Helsinki University of Technology Institute of Mathematics Research Reports
Teknillisen korkeakoulun matematiikan laitoksen tutkimusraporttisarja
Espoo 1999 A412

IDEAL GMRES CAN BE BOUNDED FROM BELOW BY THREE FACTORS

Marko Huhtanen

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

Marko Huhtanen: Ideal GMRES can be bounded from below by three factors; Helsinki University of Technology Institute of Mathematics Research Reports A412 (1999).

Abstract: *For a matrix $A \in \mathbb{C}^{n\times n}$ suppose the task is to estimate how iterative methods behave for $Ax = b$ with $b \in \mathbb{C}^n$. If A is normal, then the behavior of GMRES [20] can be forecasted by considering the spectrum of A. However, when A is not normal, the spectrum can be a poor indicator of convergence. In this paper we perturb A with a matrix F of rank $k \ll n$ so as to get bounds for $\min_{p\in\mathcal{P}_j(0)} \|p(A)\|$ by using the spectrum of the resulting matrix $A - F$. In fact, if F is such that $A - F$ is similar to a normal matrix N_F with $A - F = X_F N_F X_F^{-1}$ and $\kappa(X_F) = \|X_F\|\|X_F^{-1}\|$, then*

$$\min_{p\in\mathcal{P}_j(0)} \|p(A)\| \geq \min_{p\in\mathcal{P}_j(0)} \frac{1}{\kappa(X_F)} |\lambda_{k_j+1}(p(N_F))|, \tag{1}$$

where k_j is the dimension of the Krylov subspace $\mathcal{K}_j(A; F)$. Consequently, our problem becomes: How to form a small rank F such that the spectrum of $A - F$ is, so to speak, "maximized", and, at the same time the condition number of X_F is "reasonable"? In a nutshell, we choose to proceed as follows. First we "open up" the spectrum of A with a low rank matrix F so that $A - F = X_F N_F X_F^{-1}$ and then we construct a perturbation G of X_F to make $\kappa(X_F)$ smaller, if necessary.

AMS subject classifications: 65F10, 65F15, 15A18, 15A45 .

Keywords: iterative methods, Krylov subspace, ideal GMRES, nonnormality, pole placement, condition number, ill-conditioned eigenvalue, optimal diagonalization.

ISBN 951-22-4405-5
ISSN 0784-3143
Edita, Espoo, 1999

Helsinki University of Technology
Department of Engineering Physics and Mathematics
Institute of Mathematics
P.O. Box 1100, 02015 HUT, Finland
email: *math@hut.fi*
downloadables: *http://www.math.hut.fi/*

author's email: *Marko.Huhtanen@hut.fi*

1 Introduction and notation

Assume we have a large, possibly sparse, invertible matrix $A \in \mathbb{C}^{n\times n}$ for which we consider solving iteratively

$$Ax = b \tag{2}$$

for a vector $b \in \mathbb{C}^n$. Before executing any particular method (or after an unsuccessful attempt) it would be nice to have an estimate, or a forecast, of the convergence behavior of the approximations. It is well-known that the speed of convergence of iterative methods can be related to approximation problems on the spectrum $\sigma(A)$ of A as follows (see e.g. [8, 19]). If $\mathcal{P}_j(0)$ denotes the set of polynomials of degree at most $j \geq 1$, normalized such that $p(0) = 1$, then the first analysis of the speed of convergence of iterative methods is typically based on the classical bound

$$\min_{p\in\mathcal{P}_j(0)} \|p(A)\| \geq \min_{p\in\mathcal{P}_j(0)} \max_{\lambda\in\sigma(A)} |p(\lambda)| \tag{3}$$

for the ideal GMRES. We denote by $\|\cdot\|$ the spectral norm and by $\sigma(A)$ the spectrum of A. However, this lower bound may yield an over-optimistic estimate if, for instance, the spectrum of A is small compared with other spectral sets of A, like the field of values of A. Then, typically, the eigenvalues of A are ill-conditioned and a small perturbation of A can spread out the spectrum of A violently. The ϵ-pseudospectrum, see [22] by L. Trefethen, is based on this idea. Then A is perturbed with a set of *small norm* matrices E fulfilling $\|E\| \leq \epsilon$ and the information for the speed of convergence is the resulting union of the eigenvalues.

In this paper we study bounds for iterative methods when A is perturbed with a somehow constructed, specific, *small rank* matrix F. The suggested approach is based on a generalization of (3) we failed to notice in [14]. In fact, combining this with the results of [14], our conclusion is that there are several routes to build lower bounds for $\min_{p\in\mathcal{P}_j(0)} \|p(A)\|$ based on spectra of matrices constructed in small rank perturbations of A. To see this, let $\sigma_k(A)$ denote the k^{th} singular value of A and $\lambda_k(A)$ the k^{th} eigenvalue of A arranged in decreasing order in absolute value. We showed that, if A is perturbed with F of small rank, then

$$\min_{p\in\mathcal{P}_j(0)} \|p(A)\| \geq \min_{p\in\mathcal{P}_j(0)} \sigma_{k_j+1}(p(A-F)) \tag{4}$$

[14]. Here k_j denotes the dimension of the Krylov subspace

$$\mathcal{K}_j(A;F) = \text{span}\{F, AF, ..., A^{j-1}F\} := \text{span}_{0\leq k\leq j-1}\{A^k f_1, ..., A^k f_{\text{rank}(F)}\}, \tag{5}$$

where $f_1, ..., f_{\text{rank}(F)} \in \mathbb{C}^n$ span the range of F. In particular, if A is such that the perturbed $A - F$ is normal, then the right-hand side of (4) equals $\min_{p\in\mathcal{P}_j(0)} |\lambda_{k_j+1}(p(A-F))|$, i.e., we have an approximation problem on the spectrum of $A-F$. This was done in [14] and a related problem was studied in [13]. A drawback in this approach is that, in general, the rank of k can be large for the difference $A-F$ to be normal. Thus, we relax this assumption to the other extreme as follows: Let F be a small rank matrix such that $A-F$ is (only!) similar to a normal matrix N_F and $A-F=X_FN_FX_F^{-1}$. Then, by using the techniques of [14], it is straightforward to show that

$$\min_{p\in\mathcal{P}_j(0)}\|p(A)\|\geq\min_{p\in\mathcal{P}_j(0)}\frac{1}{\kappa(X_F)}|\lambda_{k_j+1}(p(N_F))|, \tag{6}$$

where $\kappa(X_F):=\|X_F\|\|X_F^{-1}\|$ is the condition number of the eigenbasis X_F. Consequently, the quality of this bound depends on three factors: On the growth of the dimension of the Krylov subspace $\mathcal{K}_j(A;F)$, the structure of the spectrum N_F and the size of the condition number $\kappa(X_F)$.

Based on (6), we analyze how a small rank perturbation F should be chosen in order make this bound near optimal as such. This problem of having freedom to choose F and then to form $A-F$ resembles the robust pole assignment problem in control theory, see e.g. [16, 17, 18]. Of course, our task now is *not* to place the eigenvalues to any particular position but to find a perturbation F such that the right-hand side of (6) is as large as possible. Consequently, we have a kind of multiple criteria optimization problem as, at the same time, with a small rank F the spectrum of $A-F$ should be "plentiful" but not so that $\kappa(X_F)$ becomes too large. For this purpose we suggest the following steps for bounding $\min_{p\in\mathcal{P}_j(0)}\|p(A)\|$ from below.
(0.) Perturb A with F_0 such that $A-F_0$ is a diagonalizable matrix.
(1.) Divide the eigenvectors of $A-F_0$ into "insensitive" and "sensitive" in order to locate the nonnormality of $A-F_0$.
(2.) "Open up" the sensitive eigenvalues of $A-F_0$ corresponding to the sensitive eigenvectors with F_1.
(3.) Improve the condition number of a diagonalization of $A-F_0-F_1$ with F_2.

After these steps have been completed, the perturbation of A is $F=F_0+F_1+F_2$. In particular, since we assume that A can be seriously nonnormal, we need to take all the steps into account. Step (0.) is needed since we cannot assume that A is diagonalizable, i.e., similar to a normal matrix. At step (1.) we cannot assume A to be semi-simple etc.

The paper is organized as follows. In Section 2 we present different ways to bound ideal GMRES from below. Then in Section 3 we analyze how to perturb A in order to achieve good bounds with the results of Section 2. In Subsection 3.1 we cover how to perturb A with F_0 so that it becomes a diagonalizable matrix. In Subsection 3.2 we locate the cause of nonnormality for $A-F_0$ and the rank of a matrix needed to open up its spectrum. In Subsection 3.3 we describe ways to perturb $A-F_0$ with F_1 so that its spectrum will be "plentiful" but not so that the condition number of an optimal diagonalization is increased at the same time. In Subsection 3.4 we describe how to improve the condition number of a diagonalization of $A-F_0-F_1$ with a small rank matrix F_2.

2 Ideal GMRES can be bounded from below by three factors

A natural problem related to solving a linear system

$$Ax = b \tag{7}$$

with an iterative method is the characterization of the behavior of

$$\min_{p \in \mathcal{P}_j(0)} \|p(A)\| \text{ and } \min_{p \in \mathcal{P}_j(0)} \|p(A)b\|, \tag{8}$$

see, for instance, [8, 19, 11]. These both are difficult problems, but, as to lower bounds, the former, also called ideal GMRES, can be estimated by using the classical inequality based on the spectral mapping theorem

$$\min_{p \in \mathcal{P}_j(0)} \|p(A)\| \geq \min_{p \in \mathcal{P}_j(0)} \max_{\lambda \in \sigma(A)} |p(\lambda)|. \tag{9}$$

There are, however, matrices A for which the right-hand side of (9) fails to indicate the decay of $\min_{p \in \mathcal{P}_j(0)} \|p(A)\|$ well, that is, the spectrum of A is somehow "small". For this to be true A needs to be nonnormal. Indicators predicting the behavior of ideal GMRES in these cases have also been developed. In particular, it is well-known that for nonnormal matrices the spectrum of A can be much smaller than $\mathcal{F}(A)$, the field of values of A, see [3] by M. Eiermann, or, the ϵ-pseudospectra of A, see [22] by L. Trefethen. Consequently, then estimates of how iterative methods behave for the system (7) are based on these sets. In [14] we took an approach where we linked A to a normal matrix N close to A in small rank perturbations. More precisely, we looked for F of smallest possible rank such that $N = A - F$ is normal. Then, using the spectrum of N, we were able to estimate $\min_{p \in \mathcal{P}_j(0)} \|p(A)\|$ from below as follows.

Theorem 1 *[14] Suppose $A - F \in \mathbb{C}^{n \times n}$ is normal. Then*

$$\min_{p \in \mathcal{P}_j(0)} \|p(A)\| \geq \min_{p \in \mathcal{P}_j(0)} |\lambda_{k_j+1}(p(A - F))|, \tag{10}$$

where k_j denotes the dimension of $\mathcal{K}_j(A; F) = \operatorname{span}\{F, AF, ..., A^{j-1}F\}$.

One point we did not notice in [14] and want to bring up is the following. The quantity $\min_{p \in \mathcal{P}_j(0)} \|p(A)\|$ equals $\min_{p \in \mathcal{P}_j(0)} \|p(A^*)\|$ for the adjoint of A However, the growth of the dimension of a Krylov subspace (5) can be very different for the adjoint operator and better bounds can possibly be obtained with $A^* - F^*$ while monitoring the growth of the dimension of $\mathcal{K}_j(A^*; F^*)$ instead. Let us illustrate this with an example. We denote by $\dim(V)$ the dimension of a subspace $V \subset \mathbb{C}^n$.

EXAMPLE 1. Suppose $N \in \mathbb{C}^{n \times n}$ is a diagonal (i.e., normal) and cyclic matrix. Let $e_k \in \mathbb{C}^n$ be a standard unit vector and $v \in \mathbb{C}^n$ a cyclic vector for N and set $A = N + ve_k^*$. Now $\dim(\mathcal{K}_j(N; ve_k^*)) = j$ where as for the adjoint $\dim(\mathcal{K}_j(N^*; e_k v^*)) = 1$ for all $1 \leq j \leq n$. Consequently, the bound of Theorem 1 is significantly better when applied with $A^* - e_k v^* = N^*$ instead.

A variation of Theorem 1 can be obtained by using a Jordan decomposition $A = XJX^{-1}$ of A. Namely, J is easily perturbed to normal by adding ones to the lower left-corner of each Jordan block of size larger than one, see Proposition 5 for the exact statement. Let F denote the resulting perturbation, that is, the rank of the perturbation is the number of Jordan blocks of size larger than one. With this construction we were able to show the following.

Theorem 2 *[14] Let $A = XJX^{-1}$ be a Jordan decomposition of $A \in \mathbb{C}^{n\times n}$ and F as described above. Then*

$$\min_{p\in\mathcal{P}_j(\infty)} \|p(A)\| \geq \frac{1}{\kappa(X)} \min_{p\in\mathcal{P}_j(\infty)} |\lambda_{k_j+1}(p(J+F))|, \tag{11}$$

where $k_j = \dim(\mathcal{K}_j(J; F)) \leq n-1$ and $\kappa(X) = \|X\|\|X^{-1}\|$.

The idea here was to construct a normal matrix similar to A by manipulating the block-matrix J similar to A in a Jordan canonical form. The Frobenius canonical form is another decomposition in which the resulting matrix similar to A is easy to low-rank perturb to a normal matrix. Recall that every matrix is similar to a Frobenius canonical form, see e.g. [23]. If $A = XBX^{-1}$ is such a form, then B is block-diagonal with blocks being companion matrices. Like Jordan blocks, it is straightforward to rank-one correct companion matrices to normal matrices, see Proposition 6. Even more, then the blocks are unitary.

A further useful bound we did not notice in [14] can be derived as follows. Assume we look for F of smallest (or nearly smallest) rank such that $A - F$ is only diagonalizable.

Theorem 3 *Suppose for $A \in \mathbb{C}^{n\times n}$ F is such that $A - F$ is diagonalizable with $A - F = X_F \Lambda_F X_F^{-1}$. Then*

$$\min_{p\in\mathcal{P}_j(0)} \|p(A)\| \geq \min_{p\in\mathcal{P}_j(0)} \frac{1}{\kappa(X_F)} |\lambda_{k_j+1}(p(\Lambda_F))|, \tag{12}$$

where $k_j = \dim(\mathcal{K}_j(A; F)) \leq n-1$ and $\kappa(X_F) = \|X_F\|\|X_F^{-1}\|$.

Proof. First of all, $A = A - F + F = X_F \Lambda_F X_F^{-1} + F$, so that there holds

$$\min_{p\in\mathcal{P}_j(0)} \|p(A)\| \geq \min_{p\in\mathcal{P}_j(0)} \sigma_{k_j+1}(p(A-F)),$$

see Corollary 3.10 in [14]. On the other hand, by using Theorem 3.3.16 [12] twice we have

$$\sigma_{k_j+1}(p(\Lambda_F)) = \sigma_{k_j+1}(X_F^{-1} X_F p(\Lambda_F) X_F^{-1} X_F) \leq$$

$$\sigma_{k_j+1}(X_F p(\Lambda_F) X_F^{-1}) \|X_F\| \|X_F^{-1}\| = \sigma_{k_j+1}(p(A-F)) \|X_F\| \|X_F^{-1}\|$$

and the claim follows as Λ_F is normal so that $\sigma_{k_j+1}(p(\Lambda_F)) = |\lambda_{k_j+1}(p(\Lambda_F))|$.
□

Thus, this yields another route to build lower bounds as follows: With an appropriate small rank perturbation, connect A via a similarity transformation to a normal matrix. Then solve a minimization problem on the spectrum of this resulting normal matrix. Note that the fenomenon of Example 1 applies also in this case, that is, it is possible that the growth of the Block-Krylov subspace is slower when A^* is applied to F^*. Thus, working with A^* instead can yield a better bound. The proof of the following is now straightforward since every normal matrix is unitary similar to a diagonal matrix.

Corollary 4 *Suppose $A - F = X_F N_F X_F^{-1}$ with $N_F \in \mathbb{C}^{n \times n}$ normal. Then*

$$\min_{p \in \mathcal{P}_j(0)} \|p(A)\| \geq \min_{p \in \mathcal{P}_j(0)} \frac{1}{\kappa(X_F)} |\lambda_{k_j+1}(p(N_F))|, \tag{13}$$

where $k_j = \dim(\mathcal{K}_j(A; F)) \leq n-1$ and $\kappa(X_F) = \|X_F\| \|X_F^{-1}\|$.

The usability of these theorems for lower bounds is based on the well-known fact that small rank perturbations can change the eigenvalues of a matrix $A \in \mathbb{C}^{n \times n}$ violently. This property is also widely used in control theory. For instance, in the pole assignment problem [24], with a single input, the task is to find, for a given $u \in \mathbb{C}^n$, a vector $v \in \mathbb{C}^n$ such that with $F = uv^*$ the matrix $A - F$ attains some preassigned eigenvalues. All this is based on the property that the eigenvalues of A can change strongly in small rank perturbations. This is also a partial reason for why the information provided by (9) can be misleading as iterative methods typically behave almost similarly in small rank perturbations, as long as the perturbed matrix has condition number close to the original matrix. In particular, changing the eigenvalue structure of A with a small rank F such that $A - F$ has "a rich spectrum" compared with A is the key for better bounds. At this point let us consider an example.

EXAMPLE 2. This example is from [14]. Suppose $A \in \mathbb{C}^{n \times n}$ is a translated nilpotent backward shift, i.e., A is invertible,

$$A = \begin{bmatrix} 1 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \text{ and } F = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ -1 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Thus, with this rank-one matrix $A - F$ becomes even normal, that is, $\kappa(X_F) = 1$. It is also easy to see that $k_j = j$. Furthermore, the spectrum of A is just the point 1. However, for $A - F$ the spectrum is "plentiful", that is, the set

$1 + \{z \in \mathbb{C} : z^n = 1\}$.

Essentially the bound (12) is of use in case A is from the set

$$\{A = X(N+F)X^{-1} \in \mathbb{C}^{n\times n} : \kappa(X) \leq 1+\delta, N \text{ is normal, rank}(F) \leq k\} \tag{14}$$

for some *moderate* $\delta \geq 0$ and $k \geq 0$. In particular, to our mind, only very seriously nonnormal matrices do not belong to (14) and allow to derive useful bounds. For actual bounds, some factors X, N and F that yield A need to be "inverted" based on knowing only A. This is the purpose of the following section. The corresponding inversion was covered in [14, 13] in case $\delta = 0$.

3 How to get good lower bounds for ideal GMRES with (12)

As can be seen from the bound (12), the problem is essentially how to make A more normal with a low rank matrix F so that the spectrum of A "opens up" in the perturbation, and, at the same time, to control the size of $\kappa(X_F)$. In the lack of better expression we use the intuitive description "open up" the spectrum. In this section we suggest the following steps for bounding $\min_{p\in\mathcal{P}_j(0)} \|p(A)\|$ from below.
(0.) Perturb A with F_0 such that $A - F_0$ is a diagonalizable matrix.
(1.) Divide the eigenvectors of $A - F_0$ into "insensitive" and "sensitive" in order to locate the nonnormality of $A - F_0$.
(2.) "Open up" the sensitive eigenvalues of $A - F_0$ corresponding to the sensitive eigenvectors with F_1.
(3.) Improve the condition number of a diagonalization of $A - F_0 - F_1$ with F_2.

After these steps have been completed, the perturbation of A is $F = F_0 + F_1 + F_2$.

This section is divided into 4 subsections where we cover each of these steps separately. Obviously, not all the steps are indispensable in practise. For instance, in Example 2 we only had to perform the step (0.).

3.1 Perturbing A with F_0 such that $A - F_0$ is a diagonalizable matrix

In Example 2, the matrix A was not diagonalizable. Still, in rank-one perturbation A became even normal. As we do not want to "waste" rank in perturbations, for us an important factor is the rank of F_0 that yields $A - F_0$ similar to a normal matrix. For that purpose we say that A is *k-rank diagonalizable*, if there exists F of rank k such that $A - F$ is diagonalizable and no G of rank less than k exist such that $A - G$ is diagonalizable. Two simple ways to make A diagonalizable are based on classical decompositions.

Proposition 5 *Let j be the number of the Jordan blocks of $A \in \mathbb{C}^{n\times n}$ of size larger than one. Then there exists a matrix F of rank j such that $A-F$ is diagonalizable.*

Proof. Let $A = XJX^{-1}$ be a Jordan decomposition of A. Let $J_j(\lambda_k)$ be a Jordan block of A of order larger than one related to an eigenvalue λ_k of A. It is easy to see that adding 1 to the lower left-corner of $J_j(\lambda_k)$ gives a normal matrix. This corresponds to a rank-one perturbation of A. After making these perturbations to each Jordan blocks of A of size larger than one we have $A - F = XNX^{-1}$ with N normal and F of rank j. Let $N = U\Lambda U^*$ be a diagonalization of N. Then $A - F = (XU)\Lambda U^* X^{-1}$ is a diagonalization of $A - F$. □

The Frobenius canonical form can also be a starting point for a small rank perturbation. The following can be proved in a similar vain.

Proposition 6 *Let j be the number of blocks in a Frobenius form of $A \in \mathbb{C}^{n\times n}$. Then there exists a matrix F of rank j such that $A - F$ is similar to a unitary matrix.*

Proof. In the following B_r is the Frobenius matrix and F_r is the perturbation corresponding to the r^{th} block:

$$B_r = \begin{bmatrix} b_{r-1} & b_{r-2} & \cdots & b_1 & b_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \; F_r = \begin{bmatrix} b_{r-1} & b_{r-2} & \cdots & b_1 & b_0 - 1 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Consequently, $B_r - F_r$ is diagonalizable (and even unitary). □

In particular, if A is cyclic, then at most rank-one perturbation is needed to make the companion matrix diagonalizable and thereby A. Obviously it is difficult to compute these decompositions in finite aritmethics. For computational aspects of the Jordan decomposition, see [7] by G. Golub and J. Wilkinson. Since A was perturbed similar to a unitary matrix in Proposition 6, some excess rank is possibly consumed. This can be avoided as follows.

Proposition 7 *Suppose $A \in \mathbb{C}^{n\times n}$ is k-rank diagonalizable. Then k equals*

$$\max_{\lambda\in\sigma(A)} \{\#\textit{Jordan blocks associated with } \lambda \textit{ of size larger than one}\}. \tag{15}$$

Proof. Here we use the relation between the Frobenius and the Jordan canonical forms [23]. Consider the Frobenius canonical form in the form where the elementary divisors of A are grouped together such that the corresponding Jordan submatrices of highest order in each eigenvalue gives the first block. Then those of next highest order gives the following and so on, see e.g. [23][pp. 16-17]. Then, after a block-similarity transformation, we obtain a Frobenius canonical form. Let p be the number in (15). Because of the ordering, only p first Frobenius matrices in the decomposition have multiple eigenvalues. Thus, only F of rank p corresponding to these blocks constructed as in the proof of Proposition 6 is needed to make this decomposition diagonalizable.

To see that p is also a lower-bound for diagonalizability of A in a perturbation, take $p-1$ vectors $b_1, ..., b_{p-1} \in \mathbb{C}^n$ and assume $F = \sum_{j=1}^{p-1} b_j c_j^*$ with some $c_1, ..., c_{p-1} \in \mathbb{C}^n$. Denote by $B = [b_1, ..., b_{p-1}] \in \mathbb{C}^{n \times k}$ and form a Block-Krylov subspace

$$K(A; B) = \operatorname{span}_{j \geq 0}\{A^j b_1, ..., A^j b_{p-1}\}$$

and form an orthogonal basis of $K(A; B)$ with the Arnoldi method. This does not span the whole $\mathbb{C}^n$ as (A, B) is not controllable (for the definition of controllability, see [24]). Namely, constructing a SVD decomposition of $A - \lambda_j I$ for an eigenvalue realizing (15) we obtain an eigenvalue that cannot be relocated. Complete $K(A; B)$ to a basis of $\mathbb{C}^n$ so that A equals $\begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix}$. Since A_1 is controllable, it has at most $p-1$ invariant factors different from 1, see [24][Theorem 1.2]. Now if $\alpha_1 | ... | \alpha_q$ and $\gamma_1 | ... | \gamma_n$ are the invariant factors of A_3 and A respectively, then

$$\gamma_{i+n-q-r} | \alpha_i | \gamma_{i+n-q}, \ 1 \leq i \leq q, \tag{16}$$

where r is the number of invariant factors of A_1 different from 1, see [1] by D. Carlson. In particular, $\gamma_{n-r} | \alpha_q$ so that A_3 has an invariant factor corresponding to a Jordan-block of size larger than 1 as $r \leq p-1$. Since in this representation the sum $A + F$ does not alter the rows below the size of A_1, this Jordan-block of A_3 remains unaltered for $A + F$. But then, applying a result of I. Zaballa [25] we have, if $\hat{\gamma}_1 | ... | \hat{\gamma}_n$ denote the invariant factors of $A + F$, that then $\alpha_q | \hat{\gamma}_n$ must hold. In particular, $A + F$ is not diagonalizable. Since this holds for any F with rank at most $p-1$, the claim follows. □

3.2 Dividing the eigenvectors of $A - F_0$ into "insensitive" and "sensitive" in order to locate the nonnormality of $A - F_0$

After having a diagonalizable $A - F_0$, the next step is to open up the spectrum of $A - F_0$ with a small rank F_1. Before this, we need to know better which eigenvectors really should be perturbed. In particular, we want to locate the cause of nonnormality in $A - F_0$. Second, while opening up the spectrum, the question arises whether we can move the eigenvalues to any preassigned point set of $\mathbb{C}$? As to this problem, there is a direct connection to control theory since in the pole placement problem the task is to move, with a *very low* rank matrix, the eigenvalues of A at least to the left half-plane. For this purpose, recall that a Block-Krylov subspace of $A \in \mathbb{C}^{n\times n}$ at $F \in \mathbb{C}^{n\times k}$ is defined via

$$\mathcal{K}(A;F) := \operatorname{span}\{F, AF, A^2F, ...\}, \tag{17}$$

that is, in (5) the number of steps is n, if necessary. In particular, if $\mathcal{K}(A;F) = \mathbb{C}^n$, then a system with the generator A and the input matrix F is said to be *controllable*, see e.g. [24]. Equipped with this, let us state the pole assignment theorem as follows.

Proposition 8 *[24] The spectrum of $A \in \mathbb{C}^{n\times n}$ can be placed to any n points of $\mathbb{C}$ in a k-rank perturbation if and only if $\mathcal{K}(A;F) = \mathbb{C}^n$ for a $F \in \mathbb{C}^{n\times k}$.*

Thus, the smallest rank matrix F for which $A - F$ can have any preassigned eigenvalues is determined by a Block-Krylov subspace criterion. In particular, if A is cyclic, then a rank-one F will suffice. However, combining this with Theorem 3, it is apparent that if the eigenvalues of A are transfered very far with a matrix F, for instance, far outside the numerical range of A, then the condition number $\kappa(X_F)$ must be large to compensate this growth of the spectrum. This is because the left-hand side of (12) is also an upper bound for our variables, that is, for the rank of F, the spectrum of $A - F$ and the condition number $\kappa(X_F)$. Still, for our purposes it is very important to have no restriction where to exactly place the eigenvalues since then the problem can, in a sense, be ill-posed, see [17] by V. Mehrmann and H. Xu. Because of this, exact relocation of the poles has also been relaxed in the sense the poles are only being placed to a certain region of $\mathbb{C}$, see [18] V. Mehrmann and H. Xu. This problem is clearly closer to our task.

Obviously the controllability is not that important for us per se. For instance, we are completely satisfied if our matrix is the identity as it is normal and the speed of iteration is exactly revealed by the spectrum. Still, the identity matrix is clearly not controllable unless the rank of the input matrix is the same as the dimension. To illustrate this with another example, suppose A is unitary similar to a block-diagonal matrix $B \oplus \Lambda$, where Λ is a diagonal matrix and $B \in \mathbb{C}^{(n-k)\times(n-k)}$ with $k > 0$. Then all the eigenvectors corresponding to Λ are well-conditioned and obviously F needs to be constructed for B only to make A more normal. In particular, Proposition 8 is now relevant for B as, being the cause of nonnormality for A, only its eigenvalues need to be relocated. In order to use this type of approach, we thus need to separate "good" eigenvectors from "bad". For that purpose, finding a unitary similarity that block-diagonalizes A in this manner is a too stringent operation that yields, generically, no Λ-block. And even if it did, it may not be readily performed in finite aritmethics. Instead, to really be able to locate the nonnormality of A, we need not quite so clear cut as follows. The construction is somewhat tedious and it will take the rest of this subsection (see Proposition 11 for the exact statement and (30)).

Recall that if $\lambda_j \in \mathbb{C}$ is a simple eigenvalue of A, then with a pair $x_j, y_j \in \mathbb{C}^n$ of right and left eigenvectors of A, that is,

$$Ax_j = \lambda_j x_j \text{ and } A^* y_j = \overline{\lambda}_j y_j, \tag{18}$$

the condition number of λ_j is defined as $1/s_j$, where

$$s_j := \frac{|y_j^* x_j|}{\|x_j\| \|y_j\|}, \tag{19}$$

see e.g. [23] or [6]. This is the cosine of the angle between the right and left eigenvectors corresponding to λ_j. Here we assumed that the eigenvalue λ_j is simple but we need to take into account the more general case when this does not necessarily hold. To that end we need to consider diagonalizations of $A - F_0$, that is, suppose we have constructed a perturbation F_0 such that $A - F_0$ is diagonalizable. Then, if $A - F_0$ is not semi-simple, the numbers (19) depend on the constructed similarity transformation X_{F_0} in the diagonalization $A - F_0 = X_{F_0} \Lambda_{F_0} X_{F_0}^{-1}$. For this purpose a *scaled* similarity transformation is a good choice. A similarity transformation is said to be scaled if the eigenvectors corresponding to each spectral subspace are chosen to be orthonormal. One can, however, do better. For that purpose set

$$s(X) := \sum_{j=1}^{n} |s_j|^{-1} \tag{20}$$

for a diagonalizable $A \in \mathbb{C}^{n \times n}$ corresponding to a diagonalization $A = X \Lambda X^{-1}$. That is, we have $A = \sum_{j=1}^{n} x_j \lambda_j y_j^*$, where x_j and y_j are the columns of X and X^{-*} respectively and the numbers s_j are defined via (19) with these vectors corresponding to this particular diagonalization. Let $\kappa(A)$ and $\kappa_{\mathcal{F}}(A)$ denote the condition number of a matrix A in the the spectral norm and the Frobenius norm respectively.

Proposition 9 *Let $A \in \mathbb{C}^{n \times n}$ be diagonalizable. Then for an optimal diagonalization in the Frobenius norm of $A = X \Lambda X^{-1}$*

$$\kappa_{\mathcal{F}}(X) = s(X) = \min_{A = Y \Lambda Y^{-1}} s(Y). \tag{21}$$

Proof. If all the eigenvalues of A are semi-simple, then this follows from [21]. Thus, assume there are multiple eigenvalues, say k, and $A = Y \Lambda Y^{-1}$ is a diagonalization of A. Without loss of generality, let

$$\Lambda = \operatorname{diag}(\lambda_1, ..., \lambda_1, \lambda_2, ..., \lambda_2, \lambda_k, ..., \lambda_k, \lambda_{k+1}, \lambda_{k+2}, ..., \lambda_p),$$

with $\lambda_i \neq \lambda_j$, for $k+1 \leq i, j \leq p$ and $i \neq j$. Perturb A analytically with $Y D(\alpha) Y^{-1}$, where

$$D(\alpha) = \operatorname{diag}(\alpha, 2\alpha, 3\alpha ..., (n-p)\alpha, 0, 0, ..., 0), \tag{22}$$

so that $A(\alpha) := A + YD(\alpha)Y^{-1}$ has just simple eigenvalues for small $|\alpha| > 0$. Clearly, by construction, the eigenvectors of $A(\alpha)$ are also eigenvectors of A. Now, [21] gives an optimal diagonalization Y_{opt} for $A(\alpha)$, that is, for this diagonalization it holds $s(Y) = \kappa_{\mathcal{F}}(Y_{opt}) \leq \kappa_{\mathcal{F}}(Y)$. In particular, this diagonalization is independent on α as long as $A(\alpha)$ remains semi-simple. Obviously Y_{opt} yields a diagonalization of A as well since its columns equal the columns of Y, except that they are optimally scaled [21]. Since this holds for any Y yielding a diagonalization of A, choosing an X that minimizes (20), gives an optimal. $\square$

According to the proof of the previous proposition, for any diagonalization $A = X\Lambda X^{-1}$ of A, it is arguable to scale each column x_j of X to have length $1/\sqrt{|s_j|}$, for $1 \leq j \leq n$, as the condition number will decrease in the Frobenius norm. Further, for picking x_{k_j} and computing s_{k_j}, it is not necessary to invert X. Instead, it is possible just consider angles between subspaces, see [2][page 601], where J. Demmel considers the problem of how to diagonalize *after* fixing the invariant subspaces. Here our problem is, after computing the spectral subspaces corresponding to each eigenvalue of A, how to choose "insensitive" vectors from each subspace, in particular, from those having dimension larger than 1. This is obviously an interdependent problem.

The numbers (19) provide a criterion for picking up insensitive eigenvectors for a not necessarily semi-simple matrix A. Thus, let $A = X\Lambda X^{-1} = \sum_{j=1}^{n} x_j \lambda_j y_j^*$ be a diagonalization of A and let x_{k_j} be chosen, for $1 \leq j \leq k$, to be the insensitive eigenvectors of A corresponding to this diagonalization. Let us put

$$m_k(X) := \sum_{j=1}^{k} |s_{k_j}|^{-1} \tag{23}$$

to measure the insensitiveness of the chosen eigenvectors. The sensitiveness needs to be measured in this manner since a spectral projector belonging to a cluster of eigenvalues can have moderate norm while spectral projectors belonging to a sub-cluster of a cluster can have very large norms. This is well described by W. Kahan in [15]. Furthermore, put

$$P_k(X) = \sum_{j=1}^{k} \frac{x_{k_j} y_{k_j}^*}{y_{k_j}^* x_{k_j}} \tag{24}$$

to be the *generalized spectral projector* corresponding to the chosen eigenvectors in (23) of the diagonalization $A = X\Lambda X^{-1}$. The following lemma gives reasons for this expression.

Lemma 10 *Let $A = X\Lambda X^{-1}$ be a diagonalization of $A \in \mathbb{C}^{n\times n}$ and $P_k(X)$ defined via (24). Then $P_k(X)$ is a projector and it commutes with A.*

Proof. First of all, $P_k(X)$ is a projector, since it is the spectral projector of $A(\alpha)$ defined in (22) corresponding to this diagonalization for $|\alpha| > 0$

small. This can be seen by performing the path-integration such that each eigenvalue is surrounded separately with a path of integration and then summing the result. In particular, $P_k(X)$ is independent on α. Furthermore, since $P_k(X)A(\alpha) = A(\alpha)P_k(X)$, there holds

$$\begin{aligned}\|AP_k(X) - P_k(X)A\| &\leq \|AP_k(X) - A(\alpha)P_k(X)\| + \|P_k(X)A(\alpha) - P_k(X)A\| \\ &\leq 2\|P_k(X)\|\|A - A(\alpha)\|\end{aligned}$$

which converges to zero while $\alpha \to 0$ and the claim follows since $\|P_k(X)\|$ is uniformly bounded in α. □

We need the commutativity of the projector $P_k(X)$ in the proof of the following proposition yielding a block-diagonalization that separates the insensitive and sensitive eigenvectors. Because we use just block-diagonalization, the condition number of a similarity transformation yielding such a diagonalization can obviously be much lower than the condition number of a similarity transformation that actually diagonalizes A. Note that because $P_k(X)$ is *not* a spectral projector for A, the spectra of C_1 and C_2 can overlap.

Proposition 11 *Let $A = X\Lambda X^{-1}$ be a diagonalization of $A \in \mathbb{C}^{n\times n}$ and $P_k(X)$ be defined via (24). Then there exists a corresponding block-diagonalization*

$$Q^{-1}AQ = \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix} \text{ with } C_1 \in \mathbb{C}^{(n-k)\times(n-k)} \text{ and } C_2 \in \mathbb{C}^{k\times k} \tag{25}$$

such that $\kappa(Q) \leq 4\|P_k(X)\| \leq 4m_k(X)$. Further, if the columns of $X_k \in \mathbb{C}^{n\times k}$ consist of the chosen k insensitive eigenvectors of A, then the columns of $X_2 \in \mathbb{C}^{k\times k}$ are the eigenvectors of C_2, where $[0\ X_2^*]^* = Q^{-1}X_k$.

Proof. Let us form a Schur decomposition of A such that

$$U^*AU = \begin{bmatrix} Z_1 & Z_1R - RZ_2 \\ 0 & Z_2 \end{bmatrix}, \tag{26}$$

where the block $Z_1R - RZ_2$ written in this form reveals the correlation with the generalized spectral projector

$$P_k(X) = \begin{bmatrix} 0 & -R \\ 0 & I \end{bmatrix}$$

in this co-ordinate system. Moreover, if we use the co-ordinate system where $P_k(X)$ appears as $\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}$, then we have the relation

$$\begin{bmatrix} Z_1 & Z_1R - RZ_2 \\ 0 & Z_2 \end{bmatrix} = \begin{bmatrix} I & -R \\ 0 & I \end{bmatrix}\begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix}\begin{bmatrix} I & R \\ 0 & I \end{bmatrix}.$$

(Note that since Z_1 and Z_2 may have overlapping spectra, we have the problem which eigenvalues to choose if we actually want to construct the Schur decomposition (26) such that Z_2 corresponds to chosen insensitive eigenvectors. One way to achieve this is to carry out algorithmically the Schur decomposition as follows: Assume that the eigenvectors are already ordered in X so that $x_{n-k}, \ldots, x_n$ are the chosen insensitive ones. Then after Gram-Schmidt process we have a unitary $U = XT$ with an upper-triangular matrix T. Thus if $A = X\Lambda X^{-1}$, then $A = UT^{-1}\Lambda TU^*$ gives a prescribed Schur decomposition. A further note: this Schur decomposition is smooth for $A(\alpha)$ in variable α, since U and T are constant matrices then.)

Now we use a trick from [15] by W. Kahan: with any invertible $T \in \mathbb{C}^{k\times k}$ and $S \in \mathbb{C}^{(n-k)\times(n-k)}$ we obtain a similarity transformation

$$Q = \begin{bmatrix} S & -RT \\ 0 & T \end{bmatrix} \text{ and } Q^{-1} = \begin{bmatrix} S^{-1} & S^{-1}R \\ 0 & T^{-1} \end{bmatrix},$$

so that we have $Q^{-1}U^*AUQ = \begin{bmatrix} S^{-1}Z_1S & 0 \\ 0 & T^{-1}Z_2T \end{bmatrix}$. Clearly

$$Q = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & -R \\ 0 & I \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 0 & T \end{bmatrix} \tag{27}$$

and

$$Q^{-1} = \begin{bmatrix} S^{-1} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} I & R \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & T^{-1} \end{bmatrix}. \tag{28}$$

In particular, choosing S and T such that $S^*S = \sigma^2$ and $T^*T = \tau^2$ for constants σ and τ satisfying $\frac{\sigma}{\tau} = \|P_k(X)\|$ we obtain

$$\begin{aligned}\kappa(Q) = \|Q\|\|Q^{-1}\| &\leq (\|S\| + \|P_k(X)\|\|T\|)(\|T^{-1}\| + \|S^{-1}\|\|P_k(X)\|) \\ &= (\sigma + \tau\|P_k(X)\|)(\frac{1}{\tau} + \frac{1}{\sigma}\|P_k(X)\|) = 4\|P_k(X)\|.\end{aligned}$$

The latter inequality follows immediately after using triangle inequality with (24).

For the latter part of the claim, let us consider a perturbation $A(\alpha)$ defined in (22). Then the above proved claims hold for $A(\alpha)$ as well with the same bounds since the projector $P_k(X)$ does not change in this particular perturbation. Now, for any eigenvector x corresponding to an eigenvalue λ of $C_2(\alpha)$

$$Q^{-1}U^*A(\alpha)UQx = \begin{bmatrix} C_1(\alpha) & 0 \\ 0 & C_2(\alpha) \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \lambda x_1 \\ \lambda x_2 \end{bmatrix},$$

where x is partitioned conformly. Since $A(\alpha)$ is semi-simple, λ is not an eigenvalue of $C_1(\alpha)$ and, consequently, $x_1 = 0$. Thus, we have necessarily $Q^{-1}U^*X_k = \begin{bmatrix} 0 \\ X_2 \end{bmatrix}$ and the claim follows by taking the limit as the similarity transformations had norms that were uniformly bounded in α. □

Note that a near optimal block-diagonalization in Proposition 11 can also be computed in practise, after having chosen the subspace spanned by the vectors in (23), by using orthonormal scaled diagonalization $\hat{Q}$. Namely, if Q_{opt} is optimal, then $\kappa(\hat{Q}) \leq \sqrt{2}\kappa(Q_{opt})$, see [2]. This is clearly almost optimal. As the norm of $P_k(X)$ bounds from above the condition number of Q, the similarity transformation Q is well-conditioned since the eigenvectors were chosen to be insensitive. Consequently, C_1 essentially causes the nonnormality. To state this rigorously, we need the following.

Corollary 12 *Let s_j denote the condition numbers (19) of A in a diagonalization of A and $\hat{s}_j$ the corresponding condition numbers for C_2. Then $\hat{s}_j \geq \frac{1}{\kappa(Q)} s_j$.*

Proof. Since (lower blocks of) $Q^{-1}x_{k_j}$ and $Q^* y_{k_j}$ are left and right eigenvectors of C_2, we have

$$\hat{s}_j = \frac{|y_{k_j}^* Q Q^{-1} x_{k_j}|}{\|Q^{-1}x_{k_j}\| \|Q^* y_{k_j}\|} \geq \frac{1}{\kappa(Q)} \frac{|y_{k_j}^* x_{k_j}|}{\|x_{k_j}\| \|y_{k_j}\|} \tag{29}$$

and the claim follows. $\square$

In particular, by this and Proposition 20 and Proposition 9, we can diagonalize C_2 with $Y \in \mathbb{C}^{k \times k}$ such that

$$\begin{bmatrix} I & 0 \\ 0 & Y^{-1} \end{bmatrix} Q^{-1} A Q \begin{bmatrix} I & 0 \\ 0 & Y \end{bmatrix} = \begin{bmatrix} C_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}, \tag{30}$$

where the condition number satisfies $\kappa_{\mathcal{F}}(Y) \leq \kappa(Q) m_k(X)$.

To sum up, in order to locate the nonnormality of A, we approached the problem by finding a relaxed version of the case in which A is unitary similar with $B \oplus \Lambda$ with Λ being a diagonal matrix. Here we constructed a similarity transformation, that was "almost unitary" (condition number close to 1) block-diagonalizing A such that $C_1 \oplus \Lambda_2$ with Λ_2 being a diagonal matrix. Consequently, for our purposes it is arguable to low rank perturb just C_1 so as to open up its spectrum. To that end, Proposition 8 can be stated more appropriately for bounding $\min_{p \in \mathcal{P}_j(0)} \|p(A)\|$ from below as follows: The eigenvalues of C_1 can be placed freely with $F \in \mathbb{C}^{(n-k)\times(n-k)}$ if the dimension of $\mathcal{K}(C_1; F)$ equals $n-k$.

3.3 Opening up the sensitive eigenvalues of $A - F_0$ corresponding to the sensitive eigenvectors with F_1

Assume that have a diagonalizable $A - F_0$. As analyzed in the previous subsection, the next step is to open up that part spectrum which is related to nonnormality caused by C_1, that is, we need a perturbation F_1 to open up the spectrum of C_1. This should not, however, be done at the expense of increasing the condition number in the resulting eigenbasis. Consequently, a straightforward control theoretic approach based on pole placement algorithms is not applicable. But, on the other hand, how and where to place the eigenvalues then? And, in particular, how to perform this inexpensively? A way to proceed is to use the self-commutator as a criterion in the construction of a perturbation. Namely, there holds, whenever $A-F$ is diagonalizable,

$$\kappa(X_F) \geq (1 + \frac{\|[A-F,(A-F)^*]\|_{\mathcal{F}}}{2\|(A-F)^2\|_{\mathcal{F}}})^{1/4}, \tag{31}$$

see [21]. Here $[A-F,(A-F)^*] := (A-F)(A-F)^* - (A-F)^*(A-F)$ is the self-commutator of $A-F$ and $A-F = X_F\Lambda_F X_F^{-1}$ is a diagonalization of $A-F$. Thus, (31) yields a necessary condition for $\kappa(X_F)$ to remain small (but of course always not less than 1), that is, $\|[A-F,(A-F)^*]\|_{\mathcal{F}}$ should be small. However, this is not a sufficient condition and it is easy to construct examples for which the right-hand side of (31) is arbitrarily close to 1 while the left-hand side is arbitrarily large. Still, if we construct perturbations from

$$\{F \in \mathbb{C}^{n\times n} : \text{ all the eigenvalues of } A-F \text{ are simple}\}, \tag{32}$$

then we obtain a reversed inequality.

Proposition 13 *Suppose $F \in \mathbb{C}^{n\times n}$ belongs to (32) and let $\{\lambda_j\}_{j=1}^n$ denote the eigenvalues of $A-F$ and $\delta_j := \min\{|\lambda_i-\lambda_j| : i \neq j\}$. Then for an optimal diagonalization $A-F = X_F\Lambda_F X_F^{-1}$ holds*

$$\kappa_{\mathcal{F}}(X_F) \leq \sum_{j=1}^{n} \big(1 + (\frac{n(n+1)}{12(n-1)})^{1/2} \frac{\|[A-F,(A-F)^*]\|_{\mathcal{F}}}{\delta_j^2}\big)^{(n-1)/2}.$$

Proof. Combining theorems 3 and 5 of R. A. Smith in [21] yields

$$\min\{\|X_F\|_{\mathcal{F}}\|X_F^{-1}\|_{\mathcal{F}} : A = X_F\Lambda_F X_F^{-1}\} \leq \sum_{j=1}^{n}(1+(n-1)^{-1}\delta_j^{-2}D_F^2)^{(n-1)/2},$$

where D_F denotes the departure from normality of $A-F$ of P. Henrici, and, for D_F P. Henrici has shown that $D_F^2 \leq \sqrt{(n^3-n)/12}\|[A-F,(A-F)^*]\|_{\mathcal{F}}$ [10]. Combining these yields the claimed bound for the optimal condition number. $\square$

Since this bound is in the Frobenius norm, $n \leq \kappa_{\mathcal{F}}(X_F)$ holds, which explains the growth n at least. Even though this can give a severely large overestimate, combining this with (31) provides a criterion for opening up the spectrum of C_1. Namely, using the self-commutator, we choose to minimize, for some small k

$$\min_{\text{rank}(F)\leq k} \|[A-F_0-F,(A-F_0-F)^*]\|_{\mathcal{F}}^2. \tag{33}$$

Or, alternatively, minimizing (33) sequentially means finding, at step one,

$$\min_{u,v\in\mathbb{C}^n} \|[A-F_0-uv^*,(A-F_0-uv^*)^*]\|_{\mathcal{F}}^2. \tag{34}$$

Then, repeating this sequentially as long as the self-commutator of the perturbation is reasonable, the resulting sum yields an F_1. A problem related to this, see [13]. Further, recall that the self-commutator is a measure of nonnormality. For other measures of nonnormality, see [9, 5, 4]. As to using other measures of nonnormality, rankwise they can be very misleading. For instance, the upper-triangular part of a Schur decomposition, which can be used to measure nonnormality, can have very large rank even though the matrix itself is a small perturbation of a normal matrix. This is what happens in Example 2.

Let us describe other possible approaches for opening up the spectrum of $A - F_0$ (or C_1) in some arguable manner. One possibility is to use the Ritz values that result from the Arnoldi iteration and combine this with the pole placement algorithms. Namely, when $A - F_0$ is nonnormal and the spectrum of $A - F_0$ is small compared with the field of values of A, then the Ritz values of A tend to "spend time" at the early stage of iteration in a larger set (but inside the field of values) than the spectrum of $A - F_0$. Obviously, the scheme does not need be the Arnoldi iteration method, as long as the actual eigenvalue approximation, because of nonnormality, is, so to speak, bad. Then F_1 is chosen such that $A - F_0 - F_1$ has eigenvalues in those areas where the Ritz values were located. In particular, at this point techniques of V. Mehrmann and H. Xu [18] can possibly be exploited.

And of course, one possibility is to make a small rank random perturbation. As a result the spectrum will open up, but obviously not in an optimal manner. A simple test can be performed with the matrix of Example 2.

3.4 Improving the condition number of a diagonalization of $A - F_0 - F_1$

At this final step the spectrum of the perturbed $A - F_0 - F_1$ does not change, only the condition number of the matrix $X_{F_0+F_1}$ yielding a diagonalization of $A - F_0 - F_1$ is being controlled. Obviously, the construction of F_1 in the previous section was designed so that it should also decrease the condition number simply because the self-commutator was made smaller while the spectrum was opened up. These both are necessary conditions for small condition number when $A - F_0$ stays semi-simple in the perturbation. Still, in case the condition number is too high for good bounds, we need the following.

Proposition 14 *Let $A = X\Lambda X^{-1}$ be a diagonalization of $A \in \mathbb{C}^{n\times n}$ and G such that $X + G$ is invertible. Then*

$$\operatorname{rank}(A - (X+G)\Lambda(X+G)^{-1}) \leq 2\operatorname{rank}(G). \tag{35}$$

Proof. By the Sherman-Morrison-Woodbury formula [6] we have $(X + G)^{-1} = X^{-1} - S$ with S of rank equaling $\operatorname{rank}(G)$. Consequently,

$$(X+G)\Lambda(X+G)^{-1} = X\Lambda(X+G)^{-1} + G\Lambda(X+G)^{-1}$$

$$= X\Lambda X^{-1} - X\Lambda S + G\Lambda(X+G)^{-1}$$

and the claim follows. □

Thus, for $A - F_0 - F_1$, after improving the condition number of a diagonalization $A - F_0 - F_1 = X\Lambda X^{-1}$ with G, we obtain

$$A - F_0 - F_1 - F_2 = (X+G)\Lambda(X+G)^{-1} \tag{36}$$

with $\text{rank}(F_2) \leq 2\text{rank}(G)$. Consequently, improving the condition number of X is, in general, twice as expensive in rank as perturbing A directly. Second, the spectrum of $A - F_0 - F_1$ equals the spectrum of $A - F_0 - F_1 - F_2$, that is, at this step only the condition number of the eigenbasis and the rank of perturbation are altered. Note that the improvement of the condition number is, in fact, the easy part of the construction of a perturbation. Namely, picking a G is straightforward since it is obtained from a singular value decomposition of X. More precisely, as

$$\kappa(X) = \frac{\sigma_1(X)}{\sigma_n(X)}, \tag{37}$$

then it is easy to decrease $\kappa(X)$ by rank-one updating the singular vectors that cause the ill-condition.

4 Conclusions

In this paper have considered lower bounds for ideal GMRES. The bounds depend on three factors as follows. First, on the growth of the dimension of the Krylov subspace that result from a small rank perturbation of A. Second, on the condition number that yields a diagonalization of $A - F$. And third, on a minimization problem on the spectrum of $A - F$.

In the latter part of the paper we analyzed how to construct a perturbation F such that the bounds will be good. This can be accomplished in 4 steps as follows.

(0.) Perturb A with F_0 such that $A - F_0$ is a diagonalizable matrix.
(1.) Divide the eigenvectors of $A - F_0$ into “insensitive” and “sensitive”.
(2.) “Open up” the sensitive eigenvalues corresponding to the sensitive eigenvectors with F_1.
(3.) Improve the condition number of a diagonalization of $A - F_0 - F_1$ with F_2.

After these steps have been completed, the perturbation of A is $F = F_0 + F_1 + F_2$.

References

[1] D. CARLSON, *Inequalities relating the degrees of elementary divisors within a matrix,* Simon Stevin, 44, pp. 3-10, 1970.

[2] J. DEMMEL, *The condition number of equivalence transformations that block diagonalize matrix pencils,* SIAM J. Numer. Anal., 20, pp. 599-610, 1983.

[3] M. EIERMANN, *Field of values and iterative methods,* Lin. Alg. Appl. 180, (1993) pp. 167-198.

[4] L. ELSNER AND KH.D IKRAMOV, *Normal matrices: an update,* Lin. Alg. Appl. 285: 291-303 (1998).

[5] L. ELSNER AND M. H. C. PAARDEKOOPER, *On measures of non-normality of matrices,* Lin. Alg. Appl. 92: 107-124 (1987).

[6] G. GOLUB AND C. VAN LOAN, Matrix Computations, *The John Hopkins Univ. Press*, 3^{rd} ed., 1996.

[7] G. GOLUB AND J. WILKINSON, *Ill-conditioned eigensystems and the computation of the Jordan canonical form.,* SIAM Rev., vol. 18, no. 4 pp. 578-619, 1976.

[8] A. GREENBAUM, Iterative Methods for Solving Linear Systems, *Frontiers in Applied Mathematics, SIAM, Philadelphia,* 1997.

[9] R. GRONE, C.R. JOHNSON, E.M. SA AND H. WOLKOWICZ, *Normal matrices,* Lin. Alg. Appl. 87: 213-225 (1987).

[10] P. HENRICI, *Bounds for iterates, inverses, spectral variation and field of values of nonnormal matrices,* Numer. Math., 4, pp. 24-40, 1962.

[11] M. HOCHBRUCK AND C. LUBICH, *Error analysis of Krylov methods in a nutshell,* SIAM J. Sci. Comput. 19 (1998), pp. 695-701.

[12] R. A. HORN AND C. R. JOHNSON, Topics in Matrix Analysis, *Cambridge Univ. Press* 1991.

[13] M. HUHTANEN, $[A, A^*]$ *and inversion of the sum operation in* $A =$ *"normal + small rank",* In preparation.

[14] M. HUHTANEN AND O. NEVANLINNA, *Minimal decompositions and iterative methods,* Centro Int. Math., Portugal, Preprint 3, 1998. Also to appear in Numer. Math.

[15] W. KAHAN, *Conserving confluence curbs ill-condition,* Computer Sci. Tech. Rep. 6, Univ. of Calif. 1972.

[16] J. KAUTSKY, N. K. NICHOLS AND P. VAN DOOREN, *Robust pole assignment in linear state feedback,* Int. J. Contr. vol. 41, no. 5, pp. 1129-1155, 1985.

[17] V. MEHRMANN AND H. XU, *An analysis of the pole palcement problem II. The multi-input case,*, Technical Report 97-14, Fakultät für Mathematik, TU Chemnitz-Zwickau, May 1997.

[18] V. MEHRMANN AND H. XU, *Choosing poles so that the single-input pole placement problem is well-conditioned,* SIAM J. Matr. Anal. Vol. 19, No. 3, pp. 644-681, 1998.

[19] N. M. NACHTIGAL, S. REDDY AND L. N. TREFETHEN *"How fast are nonsymmetric matrix iterations?"*, Proceedings of the Copper Mountain Conference on Iterative methods, Copper Mountain, Co, 1990.

[20] Y. SAAD AND M. H. SCHULTZ, *GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems,* SIAM J. Sci. Stat. Comp., 7 (1986), pp. 856-869.

[21] R. A. SMITH, *The condition numbers of the matrix eigenvalue problem,* Numer. Math., 10, pp. 232-240, 1967.

[22] L. N. TREFETHEN, *Pseudospectra of matrices,* in Numerical Analysis 1991, D. F. Griffiths and G. A. Watson, eds., Longman, Harlow, London, 1992.

[23] J. H. WILKINSON, The Algebraic Eigenvalue Problem, *Oxford University Press,* 1965.

[24] W. M. WONHAM, Linear Multivariable Control: A Geometric Approach, *Springer-Verlag, New-York*, 1979.

[25] I. ZABALLA, *Matrices with prescribed invariant factors,* Lin. Multilin. Alg., 27, pp. 325-343, 1990.

(continued from the back cover)

A403 Saara Hyvönen and Olavi Nevanlinna
Robust bounds for Krylov method, Nov 1998

A402 Saara Hyvönen
Growth of resolvents of certain infinite matrice, Nov 1998

A400 Seppo Hiltunen
Implicit functions from locally convex spaces to Banach spaces, Jan 1999

A399 Otso Ovaskainen
Asymptotic and Adaptive Approaches to thin Body Problems in Elasticity

A398 Jukka Liukkonen
Uniqueness of Electromagnetic Inversion by Local Surface Measurements, Aug 1998

A397 Jukka Tuomela
On the Numerical Solution of Involutive Ordinary Differential Systems, 1998

A396 Clement Ph., Gripenberg G. and Londen S-O
Hölder Regularity for a Linear Fractional Evolution Equation, 1998

A395 Matti Lassas and Erkki Somersalo
Analysis of the PML Equations in General Convex Geometry, 1998

A393 Jukka Tuomela and Teijo Arponen
On the numerical solution of involutive ordinary differential equation systems, 1998

A392 Hermann Brunner, Arvet Pedas, Gennadi Vainikko
The Piecewise Polynomial Collocation Method for Nonlinear Weakly Singular Volterra Equations, 1997

A391 Kari Eloranta
The bounded Eight-Vertex Model, 1997

A390 Kari Eloranta
Diamond Ice, 1997

HELSINKI UNIVERSITY OF TECHNOLOGY INSTITUTE OF MATHEMATICS
RESEARCH REPORTS

The list of reports is continued inside. Electronical versions of the reports are available at *http://www.math.hut.fi/reports/* .

A412 Marko Huhtanen
Ideal GMRES can be bounded from below by three factors, Jan 1999

A411 Juhani Pitkranta
The first locking-free plane-elastic finite element: historia mathematica, Jan 1999

A410 Kari Eloranta
Bounded Triangular and Kagomé Ice, Jan 1999

A408 Ville Turunen
Commutator Characterization of Periodic Pseudodifferential Operators, Dec 1998

ISBN 951-22-4405-5
ISSN 0784-3143
Edita, Espoo, 1999