has one additional degree of freedom, which is the selection of the basis.

## 4. Analysis in terms of Schur vectors.

In this section we will exploit a Schur decomposition of $A$ of the form

$$A = QRQ^H,$$

where $Q$ is unitary, $R$ is upper triangular with its (1,1) entry being the eigenvalue to which convergence is being analyzed. We thus write $R$ in the form

$$R = \begin{pmatrix} \lambda_1 & s^H \\ 0 & R_1 \end{pmatrix} \ . \tag{4.1}$$

It will be assumed that $\lambda_1$ is a simple eigenvalue, so the eigenvalues $\lambda_2, \cdots, \lambda_N$ of $R_1$ are all distinct from $\lambda_1$. Since the powers of $A$ are at the basis of Krylov methods, we examine the sequence of the powers of the matrix $R$.

Lemma 4.1. *For any $k > 0$ we have*

$$R^k = \begin{pmatrix} \lambda_1^k & s_k^H \\ 0 & R_1^k \end{pmatrix} \quad \textit{with} \quad s_k^H = s^H (\lambda_1 I - R_1)^{-1} (\lambda_1^k I - R_1^k) \ . \tag{4.2}$$

*Proof.* The proof is by induction and is straightforward. □

If we apply this result to arbitrary polynomials the following corollary will be obtained.

Corollary 4.2. *For any polynomial $p$, we have:*

$$p(R) = \begin{pmatrix} p(\lambda_1) & s^H q(R_1) \\ 0 & p(R_1) \end{pmatrix} \quad \textit{with} \quad q(\lambda) = \frac{p(\lambda_1) - p(\lambda)}{\lambda_1 - \lambda} \ . \tag{4.3}$$

It is interesting to note in passing that the above result can be applied recursively with the matrix $R$ replaced by $R_1$.

Now consider the problem of estimating the distance of the first eigenvector, which is $q_1$, the first column of $Q$, from the Krylov subspace. We need to minimize $\|q_1 - K_m y\|_2$ over all vectors $y \in \mathbb{C}^m$, where $K_m$ is the Krylov basis $K_m = [v, Av, \cdots, A^{m-1}v]$ of $\mathbb{K}_m$. However, since each basis vector is of the form $A^j v = QR^j Q^H v$, we can work in the Schur basis and minimize instead

$$\|e_1 - [z, Rz, \cdots, R^{m-1} z] y\|_2 \equiv \|e_1 - p_{m-1}(R) z\|_2 \ .$$

where $z = Q^H v$ over polynomials $p_{m-1}$ of degree $\leq m-1$.

LEMMA 4.3. *Let $R$ be given by (4.1), and $z = Q^H v$ where $v$ is of norm unity and let*

$$z = \begin{pmatrix} \eta \\ \tilde{z} \end{pmatrix}; \quad t = \begin{pmatrix} 1 \\ (\lambda_1 I - R_1)^{-H} s \end{pmatrix} .$$

*Define*

$$\epsilon_m = \min_{\substack{p \ \in \ \mathbb{P}_{m-1} \\ p(\lambda_1)=1}} \|p(R_1)\tilde{z}\|_2 .$$

*Then, assuming $(t, z) = t^H z \neq 0$ we have*

$$\min_{p \in \mathbb{P}_{m-1}} \|e_1 - p(R)z\|_2 \leq \frac{\epsilon_m}{|\cos\theta(t,z)|} . \quad (4.4)$$

*Proof.* We define $\tilde{t}$ to be the vector with bottom $n-1$ components of $t$, i.e.,

$$\tilde{t}^H = s^H(\lambda_1 I - R_1)^{-1} ,$$

and seek first an approximation to $\eta e_1$ by writing

$$\eta e_1 - p(R)z = \begin{pmatrix} (1-p(\lambda_1))\eta - \tilde{t}^H[p(\lambda_1)I - p(R_1)]\tilde{z} \\ -p(R_1)\tilde{z} \end{pmatrix} ,$$

where the previous corollary was exploited. We now select the polynomial $p_{m-1}$ such that $p_{m-1}(\lambda_1) = 1$ and $\|p_{m-1}(R_1)\tilde{z}\|_2$ is minimum. Then for this polynomial,

$$\eta e_1 - p_{m-1}(R)z = \begin{pmatrix} -\tilde{t}^H\tilde{z} + \tilde{t}^H p_{m-1}(R_1)\tilde{z} \\ -p_{m-1}(R_1)\tilde{z} \end{pmatrix} .$$

From this we get

$$(\eta + \tilde{t}^H\tilde{z})e_1 - p_{m-1}(R)z = \begin{pmatrix} \tilde{t}^H p_{m-1}(R_1)\tilde{z} \\ -p_{m-1}(R_1)\tilde{z} \end{pmatrix} .$$

Using the notation defined above for $\epsilon_m$, dividing both sides by $\eta + \tilde{t}^H\tilde{z}$, under the assumption that $\eta + \tilde{t}^H\tilde{z} \neq 0$, results in

$$e_1 - \frac{p_{m-1}(R)}{(\eta + \tilde{t}^H\tilde{z})}z = \frac{1}{(\eta + \tilde{t}^H\tilde{z})} \begin{pmatrix} \tilde{t}^H p_{m-1}(R_1)\tilde{z} \\ -p_{m-1}(R_1)\tilde{z} \end{pmatrix} .$$

Calling $p$ the polynomial $p_{m-1}(\lambda)/[\eta + \tilde{t}^H\tilde{z}]$ and taking 2-norms of both sides, gives the following upper bound, with the help of the Cauchy-Schwarz inequality,

$$\|e_1 - p(R)z\|_2 \leq \frac{\sqrt{1 + \|\tilde{t}\|_2^2}}{|\eta + \tilde{t}^H\tilde{z}|} \ \times \ \epsilon_m . \quad (4.5)$$

The numerator of the fraction in the right-hand side of (4.5) is the norm of the vector $t$, and the denominator is the absolute value of the inner product $(t, z)$. Since we assumed that $\|z\|_2 = 1$ then the factor in front of the term $\epsilon_m$ in (4.5) is the inverse of the cosine of the angle between $t$ and $z$. The minimum on the left-hand side of (4.4) does not exceed the right-hand side of (4.5), so the result is proved. □

The lemma can be translated in terms of known quantities related to the original basis.

THEOREM 4.4. *Let $w_1$ be the left eigenvector of $A$ associated with $\lambda_1$ and assume that $\cos\theta(w_1, v) \neq 0$. Let $P_1$ be the orthogonal projector onto the right eigenspace associated with $\lambda_1$, i.e., : $P_1 = q_1 q_1^H$, and let $B_1$ be the linear operator $B_1 = (I - P_1)A(I - P_1)$. Define*

$$\epsilon_m = \min_{\substack{p \,\in\, \mathbb{P}_{m-1} \\ p(\lambda_1)=1}} \|p(B_1)(I - P_1)v\|_2. \tag{4.6}$$

*Then, we have*

$$\|(I - \mathcal{P}_m)q_1\|_2 = \min_{y \,\in\, \mathbb{C}^m} \|q_1 - K_m y\|_2 \ \leq \frac{\epsilon_m}{|\cos\theta(w_1, v)|} \ . \tag{4.7}$$

*Proof.* As was stated above, the theorem is nothing but a translation of the lemma into the $Q$ basis. First, we have already seen that

$$\|(I - \mathcal{P}_m)q_1\|_2 = \min_y \|q_1 - K_m y\|_2 = \min_{p \in\, \mathbb{P}_{m-1}} \|q_1 - p(A)v\|_2 = \min_{p \in\, \mathbb{P}_{m-1}} \|e_1 - p(R)z\|_2$$

which only re-expresses the quantity $\|q_1 - p(A)v\|_2$ in the basis $Q$, so $\|q_1 - p(A)v\|_2 = \|e_1 - p(R)z\|_2$.

Second, it can be seen that the vector $w_1 = Qt$, where $t$ is defined in the lemma is a left eigenvector associated with $\lambda_1$. Indeed we have

$$w_1^H(\lambda_1 I - A) = t^H Q^H Q(\lambda_1 I - R)Q^H = [1 \quad s^H(\lambda_1 I - R_1)^{-1}] \begin{pmatrix} 0 & -s^H \\ 0 & \lambda_1 I - R_1 \end{pmatrix} Q^H = 0.$$

So, $(w_1, v) = (Qt, Qz) = (t, z)$. Finally, the scalar $\epsilon_m$ in this theorem is identical with the one in the lemma. It is just expressed with a different basis. Indeed, in the $Q$ basis the vector $(I - P_1)v$ is $\tilde{z}$. In the same basis, the operator $(I - P_1)A(I - P_1)$ is represented by the matrix $R_1$. □

The above inequality gives an analysis in terms of the variable $\epsilon_m$. It differs from other bounds by not using eigenbases or spectral expansions [11, 15]. Whether or not eigenvectors are used, there are unknown quantities. On the one hand the eigenbasis expansion can lead to large coefficients (Ill-conditioned bases for example). The Schur factorization on the other hand does not have such difficulties as it should retain the non-normality effects in the quantity $\epsilon_m$.

A number of papers give a detailed analysis of the minimum of $\|p(A)v\|_2$ or $\|p(A)\|_2$ under a normalization assumption of the form $p(0) = 1$, see for example the papers [1, 2]. In [2] for example, it is shown that there is a universal constant $\widehat{c} \leq 33.75$ such that if $W(A)$ is the field of values of $A$ then for any polynomial

$$\|p(A)\|_2 \leq \widehat{c} \ \max_{z \,\in\, W(A)} |p(z)| \ .$$

These bounds can easily be adapted to our situation to have an idea of the scalar $\epsilon_m$.

Consider, for example, the situation where $\lambda_1$ is the dominant eigenvalue and the other eigenvalues are located in a disk of center $c$ and radius $r$ not containing

$\lambda_1$. For illustration we will take instead of the optimal polynomial, a simple power polynomial, namely

$$p_k(\lambda) = \frac{(\lambda - c)^k}{(\lambda_1 - c)^k}.$$

Then,

$$p_k(\lambda_1) = 1; \quad p_k(R_1) = \left( \frac{R_1 - cI}{\lambda_1 - c} \right)^k .$$

Notice that the spectral radius of $\frac{R_1 - cI}{\lambda_1 - c}$ is

$$\rho \left( \frac{R_1 - cI}{\lambda_1 - c} \right) = \frac{r}{|\lambda_1 - c|} < 1,$$

so $(R_1 - cI)^k / (\lambda_1 - c)^k$ tends to zero. We can write

$$\hat{\epsilon}_m \equiv \| p_{m-1}(R_1) \tilde{z} \|_2 = \frac{\|(R_1 - cI)^{m-1} \tilde{z}\|_2}{|\lambda_1 - c|^{m-1}} = \delta_m \left( \frac{r}{|\lambda_1 - c|} \right)^{m-1} ,$$

where $\delta_m$ is a sequence which converges to a certain constant $\delta \leq \|\tilde{z}\|_2$. One of the difficulties of all methods based on an analysis of this sort, is the well-known fact that the sequence $\delta_m$ can become very large before settling to its limit. This is characteristic of highly non-normal matrices.

**5. Analysis in terms of eigenvectors.** A common technique for estimating $\|(I - \mathcal{P}_m) u_i\|_2$ assumes that $A$ is diagonalizable and expands $v$, the first vector of the Krylov sequence, in the eigen-basis. If $A$ is diagonizable then for some matrix of eigenvectors $U$, we have $A = U \Lambda U^{-1}$, where $\Lambda = diag(\lambda_1, \ldots, \lambda_N)$. We examine the convergence of a given eigenvalue which is indexed by 1, i.e., we consider $u_1$, the 1-st column column of $U$. The initial vector $v$ is expanded in the eigen-basis as $v = \sum_{j=1}^{N} \alpha_j u_j$. It is assumed throughout that $\alpha_1 \neq 0$ and $\|u_j\|_2 = 1$ for all $j$. In [11], it was shown that, under the above assumptions, we have in the general non-Hermitian case,

$$\|(I - \mathcal{P}_m) u_1\|_2 \leq \xi_1 \epsilon_1^{(m)} , \tag{5.1}$$

where

$$\xi_1 = \sum_{j \neq 1} \left| \frac{\alpha_j}{\alpha_1} \right| \quad \text{and} \quad \epsilon_1^{(m)} = \min_{\left\{ \substack{p \ \in \mathbf{P_{m-1}} \\ p(\lambda_1)=1} \right.} \max_{j \neq 1} |p(\lambda_j)| . \tag{5.2}$$

This result, which requires a few easy manipulations to prove, does not exploit the orthogonality of the eigenbasis in the normal case. In this particular case, the same result holds but $\xi_1$ can be sharpened to

$$\xi_{1,normal} = \frac{\sqrt{\sum_{j=2}^{n} |\alpha_j|^2}}{|\alpha_1|},$$

which represents the tangent of the angle between $u_1$ and $v$. Note that the distance $\|(I - \mathcal{P}_m) u_1\|_2$ is nothing but the sine of the angle $\angle(u_1, \mathbb{K}_m)$.

Once an estimate for $\epsilon_1^{(m)}$ is available, Theorem 2.1 can be invoked to give an idea on the residual of the exact eigenpair with respect to the approximate (projected) problem. See also [15, th. 3.10] for an alternative approach which exploits the spectral decomposition.

**5.1. The approximation theory viewpoint.** What is left to do is to estimate $\epsilon_1^{(m)}$. For the sake of notational convenience, we will consider estimating $\epsilon_1^{(m+1)}$ instead of $\epsilon_1^{(m)}$. The underlying problem is one of approximation theory. For any continuous function $f$ defined on a compact set $\Omega$, denote the uniform norm:

$$\|f\|_\infty = \max_{z \,\in\, \Omega} \; |f(z)| \; . \tag{5.3}$$

The set $\Omega$ will later be taken to be the spectrum of $A$ excluding $\lambda_1$. Estimating $\epsilon_1^{(m+1)}$ amounts to finding an upper bound for the distance, in the sense of the inf-norm just defined, between the function $f(z) = 1$ and polynomials of degree $\leq m$ of the form $p(z) = (z - \lambda_1)q(z)$, or, equivalently:

$$\epsilon_1^{(m+1)} = \min_{q \,\in\, \mathbb{P}_{m-1}} \; \|1 - (z - \lambda_1)q(z)\|_\infty \; .$$

We recall that a subspace $S$ of continuous functions on $\Omega$, generated by $k$ functions $\phi_1, \cdots, \phi_k$ satisfies the Haar condition if each function in $S$ has at most $k-1$ distinct roots. This means that any linear combination of the $\phi_i$'s vanishes iff it has $k$ distinct roots in $\Omega$. Let $f$ be a continuous function and let $p^*$ be the best uniform approximation of $f$ over $\Omega$. The difference $f - p^*$ reaches its maximum modulus at a number of *extremal points*. The characteristic property [10] of the best approximation states the following.

Theorem 5.1. *Let $f$ be a continuous function and $S$ a $k$-dimensional subspace of the space of continuous functions on $\Omega$, which satisfies the Haar condition. Then $p^* \in S$ is the best uniform approximation of $f$ over $\Omega$, iff there exist $r$ extremal points $z_i, i = 1, \cdots, r$ in $\Omega$, and positive numbers $\mu_1, \cdots, \mu_r$, with $k + 1 \leq r \leq 2k + 1$ such that*

$$\sum_{i=1}^{r} \mu_i \overline{[f(z_i) - p^*(z_i)]} \phi(z_i) = 0 \quad \forall \; \phi \; \in \; S. \tag{5.4}$$

One important point here is that the number of extremal points is only known to be between $k + 1$ and $2k + 1$ in the general complex case. That $r$ must be $\geq k + 1$ is a consequence of the Haar condition and can be readily verified. When $\Omega$ is real, then $r = k + 1$. The fact that $r$ is only known to be $\leq 2k + 1$ in the complex case, comes from Caratheodory's characterization of convex hulls which expresses a point in $co(\Omega)$, the convex hull of $\Omega$, as a convex combination of $k + 1$ points of $\Omega$ in real spaces and $2k + 1$ points of $\Omega$ in complex spaces.

We will now translate the above result for our situation. Let $\Omega = \Lambda(A) \backslash \{\lambda_1\}$ and $S = \text{span}\{\phi_j(z)\}_{j=1,\cdots,m}$ where $\phi_j(z) = (z - \lambda_1)z^{j-1}$. Then, the dimension of $S$ is $m$ and therefore the theorem states that there are $r$ eigenvalues from the set $\Omega$, with $m + 1 \leq r \leq 2m + 1$ such that

$$\sum_{k=1}^{r} \mu_k \overline{[1 - (\lambda_{k+1} - \lambda_1)q^*(\lambda_{k+1})]} \phi_j(\lambda_{k+1}) = 0 \quad j = 1, \ldots, m \; .$$

Although we do not know how many extremal points there are we can still express the best polynomial by selecting any set of $m$ extremal points. Assume without loss of generality that these points are labeled from 2 to $m+1$. Let $p^*(z) = (z - \lambda_1)q^*(z)$. We can write $1 - p^*(\lambda_k)$ at each of the extremal points $\lambda_k$ as

$$1 - p^*(\lambda_k) = \rho e^{i\theta_k}$$

where $\theta_k$ is a real number and $\rho$ is real and positive. Then it is easily seen that

$$1 - p^*(z) = \frac{\sum_{k=2}^{m+2} e^{i\theta_k} l_k(z)}{\sum_{k=2}^{m+2} e^{i\theta_k} l_k(\lambda_1)}, \tag{5.5}$$

where each $l_k(z)$ is the Lagrange polynomial:

$$l_k(z) = \prod_{\substack{j=2 \\ j\neq k}}^{m+2} \frac{z-\lambda_j}{\lambda_k - \lambda_j} \ . \tag{5.6}$$

Indeed, $1 - p^*(z)$, which is of degree $m$, takes the values $\rho e^{i\theta_k}$ at the $m+1$ points $\lambda_2, \ldots, \lambda_{m+2}$. Therefore it is uniquely determined by the Lagrange formula

$$1 - p^*(z) = \sum_{k=2}^{m+2} \rho e^{i\theta_k} l_k(z) \ .$$

In addition $1 - p^*(\lambda_1) = 1$ and this determines $\rho$ as the inverse of $\sum_{k=2}^{m+2} e^{i\theta_k} l_k(\lambda_1)$, yielding the relation (5.5). This establishes the following theorem.

Theorem 5.2. *There are $r$ eigenvalues in $\Omega = \Lambda(A)\backslash\{\lambda_1\}$, where $m+1 \le r \le 2m+1$, at which the optimal polynomial $1 - p^*(z) = 1 - (z - \lambda_1)q^*(z)$ reaches its maximum value. In addition, given any subset of $m+1$ among these $r$ eigenvalues, which can be labeled $\lambda_2, \lambda_3, \ldots, \lambda_{m+2}$, the polynomial can be represented by (5.5). In particular,*

$$\epsilon_1^{(m+1)} = \frac{1}{\sum_{k=2}^{m+2} e^{i\theta_k} \prod_{\substack{j=2 \\ j\neq k}}^{m+2} \frac{\lambda_1-\lambda_j}{\lambda_k-\lambda_j}} \ . \tag{5.7}$$

*Proof.* The result was proved above. Note that $\epsilon_1^{(m+1)}$ is equal to $\rho$ which is the inverse of the denominator in (5.5). □

In [11] it was shown that when $r = m+1$, then the sign $e^{i\theta_k}$ in the denominator of (5.7) becomes equal to the conjugate of the sign of $l_k(\lambda_1)$, which is the product term in the denominator of (5.7). In this case, (5.7) simplifies to

$$\epsilon_1^{(m+1)} = \frac{1}{\sum_{k=2}^{m+2} \prod_{\substack{j=2 \\ j\neq k}}^{m+2} \left|\frac{\lambda_1-\lambda_j}{\lambda_k-\lambda_j}\right|} \ . \tag{5.8}$$

The result in [11] was stated incorrectly for the general complex case, because the lemma on which it was based is only valid for real functions. Therefore, the result holds true only for the situation when the spectrum is real or when it is known that $r = m+1$ (e.g., when $N = m+1$). A proof of this result is given in Appendix 1.

**5.2. Optimization viewpoint.** An explicit formula for $\|(I - \mathcal{P}_m)u_1\|_2$ is obtained immediately from Lemma 3.1.

Corollary 5.3. *Let $L_{m+1}$ be the rectangular matrix of $\mathbb{C}^{N\times(m+1)}$, with column-vectors $\alpha_1\, u_1, v, A\, v, \ldots, A^{m-1}\, v$, then*

$$\|(I-\mathcal{P}_m)\,\alpha_1\, u_1\|^2 = \frac{1}{e_1^H\,(L_{m+1}^H L_{m+1})^{-1}\, e_1}. \tag{5.9}$$

*Proof.* The result follows by applying Lemma 3.1 with $w = \alpha_1\, u_1$, and $V = [v, Av, \cdots, A^{m-1}v]$. □

We now consider a factorization of the matrix $L_{m+1}$. If we set $\alpha$ to be the vector of $\mathbb{C}^m$ such that $v = U\,\alpha$, then we obtain $A^j\;\, v = U\,\Lambda^j\alpha$ for $j = 0, \ldots, m-1$. Then, we have

$$\begin{aligned} L_{m+1} &= [\alpha_1 u_1, v, A\, v, \ldots, A^{m-1}\, v] \\ &= U\,[\alpha_1\, e_1, \alpha, \Lambda\, \alpha, \ldots, \Lambda^{m-1}\, \alpha] \\ &= U\, D_\alpha\, W_{m+1}, \end{aligned}$$

with

$$D_\alpha = \begin{pmatrix} \alpha_1 & 0 & \ldots & 0 \\ 0 & \alpha_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \alpha_N \end{pmatrix} \text{ and } W_{m+1} = \begin{pmatrix} 1 & 1 & \lambda_1 & \ldots & \lambda_1^{m-1} \\ 0 & 1 & \lambda_2 & \ldots & \lambda_2^{m-1} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 1 & \lambda_i & \ldots & \lambda_i^{m-1} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 1 & \lambda_N & \ldots & \lambda_N^{m-1} \end{pmatrix}. \tag{5.10}$$

As a result,

$$L_{m+1}^H\, L_{m+1} = W_{m+1}^H\, D_\alpha^H\,(U^H U)\, D_\alpha\, W_{m+1}. \tag{5.11}$$

With theses formulas, it is now possible to express the residual norms of the Arnoldi process in terms of eigenvalues, eigenvectors, and the expression of $v$ in the eigenbasis.

Theorem 5.4.

$$\|(I-\mathcal{P}_m)\,\alpha_1\, u_1\|^2 = \frac{1}{e_1^H\,(W_{m+1}^H\, D_\alpha^H (U^H U) D_\alpha W_{m+1})^{-1}\, e_1}\;.$$

In the case of normal matrices ($A^H\, A = A\, A^H$), we have $U^H\, U = I$, so the preceding formula simplifies to

$$\|(I-\mathcal{P}_m)\,\alpha_1\, u_1\|^2 = \frac{1}{e_1^H\,(W_{m+1}^H\, D_\alpha^H D_\alpha W_{m+1})^{-1}\, e_1}. \tag{5.12}$$

It is interesting to consider first the particular situation when $m = 1$. If $m = 1$, and $A$ is normal, a little direct calculation yields,

$$\|(I-\mathcal{P}_1)\;\alpha_1 u_1\|^2 = \frac{|\alpha_1|^2 \sum\limits_{i=2}^{N} |\alpha_i|^2}{\sum\limits_{i=1}^{N} |\alpha_i|^2}.$$

Now, let $\beta$ the vector whose components are defined by

$$\beta_i = \frac{|\alpha_i|^2}{\sum_{j=1}^{N} |\alpha_j|^2}.$$

We then obtain, as a consequence of the above equality,

$$\|(I - \mathcal{P}_1)\ u_1\| = \sqrt{\sum_{i=2}^{N} \beta_i} = \sqrt{1 - \beta_1}.$$

Note that $\beta_1 = \cos^2\theta(u_1, v)$, where $\theta(u_1, v)$ is the angle between $v$ and $u_1$, and so the above relation can be expressed as $\|(I - \mathcal{P}_1)\ u_1\| = |\sin\theta(u_1, v)|$, which is another expression for the well-known relation $\|(I - \mathcal{P}_m)u_1\| = \sin\theta(u_1, K_m)$, for the case $m = 1$. Observe that the bound given in (5.1) does not exploit the orthogonality of the eigenvectors (for the normal case). Indeed, for the situation when $m = 1$, it yields

$$\|(I - \mathcal{P}_1)\ u_1\| \le \frac{\sum\limits_{i=2}^{N} |\alpha_i|}{|\alpha_1|}. \tag{5.13}$$

If we exploit this orthogonality in the proof of the result, we would obtain a slightly sharper bound in which the numerator in (5.13) is replaced by $\sqrt{\sum_{i=2}^{N} |\alpha_i|^2}$.

For the general case ($m \ge 1$), we will derive the optimal bound for

$$\frac{\|(I - \mathcal{P}_m)\ \alpha_1\, u_1\|^2}{\|\alpha\|^2} = \frac{1}{f_m(\beta)},$$

by solving the following optimization problem

$$\min_{\substack{\beta_1 \ge 0, \cdots\cdots, \beta_n \ge 0 \\ \sum\limits_{i=1}^{N} \beta_i = 1}} f_m(\beta),$$

where

$$f_m(\beta) = e_1^H \left(W_{m+1}^H\, D_\beta W_{m+1}\right)^{-1} e_1.$$

We now state the main result which gives an upper bound for the general situation.

Theorem 5.5. *If $m < N$ and $\|(I - \mathcal{P}_k)\ u_1\| \ne 0$ for $k \in \{1, \ldots, m\}$, then*

1. *If the matrix $A$ is normal then*

$$\|(I - \mathcal{P}_m)\ u_1\| \le \frac{\|\alpha\|}{|\alpha_1|} \eta_1^{(m)}\ . \tag{5.14}$$

2. *If the matrix $A$ is non normal but diagonalizable then*

$$\|(I - \mathcal{P}_m)\ u_1\| \le \frac{\sum\limits_{j=1}^{N} |\alpha_j|}{|\alpha_1|} \eta_1^{(m)}, \tag{5.15}$$

*where $\eta_1^{(m)}$ is defined as follows*

- *If all eigenvalues are real or if $m = N-1$, there exists $m$ eigenvalues, labeled $\lambda_2, \ldots, \lambda_{m+1}$ such that*

$$\eta_1^{(m)} = \frac{1}{1 + \sum\limits_{k=2}^{m+1} \prod\limits_{\substack{j=2 \\ j \neq k}}^{m+1} \frac{|\lambda_j - \lambda_1|}{|\lambda_j - \lambda_k|}}.$$

- *If at least one eigenvalue is non real and if $m < N-1$, there exist $m$ eigenvalues, labeled $\lambda_2, \ldots, \lambda_{m+1}$, and $m$ real numbers, $\theta_2, \ldots, \theta_{m+1}$ such that*

$$\eta_1^{(m)} = \frac{1}{1 - \sum\limits_{k=2}^{m+1} e^{\imath \theta_k} \prod\limits_{\substack{j=2 \\ j \neq k}}^{m+1} \frac{(\lambda_j - \lambda_1)}{(\lambda_j - \lambda_k)}} .$$

The proof is given in Appendix 2. We now make a few remarks. In order to compare the bound given in (5.1) and the bound given in Theorem 5.5, we need to unravel a relationship between $\epsilon_1^{(m)}$ and $\eta_1^{(m)}$. From the expression of $\epsilon_1^{(m)}$ we deduce that

$$\eta_1^{(m)} = \frac{\epsilon_1^{(m)}}{1 + \epsilon_1^{(m)}}.$$

Hence if the matrix is normal, the inequality (5.14) becomes

$$\|(I - \mathcal{P}_m)\ u_1\| \leq \frac{\|\alpha\|}{|\alpha_1|} \frac{\epsilon_1^{(m)}}{1 + \epsilon_1^{(m)}} \ . \tag{5.16}$$

A second remark is that if $m = N-1$ and $\lambda_k = (k-1)/(N-1), k = 1, \ldots, N$ (Uniform distribution) then (5.16) reduces to

$$\|(I - \mathcal{P}_{N-1})\ u_1\| \leq \frac{\sqrt{\sum_{i=1}^N |\alpha_i|^2}}{|\alpha_1|} \frac{1}{2^{N-1}},$$

while (5.1) becomes

$$\|(I - \mathcal{P}_{N-1})\ u_1\| \leq \frac{\sum_{i=2}^N |\alpha_i|}{|\alpha_1|} \frac{1}{2^{N-1} - 1},$$

If the components of $v$ in the eigen-decomposition are of equal size, i.e., $\alpha_i = w, \forall i \in \{1, \ldots, N\}$ then we obtain for normal matrices

$$\|(I - \mathcal{P}_{N-1})\ u_1\| \leq \frac{\sqrt{N}}{2^{N-1}},$$

while (5.1) becomes

$$\|(I - \mathcal{P}_{N-1})\ u_1\| \leq \frac{N-1}{2^{N-1} - 1}.$$

In general, for normal matrices the bound (5.16) is a slight refinement of (5.1). This comes from the fact that the bound (5.1) restricts the polynomials with the constraint $p(\lambda_1) = 1$ to obtain a simple result. There is no such restriction with the above result.

**Example 1.** We now examine an example to illustrate Theorem 5.5 in the complex case. We consider the following diagonal matrix, $A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2+\imath & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}$, and let $v = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^T$. We have $u_j = e_j$ for $j = 1, \ldots, 4$. Let us set $\lambda_1 = 1$ and consider the step $m = 2$. Using (5.12), we have

$$\|(I - \mathcal{P}_2)\, \alpha_1\, e_1\|^2 = \|\alpha\|^2 \, \frac{1}{f_2(\beta_1, \beta_2, \beta_3, \beta_4)},$$

where

$$f_2(\beta_1, \beta_2, \beta_3, \beta_4) = \frac{\beta_1\beta_2 + 2\beta_1\beta_3 + \beta_2\,\beta_3 + 4\beta_1\beta_4 + \beta_2\,\beta_4 + 2\beta_3\,\beta_4}{\beta_1(\beta_2\,\beta_3 + \beta_2\,\beta_4 + 2\beta_3\,\beta_4)}.$$

Solving the optimization problem

$$\min_{\substack{\beta_1 \geq 0, \beta_2 \geq 0, \beta_3 \geq 0, \beta_4 \geq 0 \\ \beta_1 + \beta_2 + \beta_3 + \beta_4 = 1}} f_2(\beta),$$

we obtain:

1. $f_2(\beta^*) = 6 + 2\sqrt{5} = (1 + \sqrt{5})^2$ and
$$\beta_1^* = \frac{\sqrt{5} - 1}{4}, \quad \beta_2^* = \frac{5 - \sqrt{5}}{10}, \quad \beta_3^* = \frac{5 - \sqrt{5}}{10}, \quad \beta_4^* = \frac{5 - \sqrt{5}}{20}\ .$$
2. The optimal bounds are
$$\frac{\|(I - \mathcal{P}_1)\, \alpha_1\, e_1\|}{\|\alpha\|} \leq \frac{1}{2} \quad \text{and} \quad \frac{\|(I - \mathcal{P}_2)\, \alpha_1\, e_1\|}{\|\alpha\|} \leq \frac{1}{1 + \sqrt{5}}.$$
3. The real numbers $\theta_2, \theta_3, \theta_4$ are such that
$$e^{\imath\,\theta_2} = -\frac{2 + \imath}{\sqrt{5}}, \quad e^{\imath\,\theta_3} = \frac{-1 + 2\,\imath}{\sqrt{5}}, \quad \text{and} \quad e^{\imath\,\theta_4} = \frac{1 - 2\,\imath}{\sqrt{5}}.$$

It is important to remark that the vector $\beta^*$ solution of the minimization problem is unique and all its components are non zero.

If all the eigenvalues are real and $m = 2$, we can show that there exists a solution vector with only three non zero components, as we will see in the proof of Theorem 5.5.

**6. Conclusion.** The analysis of convergence of the Arnoldi process for computing eigenvalues and eigenvectors is difficult and results in the non-normal case are bound to be limited in scope. This is essentially because the behavior of polynomials of $A$ or simply of the succesive powers $A^k$, can by itself be difficult to predict. All three forms of analysis covered in this paper (based on projectors, or the Schur form of $A$, or the diagonalization of $A$), confront this limitation. What is important is that some insight can be gained from each of these forms of analysis. The Schur form tells us that we can reduce the problem to that of estimating $\min p(B_1)\hat{v}$ where $B_1$ is a reduced Schur form of $A$, and $\hat{v}$ is the orthogonal projection of the initial vector $v$ onto the orthogonal of the eigenvector $u_1$. The analysis based on eigenvectors works essentially in the eigenbasis (assuming there is one) and converts the problem of estimating the error into a min-max problem. Here, there are three limitations. The first comes from the nature of Theorem 2.1 which expresses the residual for the projected problem of the exact eigenpair, not the other way around as desired. The second one comes from the fact that the bounds utilize the eigen-coefficients $\alpha_i$ of the initial vector in the eigenbasis. These coefficients can be very large in case the basis is ill conditioned. The last problem comes from the fact that the min-max quantities required by the bounds $(\epsilon_1^{(m)}, \eta_1^{(m)})$ can themselves very difficult to estimate.

## REFERENCES

[1] B. Beckermann. Image numérique, GMRES et polynomes de Faber. *C.R.A.S. (Proceedings of the French Academy of Sciences)*, 2006. In French.

[2] M. Crouzeix. Numerical range, holomorphic calculus, and applications, 2005. Manuscript.

[3] M. Eiermann. Fields of values and iterative methods. *Linear Algebra and its Applications*, 180:167–197, 1993.

[4] M. Eiermann and O. G. Ernst. Geometric aspects of the theory of krylov subspace methods. *Acta Numerica*, 10:251–312, 2001.

[5] M. Eiermann, O. G. Ernst, and O. Schneider. Analysis of acceleration strategies for restarted minimal residual methods. *J. Comput. Appl. Math*, 123:345–357, 2000.

[6] R. Fletcher. *Practical methods of optimisation*. Wiley, Chichester, 2nd edition, 1987.

[7] I. C. F. Ipsen. Expressions and bounds for the gmres residuals. *BIT*, 40:524–535, 2000.

[8] P. Lancaster and M. Tismenetsky. *The Theory of Matrices*. Academic Press, New York, second edition, 1985.

[9] G. Meurant. *Computer solution of large linear systems*. North-Holland, Amsterdam, 1999. Vol 28, Studies in Mathematics and its Applications.

[10] T. J. Rivlin. *The Chebyshev Polynomials: from Approximation Theory to Algebra and Number Theory*. J. Wiley and Sons, New York, 1990.

[11] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Halstead Press, New York, 1992.

[12] Y. Saad. Further analysis of minimal residual iterations. *Numerical Linear Algebra with Applications*, 7:67–93, 2000.

[13] Y. Saad. *Iterative Methods for Sparse Linear Systems, 2nd edition*. SIAM, Philadelpha, PA, 2003.

[14] H. Sadok. Analysis of the convergence of the minimal and the orthogonal residual methods. *Numer. Algorithms*, 40:111–115, 2005.

[15] G. W. Stewart. *Matrix Algorithms II: Eigensystems*. SIAM, Philadelphia, 2001.

[16] F. Zhang. *Matrix Theory*. Springer Verlag, New York, 1999.

**Appendix 1.** We denote by $\mathbb{P}_m^*$ the set of polynomials $p$ of degree $\leq m$ such that $p(\lambda_1) = 0$. We seek the best uniform approximation of the function $f(z) = 1$ by polynomials of degree $\leq m$ in $\mathbb{P}_m^*$. Note that $\mathbb{P}_m^*$ is of dimension $m$. Let the set of $r$ extremal points be $\lambda_2, \ldots, \lambda_{r+1}$ (See theorem 5.1). According to Theorem 5.2, given any subset of $m+1$ among these $r$ extremal points, which we label $\lambda_2, \lambda_3, \ldots, \lambda_{m+2}$, the best polynomial can be represented by (5.5) in which $e^{i\theta_k} = sign(1 - p^*(\lambda_k))$.

Not much can be said from this result in the general situation. However, when $r = m+1$, then we can determine $\max |1 - p^*(z)|$. In this situation the necessary and sufficient conditions of Theorem 5.1 express the extremal points as follows. Let us set $\xi_j \equiv \mu_j \overline{[f(z_j) - p^*(z_j)]}$ for $j = 1, \cdots m+1$, and select any basis $\phi_1, \cdots, \phi_m$ of the polynomial subspace $\mathbb{P}_m^*$. Then, the condition (5.4) translates to

$$\sum_{k=1}^{m+1} \xi_k \phi_j(\lambda_k) = 0 \quad \text{for} \quad j = 1, \ldots, m. \tag{6.1}$$

The above equations constitute an underdetermined system of linear equations with the unknowns $\xi_k$. In fact, since the $\xi_k$'s are all nonzero, we can fix any one component, and the rest will then be determined uniquely. This is best done in a more convenient *basis* of polynomials given by:

$$\omega_j(z) = (z - \lambda_1)\hat{l}_j(z), \quad j = 2, \ldots, m+1, \tag{6.2}$$

where $\hat{l}_j$ is the Lagrange polynomial of degree $m-1$,

$$\hat{l}_j(z) = \prod_{\substack{k=2 \\ k \neq j}}^{m+1} \frac{z - \lambda_k}{\lambda_j - \lambda_k}, \quad j = 2, \ldots, m+1. \tag{6.3}$$

With this we can prove the following lemma.

Lemma 6.1. *The underdetermined linear system of $m$ equations and $m+1$ unknowns $\xi_k, k = 2, \ldots, m+2$*

$$\sum_{k=2}^{m+2} \omega_j(\lambda_k)\xi_k = 0, \quad j = 2, 3, \ldots, m+1 \tag{6.4}$$

*admits the nontrivial solution*

$$\xi_k = \prod_{\substack{j=2 \\ j \neq k}}^{m+2} \frac{\lambda_1 - \lambda_j}{\lambda_k - \lambda_j}, \quad k = 2, \ldots, m+2. \tag{6.5}$$

*Proof.* The proof requires a straightforward algebraic verification; see [11] for details. □

We can now prove the main result of this appendix.

Theorem 6.2. *Let $p^*$ be the (unique) polynomial of degree $m$ satisfying the constraint $p(\lambda_1) = 0$, and which is the best uniform approximation to the function $f(z) = 1$ on a compact set $\Omega$ consisting of at least $m+1$ points. Assume that there are $m+1$ extremal points labeled $\lambda_2, \ldots, \lambda_{m+2}$ and let $\xi_k, k = 2, \ldots, m+2$ be any solution of the linear system (6.4). Write each $\xi_k$ in the form $\xi_k = \delta_k e^{-i\theta_k}$ where $\delta_k$*

*is real and positive and $\theta_k$ is real. Then, $p^*$ can be expressed as*

$$1 - p^*(z) = \frac{\sum\limits_{k=2}^{m+2} e^{i\theta_k} l_k(z)}{\sum\limits_{k=2}^{m+2} |l_k(\lambda_1)|}, \tag{6.6}$$

*where $l_k$ is the Lagrange polynomial of degree $m$*

$$l_k(z) = \prod_{\substack{j=2 \\ j \neq k}}^{m+2} \frac{z - \lambda_j}{\lambda_k - \lambda_j}.$$

*As a consequence,*

$$\epsilon_1^{(m+1)} = \left( \sum_{j=2}^{m+1} \prod_{k=2, k\neq j}^{m+1} \frac{|\lambda_k - \lambda_1|}{|\lambda_k - \lambda_j|} \right)^{-1} . \tag{6.7}$$

*Proof.* Equation (5.4) states that

$$1 - p^*(z) = \rho \sum_{k=2}^{m+2} e^{i\theta_k} l_k(z) \quad \text{with} \quad \rho = \frac{1}{\sum_{k=2}^{m+2} e^{i\theta_k} l_k(\lambda_1)} . \tag{6.8}$$

We now apply Theorem 5.1 which states that at the extremal points (now known to be unique) there are $m+1$ positive coefficients $\mu_j$ such that

$$\xi_k \equiv \mu_k \overline{[1 - p^*(\lambda_k)]} = \rho \mu_k e^{-i\theta_k} \tag{6.9}$$

satisfy the system (6.1). As was already mentioned, the solution to (6.1) is uniquely determined if we set any one of its components. Set $\xi_{m+2} = l_{m+2}(\lambda_1)$. Then, according to Lemma 6.1, we must have $\xi_k = l_k(\lambda_1)$, for $k = 2, \ldots, m+2$. Since $\rho$ and $\mu_k$ are positive, (6.9) shows that

$$e^{-i\theta_k} = sign(l_k(\lambda_1)) \quad \rightarrow \quad e^{i\theta_k} = \frac{\overline{l_k(\lambda_1)}}{|l_k(\lambda_1)|} \rightarrow \quad \rho = \frac{1}{\sum_{k=2}^{m+2} |l_k(\lambda_1)|}.$$

The result (6.7) is an expanded version of the above expression for $\rho$. □

**Appendix 2: Proof of Theorem 5.5.** We will consider the 2 cases, $A$ normal and $A$ non-normal, separately. In this appendix, the notation for the 2-norm is changed to $\|.\|$.

**Case 1: $A$ is normal.** To prove the first part of this Theorem, we need two lemmas.

LEMMA 6.3. *Let $\omega_1, \omega_2, \ldots, \omega_m$ be $m$ distinct complex numbers and $V$ the $m \times m$ Vandermonde matrix whose entries $V_{i,j}$ are given by :*

$$v\ (i,j) = \omega_j^{i-1}\ , \quad i,j = 1, \ldots, m.$$

*and let $y = (1, \rho, \rho^2, \ldots, \rho^{m-1})^T$. Then, the dual Vandermonde system $V^T\ x = y$ admits the unique solution $x = (x_1, x_2, \ldots, x_m)^T$, where*

$$x_i = \prod_{\substack{1 \le j \le m \\ j \ne i}} \left( \frac{\rho - \omega_j}{\omega_i - \omega_j} \right).$$

The proof is obvious.

Recall that $\beta$ is the vector whose components are $\beta_i = \dfrac{|\alpha_i|^2}{\sum_{j=1}^N\ |\alpha_j|^2}$ and define the matrix

$$Z_{m+1}(\beta) = W_{m+1}^H D_\beta W_{m+1} \in \mathbb{C}^{m+1,m+1},$$

where $D_\beta$ and $W_{m+1}$ have been defined in (5.10). If the matrix function $Z_{m+1}(\beta)$ is nonsingular, we define the functions $f_m$ by

$$f_m(\beta) = e_1^H\ Z_{m+1}^{-1}(\beta)\ e_1,$$

with

$$\beta \in S = \left\{ \beta = (\beta_1, \cdots\cdots, \beta_N) \in [0, \quad 1]^N\ \ /\ \sum_{i=1}^N \beta_i = 1 \right\}.$$

Then we have the following result.

LEMMA 6.4. *If the matrix function $Z_{m+1}(\beta)$ is nonsingular, then the following properties hold :*

1. *$f_m$ is homogeneous of degree $-1$, i.e., $f(r\beta) = r^{-1}\ f(\beta)$ for $r > 0$.*
2. *$f_m$ is a convex function defined on the closed convex set $S$.*
3. *$f_m$ is differentiable at $\beta \in S$ and we have*

$$\frac{\partial f_m}{\partial \beta_i}(\beta)\ = -\,|e_i^H\, W_{m+1}\, t|^2,$$

*where $t = (t_1,\ t_2, \ldots\ldots,\ t_{m+1})^T$ is such that $Z_{m+1}(\beta)\ t =\ e_1$.*

*Moreover, we have $\sum_{i=1}^N\ \beta_i \dfrac{\partial f_m(\beta)}{\partial \beta_i} = -f_m(\beta)$.*

*Proof.* The proof will proceed in 3 steps.

1. Let $r$ be some positive real. Since $D_{r\,\beta} = r\,D_\beta$, then

$$f_m(r\beta) = e_1^H\ (r\,(W_{m+1}^H D_\beta W_{m+1}))^{-1}\, e_1 = r^{-1}\, f_m(\beta).$$

2. It is easy to verify that the set $S$ is convex. By taking $x = e_1,\ G_1 = Z_{m+1}(r\beta)$ and $G_2 = Z_{m+1}((1-r)\beta')$ where $\beta, \beta' \in S$ and $0 < r < 1$ in the inequality

$$x^H \left( rG_1 + (1-r)\, G_2 \right)^{-1} x \leq x^H r \left(G_1\right)^{-1} x \ + x^H \left(1-r\right) \left(G_2\right)^{-1} x$$

where $G_1$ and $G_2$ are positive definite hermitian matrices [16, p. 174], we obtain

$$f_m(r\beta + (1-r)\beta') \leq r\, f_m(\beta) \ + (1-r)\, f_m(\beta').$$

Hence $f_m$ is convex.
3. Clearly, $f_m$ is differentiable at $\beta \in S$. By using the derivative of the inverse of the matrix function, we have

$$\frac{\partial Z_{m+1}^{-1}(\beta)}{\partial \beta_i} = -Z_{m+1}^{-1}(\beta)\ \frac{\partial Z_{m+1}(\beta)}{\partial \beta_i}\ Z_{m+1}^{-1}(\beta).$$

It follows that

$$\frac{\partial Z_{m+1}^{-1}(\beta)}{\partial \beta_i} = -Z_{m+1}^{-1}(\beta)\ W_{m+1}^H\ E_i\ W_{m+1}\ Z_{m+1}^{-1}(\beta)$$

where $E_i = e_i e_i^T \in \mathbb{R}^{N,N}$. Therefore,

$$\frac{\partial f_m}{\partial \beta_i}(\beta) = -|e_i^T\ W_{m+1}\ t|^2.$$

The last equality follows from a simple algebraic manipulation, noting that

$$\sum_{i=1}^{N} \beta_i \frac{\partial f_m(\beta)}{\partial \beta_i} = -\sum_{i=1}^{N} t^H W_{m+1}^H (\beta_i e_i e_i^H) W_{m+1}\, t = -t^H W_{m+1}^H D_\beta W_{m+1}\, t = -f(\beta).$$

□

*Proof of Theorem 5.5..* From (5.12) we deduce that

$$\|(I - \mathcal{P}_m)\, \alpha_1\, u_1\|^2 = \|\alpha\|^2 \frac{1}{e_1^H\, (W_{m+1}^H\, D_\beta W_{m+1})^{-1}\, e_1}.$$

Since $\beta_i \geq 0$ and $\sum_{i=1}^N\ \beta_i = 1$, in order to get an optimal bound of

$$\frac{\|(I - \mathcal{P}_m)\, \alpha_1\, u_1\|^2}{\|\alpha\|^2} = \frac{1}{f_m(\beta)},$$

we must solve the following optimization problem

$$\min_{\substack{\beta_1 \geq 0, \cdots\cdots, \beta_N \geq 0 \\ \sum\limits_{i=1}^{N} \beta_i = 1}} \ f_m(\beta). \tag{6.10}$$

We introduce the Lagrangian function for this problem :

$$\mathcal{L}_m(\beta, \delta, \mu) = f_m(\beta) - \delta\ \left(1 - \sum_{i=1}^{N} \beta_i\right) - \sum_{i=1}^{N} \mu_i \beta_i,$$

where $\delta \in \mathbb{R}$ ; $\mu = (\mu_1, \cdots\cdots, \mu_N) \in \mathbb{R}^N$. According to the Karush-Kuhn-Tucker (KKT) conditions, if $f$ has a local minimizer $\beta^*$ in $S$, then there exist Lagrangian multipliers $\delta^*, \mu^* = (\mu_1^*, \cdots\cdots, \mu_N^*)$, such that $(\beta^*,\ \delta^*, \mu^*)$, satisfy the following conditions:

(i) $\dfrac{\partial \mathcal{L}_m}{\partial \beta_i}( \beta^*, \delta^*, \mu^*) = 0$ , for $i = 1, \ldots, N$,

(ii) $(1 - \sum\limits_{i=1}^{N} \beta_i^*) = 0$ and $\beta_i^* \geq 0$, for $i = 1, \ldots, N$.

(iii) $\mu_i^* \beta_i^* = 0$ for $i = 1, \ldots, N$,

(iv) $\mu_i^* \geq 0$ for $i = 1, \ldots, N$.

We note that in this case the KKT conditions are also sufficient since the problem is convex. Condition (i) and Lemma 6.4 give

$$\frac{\partial \mathcal{L}_m}{\partial \beta_i}(\beta^*, \delta^*, \mu^*) = -\,|e_i^H\, W_{m+1}\, t^*|^2\; + \delta^* - \mu_i^* = 0, \tag{6.11}$$

where

$$t^* = Z_{m+1}^{-1}(\beta^*)\ e_1.$$

From (iii), it follows that for $i = 1, \ldots, N$, either $\mu_i^* = 0$ or $\beta_i^* = 0$. It can be shown that $\beta_1^* \neq 0$. Indeed, notice that the best approximation to $u_1$ from $\mathbb{K}_m$ is $\mathcal{P}_m u_1 = K(K^T K)^{-1}(K^T u_1)$ and its first component $\beta_1 = u_1^T K(K^T K)^{-1}(K^T u_1)$ is clearly zero iff $K^T u_1 = 0$. So if $\alpha_1 \neq 0$ then $\beta_1^* \neq 0$.

We also notice that the vector $\beta^* = (\beta_1^*, \ldots, \beta_N^*)$ has exactly $m + s$ non zeros components which will be labeled $\beta_1^*, \ldots, \beta_{m+s}^*$, with $s \geq 1$. It follows that $\beta^* = (\beta_1^*, \ldots, \beta_{m+s}^*, 0, \ldots, 0)$ and $\mu_i^* = 0$ for $i = 1, \ldots, m + s$. Substituting in (6.11) it follows that

$$|e_i^H\, W_{m+1}\, t^*|^2 \;=\; \delta^*, \qquad \text{for } i = 1, \ldots, m + s, \tag{6.12}$$

where $t^* = (t_1^*, \ldots, t_m^*)^T = (W_{m+1}^H\, D_{\beta^*}\, W_{m+1})^{-1}\, e_1$. Hence, we have

$$e_i^H\, W_{m+1}\, t^* \;=\; e^{\imath\theta_i}\, \sqrt{\delta^*}, \quad for \quad i = 1, \ldots, m + s, \tag{6.13}$$

where $\imath = \sqrt{-1}$ and $\theta_i \in \mathbb{R}$. We then obtain the following linear system

$$\left\{ \begin{array}{rcl} t_1^* + t_2^* + \lambda_1\; t_3^* + \ldots\ldots + \lambda_1^{m-1}\; t_{m+1}^* & = & e^{\imath\theta_1}\,\sqrt{\delta^*} \\ t_2^* + \lambda_2\; t_3^* + \ldots\ldots + \lambda_2^{m-1}\; t_{m+1}^* & = & e^{\imath\theta_2}\,\sqrt{\delta^*} \\ \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \ldots\ldots \\ \\ t_2^* + \lambda_{m+s}\; t_3^* + \ldots\ldots + \lambda_{m+s}^{m-1}\; t_{m+1}^* & = & e^{\imath\theta_{m+s}}\,\sqrt{\delta^*}. \end{array} \right. \tag{6.14}$$

From Lemma 6.4 , we can easily show that $t_1^* = \delta^*$

$$\sqrt{f_m(\beta^*)} = \sqrt{\delta^*} = \frac{\begin{vmatrix} e^{\imath\theta_1} & 1 & \lambda_1 & \ldots\ldots\ldots & \lambda_1^{m-1} \\ e^{\imath\theta_2} & 1 & \lambda_2 & \ldots\ldots\ldots & \lambda_2^{m-1} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ e^{\imath\theta_{m+1}} & 1 & \lambda_{m+1} & \ldots\ldots\ldots & \lambda_{m+1}^{m-1} \end{vmatrix}}{\begin{vmatrix} 1 & \lambda_2 & \ldots\ldots\ldots & \lambda_2^{m-1} \\ 1 & \lambda_3 & \ldots\ldots\ldots & \lambda_3^{m-1} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & \lambda_{m+1} & \ldots\ldots\ldots & \lambda_{m+1}^{m-1} \end{vmatrix}}. \tag{6.15}$$

Now, since $W_{m+1}^H \, D_{\beta^*} W_{m+1} t^* = e_1$, we deduce that

$$e^{\imath\theta_1} \, \beta_1^* = \frac{1}{\sqrt{\delta^*}}, \quad \text{and} \quad \sum_{i=1}^{m+s} \overline{\lambda_i^{j-1}} \, e^{\imath\theta_i} \, \beta_i^* = 0, \text{ for } j = 1, \ldots, m. \tag{6.16}$$

Since $\beta_1^*$ is a positive real number, we obviously have: $\beta_1^* = \dfrac{1}{\sqrt{\delta^*}}$ and $e^{\imath\theta_1} = 1$. Hence, expanding the numerator of (6.15) with respect to its first column, gives

$$\sqrt{\delta^*} = 1 - \sum_{k=2}^{m+1} e^{\imath\theta_k} \prod_{\substack{j=2 \\ j \neq k}}^{m+1} \frac{(\lambda_j - \lambda_1)}{(\lambda_j - \lambda_k)} = 1 - \sum_{k=2}^{m+1} e^{\imath\theta_k} \, \mathrm{l}_k(\lambda_1), \tag{6.17}$$

where $\mathrm{l}_k(\omega) = \displaystyle\prod_{\substack{j=2 \\ j \neq k}}^{m+1} \frac{(\lambda_j - \omega)}{(\lambda_j - \lambda_k)}$. It is obvious that $\theta_2, \ldots, \theta_{m+1}$ are such that $1 - \sum_{k=2}^{m+1} e^{\imath\theta_k} \, \mathrm{l}_k(\lambda_1)$ is real. The vector $(e^{\imath\theta_2} \, \beta_2^*, \ldots, e^{\imath\theta_{m+s}} \, \beta_{m+s}^*)^T$ satisfies (6.16), which is equivalent to the following linear system

$$\left\{ \begin{array}{l} e^{-\imath\theta_2} \, \beta_2^* + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + e^{-\imath\theta_{m+s}} \, \beta_{m+s}^* = -\dfrac{1}{\sqrt{\delta^*}}, \\ \lambda_2 \, e^{-\imath\theta_2} \, \beta_2^* + \ldots\ldots\ldots\ldots\ldots + \lambda_{m+s} \, e^{-\imath\theta_{m+s}} \, \beta_{m+s}^* = -\dfrac{\lambda_1}{\sqrt{\delta^*}}, \\ \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \lambda_2^{m-1} \, e^{-\imath\theta_2} \, \beta_2^* + \ldots\ldots\ldots\ldots + \lambda_{m+s}^{m-1} \, e^{-\imath\theta_{m+s}} \, \beta_{m+s}^* = -\dfrac{\lambda_1^{m-1}}{\sqrt{\delta^*}}. \end{array} \right. \tag{6.18}$$

To obtain $\beta_l^*$ and $e^{\imath\theta_l}$ for $l = 2, \ldots, m+s$, we have to solve (6.18). Let us derive the Lagrange multipliers $\mu_l^*$. We have $\mu_l = 0$, for $l = 1, \ldots, m+s$ and the other components are deduced from (6.11):

$$\mu_l^* = \delta^* - |t_2^* + \lambda_l \, t_3^* + \ldots\ldots + \lambda_l^{m-1} \, t_{m+1}^*|^2 \quad \text{for} \quad l = m+s+1, \ldots, N.$$

In view of (6.14), and using Lemma 6.3, we obtain

$$\mu_l^* = \delta^* \left( 1 - \left| \sum_{k=2}^{m+1} e^{\imath\theta_k} \prod_{\substack{j=2 \\ j \neq k}}^{m+1} \frac{(\lambda_j - \lambda_l)}{(\lambda_j - \lambda_k)} \right|^2 \right) \geq 0 \quad \text{for} \quad l = m+s+1, \ldots, N.$$

Hence the eigenvalues $\lambda_2, \ldots, \lambda_{m+1}$ and the scalars $\theta_2, \ldots, \theta_{m+1}$ must be chosen such that

$$\left| \sum_{k=2}^{m+1} e^{\imath\theta_k} \, \mathrm{l}_k(\lambda_l) \right| \leq 1, \text{ for } \; l = m+s+1, \ldots, N.$$

We have $\sqrt{\delta^*} = p_m(\lambda_1)$, where $p_m$ is the complex valued polynomial of degree $m-1$ satisfying the following conditions :

$$\left\{ \begin{array}{l} p_m(\omega) = 1 - \sum_{k=2}^{m+1} \, e^{\imath\theta_k} \, \mathrm{l}_k(\omega), \\ p_m(\lambda_1) = \sqrt{\delta^*} \geq 2, \\ p_m(\lambda_j) = 1 - e^{\imath\theta_j} \quad \text{for} \quad j = 2, \ldots, m+s, \\ |1 - p_m(\lambda_j)| \leq 1 \quad \text{for} \quad j = m+s+1, \ldots, N. \end{array} \right. \tag{C}$$

To obtain the vector solution of the optimization problem we need to solve the linear system (6.18). This can readily be solved by Cramer's rule if for example $s = 1$. If $s > 1$ and all the eigenvalues are real the system (6.18) will have infinitely many solutions. But if one of the eigenvalues $\lambda_l$ is complex, this system will have a unique solution.

We will now investigate the possible values that $s$ can take. There are 3 distinct possibilities.

1) Case $s = 1$. This means that the system (6.18) has a solution with exactly $m+1$ non zeros elements. Using Lemma 6.3 for solving (6.18), we obtain

$$e^{-\imath\theta_l}\,\beta_l^* = -\frac{\mathrm{l}_l(\lambda_1)}{\sqrt{\delta^*}} \quad for \quad l = 2, \ldots, m+1.$$

We can deduce $\beta_l^*$ and $e^{\imath\theta_l}$, by noticing that $\beta_l^* \geq 0$. So we get for $l = 2, \ldots, m+1$

$$\beta_l^* = \frac{|\mathrm{l}_l(\lambda_1)|}{\sqrt{\delta^*}} \quad and \quad e^{\imath\theta_l} = -\frac{\overline{\mathrm{l}_l(\lambda_1)}}{|\mathrm{l}_l(\lambda_1)|}. \tag{6.19}$$

Moreover $\beta_1^* = \dfrac{1}{\sqrt{\delta^*}}$ and $\beta_l^* = 0$, for $l = m+2, \ldots, N$. Now, since $\displaystyle\sum_{l=1}^{N} \beta_l^* = 1$, then

$$\sqrt{\delta^*} = 1 + \sum_{l=2}^{m+1} \prod_{\substack{j=2\\ j\neq l}}^{m+1} \frac{|\lambda_j - \lambda_1|}{|\lambda_j - \lambda_l|}. \tag{6.20}$$

Since $\delta^*$ is the minimum of the function $f_m(\beta)$, we deduce that $\lambda_2, \ldots, \lambda_{m+1}$ are such that

$$\sum_{l=2}^{m+1} \prod_{\substack{j=2\\ j\neq l}}^{m+1} \frac{|\lambda_j - \lambda_1|}{|\lambda_j - \lambda_l|} = \min_{\lambda_{i_1}, \ldots, \lambda_{i_{m+1}}} \sum_{l=2}^{m+1} \prod_{\substack{j=2\\ j\neq l}}^{m+1} \frac{|\lambda_{i_j} - \lambda_1|}{|\lambda_{i_j} - \lambda_{i_l}|}.$$

Since $s = 1$, the minimum is attained only for the set $\{\lambda_2, \ldots, \lambda_{m+1}\}$ which will be called the extremal set.

2) Case when $s > 1$ and all the eigenvalues are real. In this case $e^{\imath\theta_k} = \pm 1$. Let us set $\sigma_j = \zeta_j\,\beta_j^*$, where $\zeta_j = \pm 1$. the system (6.18) can be written as

$$\left\{\begin{array}{c} \sigma_2 + \ldots + \sigma_{m+1} = -\dfrac{1}{\sqrt{\delta^*}} - \sigma_{m+2} \ldots - \sigma_{m+s}, \\ \lambda_2\sigma_2 + \ldots + \lambda_{m+1}\sigma_{m+1} = -\dfrac{\lambda_1}{\sqrt{\delta^*}} - \lambda_{m+2}\sigma_{m+2} \ldots - \lambda_{m+s}\sigma_{m+s}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \lambda_2^{m-1}\sigma_2 + \ldots + \lambda_{m+1}^{m-1}\sigma_{m+1} = -\dfrac{\lambda_1^{m-1}}{\sqrt{\delta^*}} - \lambda_{m+2}^{m-1}\sigma_{m+2} \ldots - \lambda_{m+s}^{m-1}\sigma_{m+s}. \end{array}\right. \tag{6.21}$$

Using lemma 6.3, we obtain for $l = 2, \ldots, m+1$

$$\zeta_l\,\beta_l^* = \sigma_l = -\frac{1}{\sqrt{\delta^*}}\,\mathrm{l}_l(\lambda_1) - \sigma_{m+2}\,\mathrm{l}_l(\lambda_{m+2}) - \ldots\ldots - \sigma_{m+s}\,\mathrm{l}_l(\lambda_{m+s}).$$

On the other hand, the conditions (C) can be written in this case

$$\left\{\begin{array}{l} p_m(\omega) = 1 - \sum_{k=2}^{m+1} \zeta_k \, \mathrm{l}_k(\omega), \\ p_m(\lambda_1) \geq 2, \\ p_m(\lambda_j) = 2 \quad \text{for } j \in \Theta_- = \{j \in \{2,\ldots,m+s\},\ \zeta_j = -1\} \\ p_m(\lambda_j) = 0 \quad \text{for } j \in \Theta_+ = \{j \in \{2,\ldots,m+s\},\ \zeta_j = 1\} \\ p_m(\lambda_j) \in [0,2] \quad \text{for} \quad j = m+s+1,\ldots,N. \end{array}\right. \tag{C1}$$

First, we dispose of the case $m = 2$. If $m = 2$, conditions (C1) imply that $s = 1$ or $p_m \equiv 2$. We assume now that $m \geq 3$. In this case the polynomial $p_m$ is not constant and we have $p_m(0) > 2$. The sets $\Theta_-$ and $\Theta_-$ contain at least one element. Otherwise $p_m \equiv 0$ or $p_m \equiv 2$, which is impossible since $p_m(\lambda_1) > 2$. Moreover, since $\{2,\ldots,m+s\} = \Theta_+ \bigcup \Theta_-$, the polynomials $p_m$ and $2 - p_m$ are of degree $m-1$ and will have at most $m-1$ zero. Therefore, the number of elements of $\Theta_+$ and $\Theta_-$ is less or equal than $m-1$. Hence $s \leq m-1$. The polynomial $p'_m$ is of degree $m-2$ and will have $m-2$ zeros. If we assume that $\lambda_2 < \lambda_3 < \ldots < \lambda_{m+1}$, there exists an extremal set of eigenvalues $\{\lambda_2,\ldots,\lambda_{m+1}\}$ with alternating sign, i.e., $\zeta_j = (-1)^{j-1}\zeta_2$. So in the real case (all the eigenvalues are real), there exists a solution $\beta^*$ of the system (6.21), with only $m+1$ non zeros components:

$$\beta^*_l = \frac{|\mathrm{l}_l(\lambda_1)|}{\sqrt{\delta^*}} \quad \text{for} \quad l = 2,\ldots,m+1.$$

Invoking the fact that $\sum_{k=2}^{N} \beta^*_k = 1$ we obtain

$$\sqrt{\delta^*} = 1 + \sum_{l=2}^{m+1} |\mathrm{l}_l(\lambda_1)|.$$

And the extremal $\{\lambda_2,\ldots,\lambda_{m+1}\}$ is such that

$$\sum_{l=2}^{m+1} \prod_{\substack{j=2 \\ j\neq l}}^{m+1} \frac{|\lambda_j - \lambda_1|}{|\lambda_j - \lambda_l|} = \min_{\lambda_{i_1},\ldots,\lambda_{i_{m+1}}} \sum_{l=2}^{m+1} \prod_{\substack{j=2 \\ j\neq l}}^{m+1} \frac{\left|\lambda_{i_j} - \lambda_1\right|}{\left|\lambda_{i_j} - \lambda_{i_l}\right|}.$$

This extremal set is not the unique set verifying the preceding property.

3) Case when $s > 1$ and one of the eigenvalues is complex. Here the situation is more complicated. If we consider the case $m = 2$. The conditions (C) now give

$$\left\{\begin{array}{l} p_2(\omega) = 1 - e^{\imath\theta_2} \, \mathrm{l}_2(\omega) - e^{\imath\theta_3} \, \mathrm{l}_3(\omega), \\ p_2(\lambda_1) = \sqrt{\delta^*} \geq 2, \\ p_2(\lambda_j) = 1 - e^{\imath\theta_j} \quad \text{for} \quad j = 2,\ldots,2+s, \\ |1 - p_2(\lambda_j)| \leq 1 \quad \text{for} \quad j = s+3,\ldots,N. \end{array}\right. \tag{C2}$$

We can have $s = 2$ (see the example). And it is not easy to explicit in general the real numbers $\theta_2,\ldots,\theta_{m+s}$. We can only conclude that if there exists a solution of the optimization problem (6.10) with only $m+1$ non zeros components then the optimal solution is given by (6.20). Otherwise the optimal solution satisfies (6.17).

**Case 2: $A$ is non normal but diagonalizable.** Let $\Delta$ be the diagonal matrix defined by

$$\Delta = \frac{1}{\sqrt{\sum_{j=1}^{N} |\alpha_j|}} \left( \begin{array}{cccc} \sqrt{|\alpha_1|} & 0 & \ldots & 0 \\ 0 & \sqrt{|\alpha_2|} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \sqrt{|\alpha_N|} \end{array} \right).$$

By Theorem 5.4, we may write

$$\|(I - \mathcal{P}_m)\, \alpha_1\, u_1\|^2 = \frac{1}{e_1^H \left(W_{m+1}^H\, \Delta\, \Delta^{-1} D_\alpha^H\, U^H U\, D_\alpha \Delta^{-1}\, \Delta W_{m+1}\right)^{-1} e_1}.$$

Using the Courant-Ficher theorem, we obtain

$$\|(I - \mathcal{P}_m)\, \alpha_1\, u_1\|^2 \leq \frac{\|U\, D_\alpha \Delta^{-1}\|^2}{e_1^H \left(W_{m+1}^H\, \Delta^2 W_{m+1}\right)^{-1} e_1}.$$

Moreover, for all vector $x = (x_1, \ldots, x_N)^T$, we have

$$\|U\, D_\alpha \Delta^{-1} x\| = \left(\sqrt{\sum_{j=1}^{N} |\alpha_j|}\right)\, \|\sum_{j=1}^{N} \frac{\alpha_j}{\sqrt{|\alpha_j|}}\, x_j u_j\|.$$

Since $\|u_i\| = 1$, we have by the triangle inequality

$$\|U\, D_\alpha \Delta^{-1} x\| \leq \left(\sqrt{\sum_{j=1}^{N} |\alpha_j|}\right) \sum_{j=1}^{N} \frac{|\alpha_j|}{\sqrt{|\alpha_j|}}\, |x_j| = \left(\sqrt{\sum_{j=1}^{N} |\alpha_j|}\right) \sum_{j=1}^{N} \sqrt{|\alpha_j|}\, |x_j|.$$

The Cauchy-Schwarz inequality implies that $\|U\, D_\alpha \Delta^{-1}\| \leq \left(\sum_{j=1}^{N} |\alpha_j|\right)$, since

$$\|U\, D_\alpha \Delta^{-1} x\| \leq \left(\sum_{j=1}^{N} |\alpha_j|\right)\, \|x\|.$$

We thus have

$$\|(I - \mathcal{P}_m)\, \alpha_1\, u_1\|^2 \leq \frac{\left(\sum_{j=1}^{N} |\alpha_j|\right)^2}{f_m(\rho_1, \ldots, \rho_N)}, \quad \text{where} \quad \rho_j = \frac{|\alpha_j|}{\sum_{k=1}^{N} |\alpha_k|}.$$

Since $\rho_j \geq 0$ and $\sum_{k=1}^{N} \rho_k = 1$, we conclude that

$$\|(I - \mathcal{P}_m)\, \alpha_1\, u_1\|^2 \leq \frac{\left(\sum_{j=1}^{N} |\alpha_j|\right)^2}{\min_{\substack{\rho_1 \geq 0, \ldots, \rho_N \geq 0 \\ \rho_1 + \ldots + \rho_N = 1}} f_m(\rho_1, \ldots, \rho_N)} = \frac{\left(\sum_{j=1}^{N} |\alpha_j|\right)^2}{\delta^*}.$$

This completes the proof by denoting $\eta_1^{(k)} = \dfrac{1}{\sqrt{\delta^*}}$. Let us consider the following examples for illustrating the conditions (C1) given in the preceding proof

*Example 2.* If $\lambda_k = k, k = 1, \ldots, N$.

1. If $N$ is even ( $N = 2n > 4$).
   - If $m = 2$, then $s = 1$ and the extremal set is unique and is $\{2, N\}$, the polynomial verifying the conditions (C1) is $p_2(w) = 2 - 2\dfrac{w-2}{N-2}$ and we have $p_2(1) = \dfrac{2n-1}{n-1}$, $\beta_1^* = \dfrac{n-1}{2n-1}, \beta_2^* = \dfrac{1}{2}, \beta_3^* = \dfrac{1}{2(2n-1)}$, therefore
   $$\frac{\|(I - \mathcal{P}_2)\, \alpha_1\, u_1\|}{\|\alpha\|} \leq \frac{n-1}{2n-1}.$$
   - If $m = 3$, then $s = 1$ and the extremal set is $\{2, n+1, N\}$, and we have $p_3(w) = 2 - 2\dfrac{(w-2)(w-N)}{(n-1)(n+1-N)}$, and
   $$\frac{\|(I - \mathcal{P}_3)\, \alpha_1\, u_1\|}{\|\alpha\|} \leq \frac{(n-1)}{2(n+1)}.$$
2. If $N$ is odd ($N = 2n+1$).
   - If $m = 2$, then $s = 1$ and the extremal set is unique and is $\{2, N\}$, the polynomial verifying the conditions (C1) is $p_2(w) = 2 - 2\dfrac{w-2}{N-2}$ and we have
   $$\frac{\|(I - \mathcal{P}_2)\, \alpha_1\, u_1\|}{\|\alpha\|} \leq \frac{2n-1}{4n}.$$
   - If $m = 3$, then $s = 2$. The sets $\{2, n+1, 2n+1\}$ and $\{2, n+2, 2n+1\}$ are the two extremal sets with alternating sign. The polynomial $p_3$ is given by $p_3(w) = 2 - 2\dfrac{(w-2)(w-N)}{(n-1)(n+1-N)}$, and we have
   $$\frac{\|(I - \mathcal{P}_3)\, \alpha_1\, u_1\|}{\|\alpha\|} \leq \frac{2n+2}{2(n-1)}.$$

     Notice that $p_3(n+1) = p_3(n+2) = 0$. And the solution of the system 6.18, is
   $$\begin{cases} \beta_1^* = \dfrac{2n+2}{2(n-1)} \\ \beta_2^* = \dfrac{n^2}{2n^2+n-1} - \beta_4^* \, \dfrac{n+1}{2n^2+n-1} \\ \beta_3^* = \dfrac{1}{n+1} - \beta_4^* \\ \beta_5^* = \dfrac{n-1}{2(2n^2+n-1)} + 2\, \beta_4^* \, \dfrac{n+1}{2(2n^2+n-1)} \\ \beta_4^* \in \left[0, \dfrac{1}{n+1}\right] \end{cases}$$

     We remark that neither $\beta_2^*$ nor $\beta_5^*$ can be zero.

*Example 3.* Here we reconsider Example 1 ( $N = 4$, $\lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 2 + \imath, \lambda_4 = 3$).

- For $m = 2$, we have $s = 2$ and the polynomial $p_2$ verifying conditions (C2) is $p_2(w) = 1 + \dfrac{8-i}{\sqrt{5}} - \dfrac{3-i}{\sqrt{5}} w$, and we have $p_2(1) = 1 + \sqrt{5}$.