Helsinki University of Technology Institute of Mathematics Research Reports
Teknillisen korkeakoulun matematiikan laitoksen tutkimusraporttisarja
Espoo 1999 A422

A MATRIX NEARNESS PROBLEM RELATED TO ITERATIVE METHODS

Marko Huhtanen

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

Marko Huhtanen: A matrix nearness problem related to iterative method ; Helsinki University of Technology Institute of Mathematics Research Reports A422 (1999).

Abstract: *We consider a matrix nearness problem arising from an analysis of the speed of convergence of GMRES for a linear system* $Ax = b$ *with* $A \in \mathbb{C}^{n \times n}$ *and* $b \in \mathbb{C}^n$*. More precisely, denoting by* $\mathcal{F}_k$ *the set of matrices of rank-k at most, we solve*

$$\min_{S \in \mathcal{S}, F_k \in \mathcal{F}_k} \|A - S - F_k\|,$$

where $\mathcal{S} \subset \mathbb{C}^{n \times n}$ *denotes the set matrices of the form* $e^{i\theta}H - \lambda I$*, with* $\theta \in [0, 2\pi)$*,* $\lambda \in \mathbb{C}$ *and* H *belongs to the set of Hermitian matrix. As to iterative methods, the set* $\mathcal{S}$ *is of interest in a larger context. To give an example, with a regular splitting* $A = S + F_k$ *of* A*, the system* $Ax = b$ *can be solved using a* $(k + 3)$*-term recurrence with inner-outer iterations.*

AMS subject classifications: 65F10, 15A23.

Keywords: Nearness problem, iterative methods, short term recurrence.

ISBN 951-22-4764-X
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

The convergence behavior of GMRES [17], with an invertible *normal* $A \in \mathbb{C}^{n\times n}$ and $b \in \mathbb{C}^n$, for solving the linear system

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

can be considered to be understood, see e.g. [4, 3, 14], whereas for a *nonnormal* A this is quite not so. In the latter case one alternative is to relate A to the set of normal matrices $\mathcal{N}$ in an appropriate way. One could, for instance, consider the distance of A to $\mathcal{N}$ and then take a closest normal matrix to A and "simulate" A with this normal matrix. This, however, does not seem to lead to any interesting conclusions. Instead, in a series of papers [13, 10, 12] different "rankwise" distances of A to $\mathcal{N}$ were introduced. A starting point for this approach was the following folklore among the scientific computing community: Small rank perturbations of the original matrix do not essentially alter the convergence behavior of GMRES as long as the condition number of the resulting system does not change significantly. Thus, because for normal matrices the convergence is understood, it is of interest to find out how close is A to $\mathcal{N}$ *modulo* small rank perturbations.

In this paper we consider a simplifed version of this problem. For that purpose, denote by $\mathcal{F}_k$ the set of matrices of rank-k at most, and, by $\mathcal{S} \subset \mathbb{C}^{n\times n}$ the set of matrices of the form $e^{i\theta}H - \lambda I$, with $\theta \in [0, 2\pi)$, $\lambda \in \mathbb{C}$ and where H belongs to the set of Hermitian matrices. With this notation, for a given $A \in \mathbb{C}^{n\times n}$, we look for

$$\min_{S\in\mathcal{S}, F_k\in\mathcal{F}_k} \|A - S - F_k\|, \tag{2}$$

for $k = 0, 1, ...n-1$. By $\|\cdot\|$ we denote the spectral norm throughout this paper. Obviously $\mathcal{S} \subset \mathcal{N}$ holds so that we get, via solving this problem, also un upper bound to the distance of A to $\mathcal{N}$ modulo $\mathcal{F}_k$.

Knowing the behavior of the quantities (2) is not only of use for understanding the convergence of GMRES better. Another reason for restricting, in particular, to $\mathcal{S}$ is that as a set it is extremely well-suited for all sorts of iterative computations. To give an example, for an invertible matrix belonging to $\mathcal{S}$ the corresponding linear system can be solved using a 3-term recurrence see [2] or [4][Exer. 6.4 in particular]. Using this, the problem (2) is decisive whenever one wants to solve (1) with a *minimum* amount of inner-outer iterations. For this purpose one needs to have, for as small k as possible, a given $A \in \mathbb{C}^{n\times n}$ decomposed as $A = S + F_k$ with $S \in \mathcal{S}$ and $F_k \in \mathcal{F}_k$. Namely, having $A = S + F_k$ with an invertible S, i.e., the splitting is *regular*, means that (1) is equivalent to solving

$$(I + S^{-1}F_k)x = S^{-1}b. \tag{3}$$

Of course S^{-1} is not computed in practise. Instead, the corresponding linear systems are solved iteratively which gives rise to the inner-iteration loop. As

$S \in \mathcal{S}$, this can be done with a 3-term recurrence. Since $S^{-1}F_k$ is of rank-k only, then, taking into account all the inner-outer iterations, only $(k+3)$-term recurrence is needed to solve (3).

Before solving (2) we consider the distance of a matrix $A \in \mathbb{C}^{n\times n}$ to the set of Hermitian matrices $\mathcal{H}$ modulo $\mathcal{F}_k$, that is,

$$\min_{M\in\mathcal{H}, F_k\in\mathcal{F}_k} \|A - M - F_k\|. \tag{4}$$

To this problem a trivial and *not* necessarily optimal approximative solution is to sum the Hermitian part of A with the closest rank-k approximation to the skew-Hermitian part of A. We show that to get an optimal solution, there is an exact condition on the eigenvalues of the skew-Hermitian part of A under which, roughly speaking, the approximation is *twice* better than the trivial one just described. These solutions are realizable in a numerically stable way.

Understanding the nearness problem (4) is the key to solving (2). First we demonstrate, with the help of (4), that any given matrix $A \in \mathbb{C}^{n\times n}$ can be decomposed as $A = S + F_k$ with $S \in \mathcal{S}$ *and* with $F_k \in \mathcal{F}_k$ for $k \leq [\frac{n}{2}]$. We find this a fairly good worst case bound as opposed to the trivial $k \leq n$. Second, we show that the decay of numbers (2) depends on a certain kind of *two-sided* convergence of the eigenvalues of $e^{i\theta}A - e^{-i\theta}A^*$ for $\theta \in [0, 2\pi)$, see Corollary 6 for the exact statement. These solutions are also realizable in a numerically stable manner yielding, consequently, a decomposition of A as $A = S + \Delta_k + F_k$ with $\|\Delta_k\| = \mu_{k+1}(A)$. Furthermore, sparse methods are extremely well suited for computing this decomposition. This is of particular interest since decomposing A in this manner is mainly motivated by large problems.

This paper is organized as follows. In Section 2 we introduce matrix nearness problems related to iterative methods. Then we characterize matrices solving two of them, namely (4) and (2). Also, we consider how the problem (4) is related to a number of issues in numerical linear algebra. In Section 3 we consider computational aspects related to (2). We end the paper with numerical examples.

2 A matrix nearness problem related to iterative methods

In the 90's a lot of work has been done in order to comprehend the convergence of GMRES for nonnormal matrices, see e.g. [4, 14] and references therein. Since the convergence of GMRES for normal matrices $\mathcal{N}$ is understood, a natural approach is to associate a given matrix $A \in \mathbb{C}^{n\times n}$ with $\mathcal{N}$ in some manner. A somewhat naive possibility is to find a closest normal matrix to A and then use this matrix for analyzing GMRES for A. Note that this approach *is* feasible as a closest normal matrix can be computed by using Ruhe's algorithm [16]. This approach does not seem to lead to any interesting conclusions, however.

In addition to understanding the convergence of GMRES for $\mathcal{N}$, empirically it is known that if A is perturbed by a small rank matrix F, then the convergence of GMRES for $Ax = b$ and $(A+F)y = b$, with $b \in \mathbb{C}^n$, is essentially similar provided the condition numbers of A and $A+F$ are close (i.e. if one system is ill-posed and the other one is not, then the convergence behavior cannot, of course, be the same anymore). Thus, combining these two phenomena, finding a normal matrix N that is close to A *modulo* small perturbations, yields a way to comprehend GMRES for a nonnormal A. Formulated as such we obtain a nearness problem

$$\mu_{k+1}(A) := \min_{N \in \mathcal{N}, F_k \in \mathcal{F}_k} \|A - N - F_k\|, \tag{5}$$

for $k = 0, 1, ..., n-1$, where $\mathcal{F}_k$ denotes the matrices of rank-k at most.

2.1 Distance of A to the Hermitian matrices modulo $\mathcal{F}_k$

Characterizing the matrices that realize the numbers $\mu_k(A)$ seems fairly difficult. Instead, we start with a nearness problem where N is constrained to belong to the set of Hermitian matrices $\mathcal{H} \subset \mathbb{C}^{n\times n}$. That is, we consider

$$\min_{M \in \mathcal{H}, F_k \in \mathcal{F}_k} \|A - M - F_k\|. \tag{6}$$

These quantities are readily approximable by using the Toeplitz decomposition $A = \frac{1}{2}(A + A^*) + \frac{1}{2}(A - A^*)$ of A. Let $\sigma_k(A)$ denote the k^{th} singular value of A.

Proposition 1 *Take $M = \frac{1}{2}(A + A^*)$ and F_k to be the closest rank-k approximation to $\frac{1}{2}(A - A^*)$ to have $\min_{M \in \mathcal{H}, F \in \mathcal{F}_k} \|A - M - F\| \leq \sigma_{k+1}(\frac{1}{2}(A - A^*))$.*

Since the closest Hermitian matrix to $B \in \mathbb{C}^{n\times n}$ is its Hermitian part $\frac{1}{2}(B + B^*)$, see e.g. [5], the nearness problem (6) is equivalent to

$$\min_{F_k \in \mathcal{F}_k} \|A - A^* - (F_k - F_k^*)\|. \tag{7}$$

Considering this formulation allows to demonstrate that taking F_k as Proposition 1 suggests does not need to yield an optimal solution.

Theorem 2 *Let $M \in \mathbb{C}^{n\times n}$ be skew-Hermitian and of rank-2 with nonzero eigenvalues λ_1 and λ_2. Then there exists $u, v \in \mathbb{C}^n$ such that $uv^* - vu^* = M$ if and only if $\lambda_1 \lambda_2 > 0$.*

Proof. Assume first that there exists $u, v \in \mathbb{C}^n$ such that $uv^* - vu^* = M$. As the rank of M is nonzero, u and v are linearly independent. Since span$\{u, v\}$ is an invariant subspace of $uv^* - vu^*$, we consider the restriction of $uv^* - vu^*$ on this subspace. Since the range of $uv^* - vu^*$ belongs span$\{u, v\}$, we notice that this restriction has eigenvalues λ_1 and λ_2. This restriction is represented by

$wz^* - zw^*$, for some $w, z \in \mathbb{C}^2$, which can be seen by applying an orthogonal projector onto span$\{u, v\}$ and the resulting matrix representation. Thus, with $w = [\bar{w}_1\ \bar{w}_2]^*$ and $z = [\bar{z}_1\ \bar{z}_2]^*$

$$wz^* - zw^* = \begin{bmatrix} w_1\bar{z}_1 - z_1\bar{w}_1 & w_1\bar{z}_2 - z_1\bar{w}_2 \\ w_2\bar{z}_1 - z_2\bar{w}_1 & w_2\bar{z}_2 - z_2\bar{w}_2 \end{bmatrix}.$$

In particular, using the spectral projector yielding a diagonalization gives us the condition $w_2\bar{z}_1 - z_2\bar{w}_1 = 0$, or $w_2 = \frac{z_2}{\bar{z}_1}\bar{w}_1$. Here we use the fact that $w_1 \neq 0$ which must hold since λ_1 and λ_2 are nonzero. After substitution we obtain

$$w_2\bar{z}_2 - z_2\bar{w}_2 = \frac{|z_2|^2}{|z_1|^2}(z_1\bar{w}_1 - \bar{z}_1 w_1), \tag{8}$$

so that, after dividing by i, we have opposite sign for (8) from that of $\frac{1}{i}(w_1\bar{z}_1 - z_1\bar{w}_1)$.

For the converse, assume M is a skew-Hermitian matrix of rank-2 that has two eigenvalues that fulfill the condition $\lambda_1\lambda_2 > 0$ and consider the problem of finding two vectors $u, v \in \mathbb{C}^n$ yielding $uv^* - vu^* = M$. We solve the 2×2 case first as the general case follows from that immediately after considering the restriction of M to the corresponding spectral subspace. For that purpose, with a diagonalization $\Lambda = WMW^*$ of M, the governing equations become

$$\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = W^*uv^*W - W^*vu^*W, \tag{9}$$

where $\frac{\lambda_1}{i} < 0$ for instance (if this is not the case, then multiply both sides of (9) by $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ from the left and right and proceed analogously). Denoting by

$$W^*u = \begin{bmatrix} x_1 + iy_1 \\ x_2 + iy_2 \end{bmatrix} \text{ and } W^*v = \begin{bmatrix} a_1 + ib_1 \\ a_2 + ib_2 \end{bmatrix},$$

we get 4 conditions from (9), namely $2i(y_1a_1 - x_1b_1) = \lambda_1$, $2i(x_2a_2 - y_2b_2) = \lambda_2$, $y_2a_1 + x_2b_1 - a_2x_1 - b_2y_1 = 0$ and $x_2a_1 - b_1y_2 - b_2x_1 + a_2y_1 = 0$. These equations are not uniquely solvable, but, for example, choices

$$x_1 = x_2 = b_1 = b_2 = 0,\ y_1 = 1,\ y_2 = \sqrt{\frac{-\lambda_2}{\lambda_1}},\ a_1 = \frac{\lambda_1}{2i},\ a_2 = \sqrt{\frac{\lambda_1\lambda_2}{4}} \tag{10}$$

yield a solution. Then applying W to W^*u and W^*v yields u and v.

If $M \in \mathbb{C}^{n\times n}$, then all the computations are repeated with a diagonalization $\Lambda = WMW^*$ taking $W_2 \in \mathbb{C}^{n\times 2}$ in (9) with orthonormal colunms corresponding to eigenvalues λ_1 and λ_2. $\square$

EXAMPLE 1. For an illustration, consider diagonal $M = \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix}$. Then with the choices (10) we have $u = \begin{bmatrix} i \\ i \end{bmatrix}$ and $v = \begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$. Thus,

$$uv^* - vu^* = \begin{bmatrix} -\frac{i}{2} & \frac{i}{2} \\ -\frac{i}{2} & \frac{i}{2} \end{bmatrix} - \begin{bmatrix} \frac{i}{2} & \frac{i}{2} \\ -\frac{i}{2} & -\frac{i}{2} \end{bmatrix} = M.$$

Note that the matrix uv^* itself is not skew-Hermitian.

REMARK 1. If the skew-Hermitian part of $A \in \mathbb{C}^{n\times n}$ has exactly two nonzero eigenvalues λ_1 and λ_2 such that $i\lambda_1$ and $i\lambda_2$ have different signs, then $A - uv^*$ is Hermitian with an appropriate choice of vectors $u, v \in \mathbb{C}^n$ yielding $uv^* - vu^* = A - A^*$. In particular, setting $H = A - uv^* \in \mathcal{H}$, we have $A = H + uv^*$.

REMARK 2. The matrix uv^* cannot be skew-Hermitian since $uv^* - vu^*$ would then be of rank-one. Consequently, the Hermitian parts of A and $A - uv^*$ will differ. The choice provided by Proposition 1 *is* skew-Hermitian.

REMARK 3. Note that the proof of Theorem 2 is constructive as well as amenable for practical computations, that is, it yields a *numerically stable* algorithm for finding vectors $u, v \in \mathbb{C}^n$ such that $M = uv^* - vu^*$. Note that the solution is *not* unique. The actual vectors u and v we provide correspond to the particular choices made in (10).

Recall that the inertia $I(A)$ of a matrix $A \in \mathbb{C}^{n\times n}$ is a vector with three components, where the first component equals the number of eigenvalues with positive real part, the second component equals the number of eigenvalues with negative real part and the third component equals the number of eigenvalues with zero real part (multiplicities are counted) [9].

Corollary 3 *Assume M is skew-Hermitian with $I(\frac{1}{i}M) = [l_1, l_2, n - l_1 - l_2]$, where $l_1 \geq l_2$. Then there exists a rank-l_2 matrix F_{l_2} with $I(\frac{1}{i}(M - (F_{l_2} - F_{l_2}^*))) = [l_1 - l_2, 0, n - l_1 + l_2]$.*

Proof. Collect pairwise $2l_2$ of the eigenvalues of $\frac{1}{i}M$ with negative and positive real parts (in any manner). Corresponding to each pair, construct vectors u_j and v_j, for $1 \leq j \leq l_2$, as in the proof Theorem 2 and then set $F_{l_2} = \sum_{j=1}^{l_2} u_j v_j^*$. $\square$

Obviously the same construction holds if $l_2 \geq l_1$. Furthermore, analogous claims hold for the case M is Hermitian.

Let $[r]$ denote the largest integer less than or equal to $r \in \mathbb{R}$. Further, for $k = 1, ..., n$, let $\lambda_k(M)$ denote the eigenvalues of a *Hermitian* matrix M arranged according to *increasing* size (counting multiplicities). As long as $k \leq [\frac{n}{2}]$, it is fairly easy to see that the rank of $UV^* - VU^*$ is generically $2k$ for a pair of matrices $U, V \in \mathbb{C}^{n\times k}$. In this case the positive-negative eigenvalue pairing of Theorem 2 continues up to $[\frac{n}{2}]$ in the following sense.

Theorem 4 *Assume $U, V \in \mathbb{C}^{n\times k}$, with $k \leq [\frac{n}{2}]$, are such that $UV^* - VU^*$ has rank $2k$. Then $I(\frac{1}{i}(UV^* - VU^*)) = [k, k, 0]$.*

Proof. By Theorem 2 the claim is true for $uv^* - vu^*$ with $u, v \in \mathbb{C}^n$. We use this with the inequalities of Weyl [8][Theorem 4.3.7] to proceed by induction. First we assume that $k = 2$ so that by the inequalities of Weyl

$$\lambda_{n-2}(\sum_{j=1}^{2} \frac{1}{i}(u_j v_j^* - v_j u_j^*)) \leq \lambda_{n-1}(\frac{1}{i}(u_1 v_1^* - v_1 u_1^*)) + \lambda_{n-1}(\frac{1}{i}(u_2 v_2^* - v_2 u_2^*)) = 0$$

and

$$0 = \lambda_2(\frac{1}{i}(u_1 v_1^* - v_1 u_1^*)) + \lambda_2(\frac{1}{i}(u_2 v_2^* - v_2 u_2^*)) \leq \lambda_3(\sum_{j=1}^{2} \frac{1}{i}(u_j v_j^* - v_j u_j^*)).$$

Thus, this forces $\lambda_3(\sum_{j=1}^{2} \frac{1}{i}(u_j v_j^* - v_j u_j^*)) = \cdots = \lambda_{n-2}(\sum_{j=1}^{2} \frac{1}{i}(u_j v_j^* - v_j u_j^*)) = 0$. As the rank is $2k = 4$ the remaining eigenvalues are necessarily nonzero and the claim follows. Then, if the claim is true for $s - 1$, we have again by the inequalities of Weyl

$$\lambda_{n-s}(\sum_{j=1}^{s} \frac{1}{i}(u_j v_j^* - v_j u_j^*)) \leq$$

$$\lambda_{n-(s-1)}(\sum_{j=1}^{s-1} \frac{1}{i}(u_j v_j^* - v_j u_j^*)) + \lambda_{n-1}(\frac{1}{i}(u_s v_s^* - v_s u_s^*)) = 0$$

and

$$0 = \lambda_2(\frac{1}{i}(u_s v_s^* - v_s u_s^*)) + \lambda_s(\sum_{j=1}^{s-1} \frac{1}{i}(u_j v_j^* - v_j u_j^*)) \leq \lambda_{s+1}(\sum_{j=1}^{s} \frac{1}{i}(u_j v_j^* - v_j u_j^*))$$

and the claim follows again by exactly similar arguments as in the case $k = 2$. $\square$

To sum up, we have a complete solution to (6) as follows.

Algorithm 1. ”For solving (6)”.

If $k \leq \min\{l_1, l_2\}$, where $I(\frac{1}{i}(A - A^*)) = [l_1, l_2, n - l_1 - l_2]$, choose $F_k \in \mathcal{F}_k$ in such a way that

$$I(i(A - A^* - (F_k - F_k^*))) = [l_1 - k, l_2 - k, n - l_1 - l_2 + 2k], \qquad (11)$$

with the property that the k largest eigenvalues in absolute value from both of the first two components of the inertia of $\frac{1}{i}(A - A^*)$ are removed. The matrix M is taken to be the Hermitian part of $A - F_k$.

If $k > \min\{l_1, l_2\}$, for instance if $l_1 \geq l_2$, then take F_{l_2} for $A - A^*$ as in (11) above and take F_{k-l_2} as in Proposition 1 for $\frac{1}{2}\Big(A - A^* - (F_{k-l_2} - F_{k-l_2}^*)\Big)$. Then set $F_k = F_{l_2} + F_{k-l_2}$. The matrix M is taken to be the Hermitian part of $A - F_k$.

Let us illustrate the above claims with an example (using Matlab).

EXAMPLE 2. Take $A \in \mathbb{C}^{9\times 9}$ to be the nilpotent forward shift. One easily verifies that $I(\frac{1}{i}(A - A^*)) = [4, 4, 1]$. Computing the rank-one corrections with the choices (10) yields a Hermitian

$$A - \sum_{j=1}^{4} u_j v_j^* = \begin{bmatrix} -0.3775 & 0.5693 & 0.1706 & -0.1115 & 0.1902 & 0.1119 & -0.0530 & -0.0691 & 0.0697 \\ 0.5693 & 0.5481 & 0.3191 & 0.0196 & 0.2235 & -0.2432 & -0.1810 & 0.1227 & 0.0691 \\ 0.1706 & 0.3191 & -0.3579 & 0.7928 & -0.0727 & -0.2926 & 0.3130 & 0.1810 & -0.0530 \\ -0.1115 & 0.0196 & 0.7928 & 0.3048 & 0.1381 & 0.1424 & 0.2926 & -0.2432 & -0.1119 \\ 0.1902 & 0.2235 & -0.0727 & 0.1381 & -0.2351 & 0.8619 & -0.0727 & -0.2235 & 0.1902 \\ 0.1119 & -0.2432 & -0.2926 & 0.1424 & 0.8619 & 0.3048 & 0.2072 & 0.0196 & 0.1115 \\ -0.0530 & -0.1810 & 0.3130 & 0.2926 & -0.0727 & 0.2072 & -0.3579 & 0.6809 & 0.1706 \\ -0.0691 & 0.1227 & 0.1810 & -0.2432 & -0.2235 & 0.0196 & 0.6809 & 0.5481 & 0.4307 \\ 0.0697 & 0.0691 & -0.0530 & -0.1119 & 0.1902 & 0.1115 & 0.1706 & 0.4307 & -0.3775 \end{bmatrix}.$$

The quantities (6) are 0.9511, 0.8090, 0.5878, 0.3090 and 0 for $k = 0, ..., 4$.

2.2 Finding $\min \|A - S - F_k\|$ and its applications

With the help of a complete solution to (6) we consider a simplified version of the nearness problem (5). For that purpose, we denote by $\mathcal{S} \subset \mathbb{C}^{n\times n}$ the set of matrices of the form $e^{\theta i}H - \lambda I$, with $\theta \in [0, 2\pi)$, $\lambda \in \mathbb{C}$ and $H \in \mathcal{H}$. Obviously $\mathcal{S}$ is a subset of $\mathcal{N}$. With this notation, for a given $A \in \mathbb{C}^{n\times n}$, set

$$\nu_{k+1}(A) := \min_{S\in\mathcal{S}, F_k\in\mathcal{F}_k} \|A - S - F_k\|, \tag{12}$$

for $k = 0, 1, ..., n - 1$.

As $\mathcal{S} \subset \mathcal{N}$, un upper bound to $\mu_k(A)$ is obtained via solving (12). However, replacing $\mathcal{N}$ with $\mathcal{S}$ is not solely of interest for understanding (5) better and thereby GMRES for A. Rather, as a set $\mathcal{S}$ is extremely well suited for all kinds of iterative computations. First, a linear systems $Sx = b$ with an invertible $S \in \mathcal{S}$ can be solved using a 3-term recurrence [2]. Second, the Hessenberg matrix resulting from the Arnoldi method for $S \in \mathcal{S}$ is tridiagonal (and thereby can also be computed with a 3-term recurrence). Third, eigenapproximations with the Arnoldi method for matrices from $\mathcal{S}$ are structure preserving. Namely, truncated Hessenberg matrices also remain in $\mathcal{S}$ with the corresponding, smaller, dimension of course.

With $\mathcal{S}$ it is possible to demonstrate that $\mu_k(A) = 0$, for $k \geq [\frac{n}{2}] + 1$.

Theorem 5 *Assume $A \in \mathbb{C}^{n\times n}$. Then there exists $F \in \mathcal{F}_{[\frac{n}{2}]}$ and $S \in \mathcal{S}$ such that $A = S + F$. Or, equivalently, $\nu_{[\frac{n}{2}]+1}(A) = 0$.*

Proof. The parameter θ is in fact redundant and, consequently, we set $\theta = 0$ in our matrix $e^{\theta i}H - \lambda I \in \mathcal{S}$ that follows.

Assume first that $A - A^*$ has n distinct eigenvalues. Choose a purely imaginary λ such that for the inertia $I\Big(\frac{1}{i}((A-\lambda I)-(A-\lambda I)^*)\Big) = ([\frac{n}{2}], [\frac{n}{2}], a)$ holds with

$$a = \begin{cases} 0 & n \text{ is even} \\ 1 & n \text{ is odd} \end{cases}.$$

Now taking $F = UV^*$ of rank $[\frac{n}{2}]$ for $(A - \lambda I) - (A - \lambda I)^*$ as in Corollary 3 we have for $H = A - \lambda I - F \in \mathcal{H}$ equality $A = H + \lambda I + F$ with desired properties.

Assume now that $\frac{1}{i}(A - A^*)$ has less than n distinct eigenvalues, say p. Order these eigenvalues $\lambda_1 < \lambda_2 < ... < \lambda_p$ and let $l_1, ..., l_p$ denote their multiplicities. Let s be the largest index for which there still holds

$$\sum_{k=1}^{s} l_k \leq [\frac{n}{2}]. \tag{13}$$

Now choose λ such that the eigenvalue λ_{s+1} is translated to zero so that we have $I(\frac{1}{i}((A - \lambda I) - (A - \lambda I)^*)) = [\sum_{k=1}^{s} l_k, \sum_{k=s+2}^{p} l_k, l_{s+1}]$. Now, for $j = |\sum_{k=1}^{s} l_k - \sum_{k=s+2}^{p} l_k|$ eigenvalues of $\frac{1}{i}((A - \lambda I) - (A - \lambda I)^*)$ with the same sign, we take F_1 of rank-j for $(A - \lambda I) - (A - \lambda I)^*$ as in Proposition 1 to get rid of those. For the remaining nonzero eigenvalues we choose F_2 as in the first part of the proof as follows. Without loss of generality, assume that $\sum_{k=1}^{s} l_k - \sum_{k=s+2}^{p} l_k \geq 0$ (or order the eigenvalues other way around), so that then $l_{s+1} \geq j$, otherwise (13) does not hold. Consequently, we only need match at most $n - 2j$ eigenvalues with F_2. Thus, rank of F_2 is at most $[\frac{n-2j}{2}] = [\frac{n}{2}] - j$ and the claim follows. □

Let $\sigma(M)$ denote the spectrum of a square matrix M.

Corollary 6 *For $M_\theta := \frac{1}{i}(e^{i\theta}A - e^{-i\theta}A^*)$ choose $\theta \in [0, 2\pi)$ such that the diameter of $\sigma(M_\theta)\backslash\{\lambda_1(M_\theta), ..., \lambda_k(M_\theta), \lambda_{n-k+1}(M_\theta), ..., \lambda_n(M_\theta)\}$ attains minimum. Then, with $\lambda_\theta = \frac{i}{2}(\lambda_{k+1}(M_\theta) + \lambda_{n-k}(M_\theta))$,*

$$\nu_{k+1}(A) = \|A - e^{-i\theta}H_\theta - \frac{e^{-i\theta}}{2}\lambda_\theta I - e^{-i\theta}F_k\|,$$

where $F_k \in \mathcal{F}_k$ is taken as in (11) for $e^{i\theta}A - e^{-i\theta}A^* - \lambda_\theta I$ and H_θ is the Hermitian part of $e^{i\theta}A - F_\theta$.

Proof. Since there holds

$$\min_{S\in\mathcal{S}, F_k\in\mathcal{F}_k} \|A - S - F_k\| = \min_{\theta\in[0,2\pi), H\in\mathcal{H}, \lambda\in\mathbb{C}, F_k\in\mathcal{F}_k} \|e^{i\theta}A - H - \lambda I - F_k\|. \tag{14}$$

the claim follows. □

REMARK 4. Although the parameter θ was redundant for the proof of Theorem 5, it is indispensable in Corollary 6 for getting optimal approximations to A from $\mathcal{S}$.

REMARK 5. The matrices $e^{i\theta}A - e^{-i\theta}A^*$ arise when the field of values of A is being approximated (or, in fact, the field of values of iA which is, of course, equivalent). Namely then, for a finite number of different θ, one needs *only* the largest eigenvalue of $e^{i\theta}A - e^{-i\theta}A^*$ in order to intersect certain half-planes defined on the basis of these eigenvalues, see e.g. [9][Thm 1.5.12, 1.5.14]. Thus in a sense, for $\nu_k(A)$, more "structure" of the field of values is needed. The field of values has also been used to analyze iterative methods [1].

REMARK 6. The eigenvalues of A give *absolutely no* clue of the behavior of the numbers $\nu_k(A)$. That is, even in the simplest case when $B \in \mathbb{C}^{n\times n}$ is Hermitian and $u \in \mathbb{C}^n$ is such that $\mathcal{K}(B;u) = \text{span}\{u, Bu, ..., B^{n-1}u\} = \mathbb{C}^n$, then, with an appropriate choice of $v \in \mathbb{C}^n$, the spectrum of $A = B + uv^*$ can consist of *any* n complex numbers (multiplicities counted), see e.g. [18]. Of course already $\nu_2(A) = 0$.

REMARK 7. As a byproduct, we obtain an upper bound for the smallest rank F yielding a decomposition $A = N + F$ while N varies among $\mathcal{N}$. Since $\mathcal{N}$ is a stratified submanifold of (real) dimension $n^2 + n$ [11] and $\mathcal{F}_k$ can be shown to have (real) dimension $2k(2n-k)$, then solving somewhat naively $n^2 + n + 2k(2n-k) = 2n^2$, gives $k \approx 0.3n$ for the smaller value of k. We do not know if this value of k really sufficies to yield any matrix $A \in \mathbb{C}^{n\times n}$ decomposed as $A = N + F$.

EXAMPLE 3. The nilpotent froward shift A of Example 2 is our standard example of a matrix that is close to $\mathcal{N}$ modulo only $\mathcal{F}_1$ [13, 10]. Namely, it is rank-one far from a unitary matrix which can be seen by replacing the zero element in the righ upper-corner of A by 1. Thus, $\mathcal{S}$ can be a poor substitute for $\mathcal{N}$ in this respect as a corrrection having rank $[\frac{n}{2}]$ is needed then instead.

Although there are examples in which $\mathcal{S}$ can be a poor subsitute for $\mathcal{N}$, there is a warning signal for that. More precisely, whenever A is a small rank perturbation of a normal matrix, it is revealed by the self-commutator $[A, A^*] = AA^* - A^*A$ of A as follows.

Proposition 7 *Suppose $A = N + F$ with $N \in \mathbb{C}^{n\times n}$ normal. Then*

$$\text{rank}([A, A^*]) \leq 4 \text{ rank}(F). \tag{15}$$

Proof. A simple computation yields, since N is normal,

$$[A, A^*] = FA^* - A^*F + NF^* - F^*N = [F, A^*] + [N, F^*], \tag{16}$$

so that $\text{rank}([A, A^*]) \leq \text{rank}([F, A^*]) + \text{rank}([N, F^*])$. Since $\text{rank}([F, A^*]) \leq 2\text{rank}(F)$ and $\text{rank}([N, F^*]) \leq 2\text{rank}(F)$, the claim follows. $\square$

Extracting as much information as possible from Corollary 6 yields the following improved normal approximant based on the fact that $p(H)$ is normal for $H \in \mathcal{H}$ and $p \in \mathcal{P}$. By $\mathcal{P}$ we denote the set of polynomials.

Proposition 8 *Let H_θ and F_θ be as in Corollary 6. Then*

$$\mu_{k+1}(A) \leq \min_{p \in \mathcal{P}} \|e^{i\theta}A - F_\theta - p(H_\theta)\| \leq \nu_{k+1}(A).$$

Proof. The first inequality follows from noticing that $p(H_\theta)$ is normal for any polynomial p. The second inequality is obvious. □

The behavior of approximation numbers $\nu_k(A)$ is of interest for the following reason. Recall that a splitting $A = M + N$ of A is said to be regular whenever M is invertible.

Theorem 9 *Assume $A \in \mathbb{C}^{n\times n}$ and $\nu_{k+1}(A) = 0$ for a $k \geq 0$ such that the corresponding splitting $A = S + F_k$ is regular. Then there exists a $(k+3)$-term recurrence for solving $Ax = b$ for any $b \in \mathbb{C}^n$.*

Proof. This involves an inner-outer iteration. That is, $Ax = b$ is equivalent to solving $(I + S^{-1}F_k)x = S^{-1}b$. The vector $S^{-1}b$ can be solved using a 3-term recurrence [2] and thereafter b can be replaced with $S^{-1}b$. Since the rank of $S^{-1}F_k$ is at most k, the Krylov subspace $\mathcal{K}_n(I + S^{-1}F_k; S^{-1}b)$ is at most $k+1$ dimensional. Thus for solving $(I + S^{-1}F_k)y = c$, for any $c \in \mathbb{C}^n$, at most $k+1$ vectors are needed to be restored. For computing the $(k+1)^{th}$ vector the recursion involves 3 vectors in the inner loop and k vectors in the outer loop. This is the maximum number of vectors that one has since once the $(k+1)^{th}$ vector in outer iteration is obtained, no 3-term inner iterations are needed. □

Of course a 3-term recurrence is always obtained by considering the normal equations instead [4]. But, as it is well-known, this may not be a good alternative as the resulting system can become very ill-conditioned. As such the problem does not exist with a splitting $A = S + F_k$ of A since, at any stage, no multiplications of A by A^* are performed. That is, S and F_k are computed, in essence, by summing A and A^*. Clearly it is possible, as always with splitting methods, that the computed splitting is not regular or nearly not so. But the worst case behavior of the normal equations cannot happen with $A = S + F_k$ because of the fundamental difference between sum and multiplication.

A general form of the problem for mimimal term recurrence (with inner-outer iterations) is obtained from the result of Faber and Manteuffel [2].

Problem. For a given $k \geq 0$, find q of least degree such that

$$q(A - F_k) = A^* - F_k^* \tag{17}$$

holds for a $F_k \in \mathcal{F}_k$.

Of course, for small $k \geq 0$, it may not be possible to fing any polynomial q fulfilling (17). Namely, if (17) holds, then $A - F_k$ is normal. Consequently, the behavior of the quantities $\mu_k(A)$ is decisive for the problem (17).

Finally a word about when, for some small k, the approximation number (12) is small but not necessarilly zero. In other words, we have $A = S +$

$\Delta_k + F_k$ with $\|\Delta_k\| = \mu_{k+1}(A)$ small. As long as $\mu_{k+1}(A) > 0$, a $(k+3)$-term recurrence is not of use for solving $Ax = b$. However, since $\mu_{k+1}(A)$ is small for a small k, a Hermitian matrix can be regarded as a governing component of A in the sense that $S_k + \Delta_k$ is close to a matrix $e^{i\theta}H + \lambda I \in \mathcal{S}$ which is computable in practise. As a consequence, it is possible to utilize several preconditioning strategies, like polynomial preconditioning, for linear systems involving Hermitian matrices by adjusting them to apply to $e^{i\theta}H + \lambda I$ in an obvious way. Since these approaches posess some sort of continuity (depending on the method, of course), it is very reasonable to use these methods with a "nearly" shifted and translated Hermitian matrix $S + \Delta_k$. For the original system $(S + \Delta_k + F_k)x = b$ such a preconditioner is a good alternative as well, as long as k is small.

3 Computational aspects for $\min \|A - S - F_k\|$

The above described nearness problems have several features in common with the singular values and SVD of $A \in \mathbb{C}^{n\times n}$. That is, the singular values are also obtained from a nearness problem where there is just one component, namely $\mathcal{F}_k$, from which approximants to A are constructed. Also, the process yields a decomposition of A where the second part is from $\mathcal{F}_k$ and the first part is the remainder. In practise the reminder is typically discarded. Furthermore, the behavior of singular values of A is equivalent to the decay of the eigenvalues of AA^*. As the eigenvalues of AA^* are all positive, the decay takes place from the right towards the origin in the complex plane.

As opposed to singular values, the decay of numbers $\nu_k(A)$ depends on the "two-sided" convergence of the eigenvalues of $e^{i\theta}A - e^{-i\theta}A^*$ with $\theta \in [0, 2\pi)$. More precisely, according to Corollary 6, for a given $k \geq 0$, we need a value of $\theta \in [0, 2\pi)$ that makes the spectrum of $e^{i\theta}A - e^{-i\theta}A^*$ as concentrated as possible after removing k eigenvalues away from both ends (on the imaginary axis). The resulting composition has two parts where the first part is (hopefully) close to $\mathcal{S}$ and the second part is from $\mathcal{F}_k$. In this decomposition the first part is, by no means, meant to be discarded but to be used in various possible ways!

In what follows we use a straightforward approach to compute $\nu_{k+1}(A)$. First, for finding a good value for $\theta \in [0, 2\pi)$, we compute the spectra of $e^{i\theta_j}A - e^{-i\theta_j}A^*$ with some θ_j, with $j = 1, ..., K$. Then, once a value for θ_j is chosen, according to Corollary 6, we use Algorithm 1 to compute H_{θ_j}.

This approach may sound very expensive, in particular for large systems. And, it would be absurd to consider this approach in practise, for instance, for solving a linear system, if the complexity of finding a decomposition $A = S + F_k$, assuming it exists for some small k, drastically exceeded the complexity of solving $Ax = b$. However, for the only expensive step it is very natural to use sparse techniques. That is, as only the k rightmost and leftmost eigenvalues (i.e. extreme eigenvalues) of $\frac{1}{i}(e^{i\theta}A - e^{-i\theta}A^*)$ are being monitored for $\nu_k(A)$, sparse techniques are perfectly suited for finding a θ_j.

And, as each $\frac{1}{i}(e^{i\theta_j}A - e^{-i\theta_j}A^*)$ is Hermitian, this computational task is obviously much "easier" than dealing with the non-Hermitian A.

We consider a well-known example arising from a second order ODE. In this case the loss of normality is caused by the way the boundary conditions are set. This is not at all unusual, the same can happen also in very realistic PDE approximations, see for instance [6, 7]. The dimension and chosen discretization can play a crucial role then.

EXAMPLE 4. We consider a completely tractable integral equation

$$(I - \lambda V^2)u := u(t) - \lambda \int_0^t (t-s)u(s)ds, \tag{18}$$

with a nonzero $\lambda \in \mathbb{C}$, on $L^2([0,\pi])$. To V^2 there corresponds an initial value problem, see [15][p.130]. As V^2 is nilpotent, $I - \lambda V^2$ is invertible. We discretize (18) using a projection method with the (computationally convenient) orthonormal basis $\{\sqrt{\frac{2}{\pi}}\sin nt\}_{n=1}^{\infty}$ of $L^2([0,\pi])$. We have $\int_0^t (t-s)\sin(ks)ds = \frac{t}{k} - \frac{1}{k^2}\sin(kt)$ and $\int_0^\pi s\sin(ks)ds = \frac{-\pi\cos(k\pi)}{k}$. These computations yield a discretization

$$(V^2)_k := \begin{bmatrix} \frac{1}{1^2} & \frac{2}{1*2} & \frac{2}{1*3} & \cdots & \frac{2}{1*k} \\ \frac{-2}{2*1} & \frac{-3}{2^2} & \frac{-2}{2*3} & \cdots & \frac{-2}{2*k} \\ \frac{2}{3*1} & \frac{2}{3*2} & \frac{1}{3^2} & \cdots & \frac{2}{3*k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{-2}{k*1} & \frac{-2}{k*2} & \frac{-2}{k*3} & \cdots & \frac{-3}{k^2} \end{bmatrix} \in \mathbb{C}^{k\times k}$$

corresponding to the projector of rank k. Now, the rank of $(V^2)_k - (V^2)_k^*$ is 2, regardless of the value of $k \geq 4$. Computing u and v using the construction of Theorem 2 results in a Hermitian $(V^2)_k - uv^*$. In fact, V^2 is a rank-one correction of an operator that is Hermitian and has, in the basis $\{\sqrt{\frac{2}{\pi}}\sin nt\}_{n=1}^{\infty}$, a diagonal matrix representation [15]. Still, $(V^2)_k - uv^*$ is *not* diagonal as the rank-one correction uv^* of Theorem 2 is not unique. Thus, there is a rank-one correction that would make $(V)_k^2$ also diagonal, but that would correspond to different choices from those made in (10).

Now, to see how the spectra of $e^{i\theta}A - e^{-i\theta}A^*$ behave while θ varies, assume that our original integral operator is $I - e^{i\pi/4}V^2$. To get an illustration, we take $k = 200$ and $\theta_j = \frac{-j^2}{1+j^2}\frac{\pi}{4}$ for $j = 1, ..., 4$ so that $\theta_1 = -\frac{1}{2}\frac{\pi}{4}$ is the farest from and $\theta_4 = -\frac{16}{17}\frac{\pi}{4}$ is the closest to the "correct" value $-\frac{\pi}{4}$. The resulting spectra are depicted in Figure 3.1. The computations were performed with Matlab. As we see, already for θ_1 the spectrum is very well concentrated except for those two eigenvalues that would eventually be corrected with uv^*.

4 Conclusions

In this paper we have considered matrix nearness problems related to iterative methods. In all cases nearness is measured from a given set modulo small rank matrices. We have provided a complete solution to $\min_{S\in\mathcal{S},F_k\in\mathcal{F}_k} \|A -$

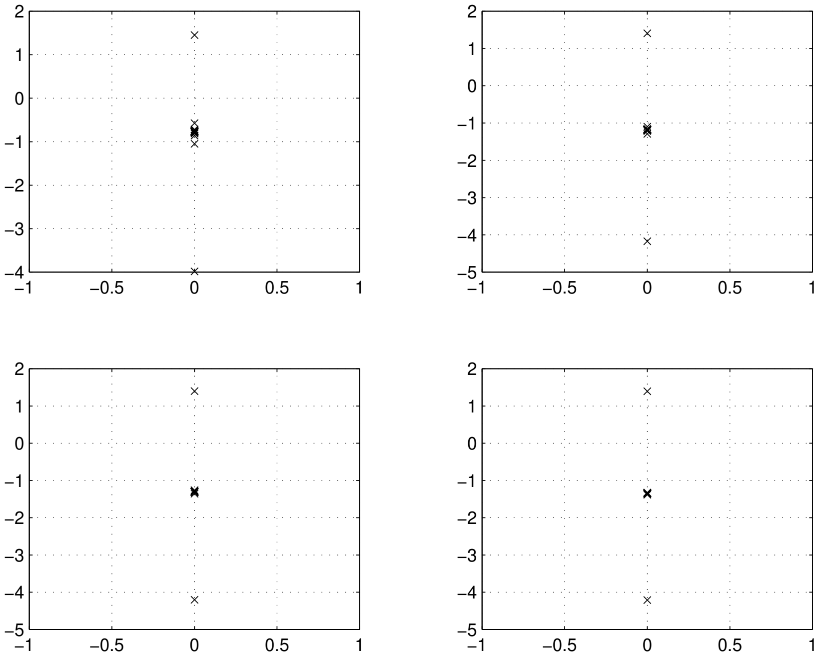

Figure 1: For Example 3 the spectra of $e^{i\theta_j}A - e^{-i\theta_j}A^*$ for $j = 1, ..., 4$.

$S - F_k\|$ which is of particular interest from several points of views as to iterative computations. We have shown that the decay of these numbers, while k grows, depends on two-sided convergence of the eigenvalues of $e^{i\theta}A - e^{-i\theta}A^*$ with $\theta \in [0, 2\pi)$, attaining zero for $k \leq [\frac{n}{2}]$. The solutions are computationally feasible and we have demonstrated the results with simple but illustrative numerical examples.

References

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

[2] V. FABER AND T. MANTEUFFEL, *Necessary and sufficient conditions for the existence of a conjugate gradiante method*, SIAM J. Numer. Anal., Vol. 21, No.2, 1984.

[3] A. GREENBAUM AND Z. STRAKOS, *Matrices that generate same Krylov residual spaces,* Recent Advances in Iterative Methods, G. Golub, A. Greenbaum and M. Luskin eds. IMA volumes in mathematics and its applications, Vol 60, Springer-Verlag, New York, 1994.

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

[5] N.J. HIGHAM *Matrix nearness problems and applications,* Inst. Math. Appl. Conf. Ser. New. Ser., 22, Oxford Univ. Press, New York, 1988.

[6] S. HOLMGREN AND K. OTTO *A framework for polynomial preconditioners based on fast transforms II: PDE applications,* Report 168 Dept. of Scientific Computing, Uppsala University, Sweden 1995.

[7] S. HOLMGREN AND K. OTTO *A framework for polynomial preconditioners based on fast transforms II: PDE applications,* BIT, Vol. 38, No. 4, pp. 721-736, 1998.

[8] R.A. HORN AND C.R. JOHNSON, Matrix Analysis, *Cambridge Univ. Press* 1985.

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

[10] M. HUHTANEN, *Ideal GMRES can be bounded from below by three factors,* Helsinki Univ. Tech. Math. Rep., A412, 1999. Also to appear in Lin. Alg. Appl.

[11] M. HUHTANEN, *A stratification of the set of normal matrices,* Helsinki Univ. Tech. Math. Rep., A414, 1999.

[12] M. HUHTANEN, *Pole assignment problems for error bounds for GMRES*, Helsinki Univ. Tech. Math. Rep., A418, 1999.

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

[14] N.M. NACHTIGAL, S. REDDY AND L.N. TREFETHEN *"How fast are nonsymmetric matrix iterations?"*, SIAM J. Matrix Anal. Appl. 13, No. 3, pp. 778-795, 1992.

[15] O. NEVANLINNA, Convergence of Iterations for Linear Equations, *Lectures in Mathematics ETH Zürich, Birkäuser Verlag*, 1993.

[16] A. RUHE, *Closest normal matrix finally found!,* BIT 27, (1987), pp. 585-598.

[17] 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.

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

This is a **draft**.

Please, use the latex in the linux-machines to obtain the final version and to add the list of earlier reports.

HELSINKI UNIVERSITY OF TECHNOLOGY INSTITUTE OF MATHEMATICS
RESEARCH REPORTS

This is a **draft**.
Please, use the latex in the linux-machines to obtain
the final version and to add the list of earlier reports.

ISBN 951-22-4764-X
ISSN 0784-3143
Edita, Espoo, 1999