ON SHORT RECURRENCES IN OPTIMAL KRYLOV SUBSPACE SOLVERS: THE FABER-MANTEUFFEL THEOREM AND EXTENSIONS

CASPER H. L. BEENTJES

Abstract. To solve large sparse linear systems of equations computational fast methods are preferable above methods such as GMRES. By exploiting structure of the involved matrix one can in certain cases use short recurrences to create an optimal Krylov subspace method. In this report we explain some of the theory regarding short-recurrence methods for solving linear systems. A central object in this theory is the Faber-Manteuffel theorem of which some extensions are given in this report as well. After the theory two actual algorithms which use these short-recurrences are discussed and tested, SUMR for shifted and scaled unitary matrices and PGMRES for nearly Hermitian matrices. Some theory is given for both methods and they are numerically tested against alternative methods.

1. Introduction

When attempting to solve a large linear system $Ax = b$ with $x, b \in \mathbb{C}^n$ and $A \in \mathbb{C}^{n\times n}$ a large sparse matrix one cannot hope of finding a solution by direct inversion of the matrix A. Instead one often tries to solve the system by using iterative methods to generate better and better approximations of the real solution. One special class of iterative methods uses so called Krylov subspaces

$$\mathcal{K}_m(A, v) = \text{span}\{v, Av, .., A^{m-1}v\}$$

to find approximations to the solution x. For obvious reasons this class of methods goes by the name of Krylov methods. In order to find a well-conditioned basis for the Krylov subspaces one often needs to perform orthogonalisation of the basis vectors. There are many algorithms known based on this procedure, of which the best known is probably Generalised Minimal Residual (GMRES) by Saad and Schultz [13]. One of the disadvantages of many of these methods (including GMRES) is that during the iterations the Krylov spaces grow and therefore a growing number of vectors has to be stored. Moreover, as the Krylov subspaces grow more and more basis vectors need to be orthogonalised and this can become computationally very costly.

Some matrices have additional structure which can be exploited in order to reduce the amount of work to construct or orthogonalise the Krylov spaces. For example if the matrix A is Hermitian then it can be proven that only 3 vectors in the Krylov subspaces need to be stored at each iteration resulting in a dramatic speed-up of the overall process. The best known method which is based on this result is the Conjugate Gradient method (CG) by Hesteness and Stiefel [9]. Since this method uses in each iteration only 3 vectors it is said to have short recurrence relations. It is now a natural question to ask if there are more 'CG-like' methods which are both based on short recurrences and are able to find an optimal approximate solution to x. Indeed Gene Golub posed this question in 1981 as a \$500 prize *"for the construction of a 3-term conjugate gradient like descent method for non-symmetric real matrices or a proof that there can be no such method."* In 1984 Faber and Manteuffel proved that in general there exists no CG-like method [6], a result now known as the Faber-Manteuffel theorem. This momentarily stopped the research in short recurrence methods for general matrices. However, in 1982 Gragg [8] proved that for another class of matrices, namely unitary matrices, it is possible to generate a Krylov space using short recurrences. On its own this is not a very practical result as linear systems with unitary matrices have a trivial solution involving the conjugate transpose. However, this result was then adapted by Jagels and Reichel in 1994 in a method (SUMR) for solving shifted unitary matrices using short recurrences [4]. This revived the research in the area of short recurrence optimal Krylov solvers

2010 *Mathematics Subject Classification.* 65F10, 65F15.

This text is based on a technical report for the 2013 Dutch Mastermath programme course Numerical Linear Algebra lectured by dr. G.L.G. Sleijpen.

resulting in a paper by Barth and Manteuffel in 2000 in which they describe a more general class of matrices for which short recurrence methods exist [3].

This report is organised as follows. In section 2 a review is given of Krylov subspace methods and optimality criteria for solving linear systems. Section 3 highlights results leading to the Faber-Manteuffel theorem and eventually the Faber-Manteuffel itself. The original proof of the theorem consists of a quite elementary 'sufficiency' part and a very hard 'necessity' part. In this paper we will proof the sufficient conditions following lines of the original arguments and adaptations from [12] but only shortly comment on the original proof and an adaptation for the necessity part following a paper by Faber, Liesen and Tichý [7]. Section 4 then highlights some results by Barth and Manteuffel which extend the Faber-Manteuffel theorem in the sense that they provide a larger class of matrices which allow short recurrence methods. In section 5 and 6 we focus on two short recurrence methods which are applicable to matrices with a structure different from Hermitian. Section 5 describes the SUMR algorithm for shifted unitary matrices and section 6 describes an algorithm for nearly Hermitian matrices. These sections numerically explore these two methods as well and comment on its strong and weak points.

2. Optimal Krylov subspace methods

As in the introduction let $A \in \mathbb{C}^{n\times n}$ and $x, b \in \mathbb{C}^n$ and let furthermore an inner product $(\cdot,\cdot)$ be given. Then the matrix A^* is defined as the unique matrix such that for all $x, y \in \mathbb{C}^n$ we have $(x, Ay) = (A^*x, y)$ and is called the adjoint of A.

Suppose we want to solve the system of linear equations:

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

An iterative method computes in every step m an approximation x_m to the solution x of (1). Ideally we want the approximation to improve every iteration. One way of measuring this is by use of the residual $r_m = b - Ax_m$. If a method minimises the residual in some norm (may be induced by the inner product or a different norm) over a given subspace $V \subset \mathbb{C}^{n\times n}$ we call the method an optimal method[1].

In general finding the x_m can be done by looking for x_m in a search space K_m which increases in every iteration, i.e. $K_{m+1} \supseteq K_m$, and which must be build iteratively as well. For reasons we will soon address a very common and useful sequence of search spaces is that of the Krylov subspaces, which are defined by a starting vector $v \in \mathbb{C}^n$

$$\mathcal{K}_m(A, v) = \text{span}\{v, Av, .., A^{m-1}v\}. \tag{2}$$

Now we note that Krylov subspaces are in particular shift and scaling invariant, i.e. $\mathcal{K}_m(A, v) = \mathcal{K}_m(aA + bI, v)$ for any $a, b \in \mathbb{C}$, a fact which will be useful in section 5. Another important observation is that the Krylov spaces cannot endlessly grow in size. Because n is finite there exists an integer $d = d(A, v)$ such that $A^d v$ is a linear combination of the linear independent vectors $v, Av, ..., A^{d-1}v$, i.e.

$$A^d v = \sum_{i=0}^{d-1} \alpha_i A^i v. \tag{3}$$

So we have that $\phi(A)v = 0$ with $\phi(\lambda) = \lambda^d - \sum_{i=0}^{d-1} \alpha_i \lambda^i$ a polynomial of degree d. This (monic) polynomial is called the minimal polynomial of v with respect to A and the degree d is called the grade of v. This generalises to a monic polynomial p for which A vanishes, $p(A) = 0$. This polynomial exists by the Cayley-Hamilton theorem and is equal to the least common multiple of minimal polynomials of $v \in V$ where V is a basis for $\mathbb{C}^{n\times n}$. Its degree $d_{\min}(A) = \deg(p)$ is called the degree of A and is always less than or equal to n by the Cayley-Hamilton theorem. A lemma from linear algebra [14] relates the degree of A to the number of distinct eigenvalues of A.

Lemma 1. *Let $A \in \mathbb{C}^{n\times n}$ and $d = d_{\min}(A)$, then A has at most d distinct eigenvalues. Furthermore, A has exactly d distinct eigenvalues if and only if A is diagonalisable.*

This will be useful later on. Using the minimal polynomial and its degree we can proof that $\mathcal{K}_d(A, v) = \mathcal{K}_m(A, v)$ for all $m \geq d$.

[1]Note that other optimal Krylov methods exist, which amongst others minimise the error $\|x - x_m\|$ every iteration.

Lemma 2. *Let $d = d_{\min}(A)$ and suppose A is non-singular, then $A^{-1} = q(A)$ with q a polynomial of degree $d-1$.*

Proof. Since $d = d_{\min}(A)$ there exists a polynomial of degree d such that $p(A) = 0$, without loss of generality we may assume that $p(\lambda) = 1 - \lambda q(\lambda)$ with q a polynomial of degree exact $d-1$. The polynomial p should have a constant term because otherwise zero would be a solution of $p(\lambda) = 0$ and thus would be an eigenvalue of A and therefore A would be singular. Furthermore, if it were smaller p would not be the minimal polynomial of A. As a result we derive that

$$0 = p(A) = I - Aq(A) \tag{4}$$

and we thus find $A^{-1} = q(A)$. □

Corollary 1. *Let $d = d_{\min}(A)$ and suppose $x_0 \in \mathbb{C}^n$ is given which initialises $r_0 = b - Ax_0$. Then the solution x to (1) is given by*

$$x = x_0 + q(A)r_0 \in x_0 + \mathcal{K}_d(A, r_0) \tag{5}$$

and furthermore $x \notin x_0 + \mathcal{K}_m(A, r_0)$ for $m < d$.

Proof. We know that $x = A^{-1}b$ is the solution of (1). This can be written as $x = x_0 + A^{-1}r_0$. Now using lemma 2 we have that $A^{-1} = q(A)$ with degree of q equal to $d-1$ and thus

$$A^{-1}r_0 = q(A)r_0 \in \mathcal{K}_d(A, r_0),$$

from which (5) follows. The last claim is a consequence of d being the grade of the minimal polynomial of A. □

This guarantees that if we iterate long enough (at most n times) we can always (at least in exact arithmetic) find the exact solution in the span of x_0 plus our Krylov spaces. The preceding illustrates the power of the Krylov subspaces, however there are some subtleties one has to consider. The first one is that if we iterate $A^m r_0$ we end up with a very ill conditioned basis for $\mathcal{K}_m(A, r_0)$, a result which can be most easily seen by relating the iterations to the power method. Since iterating A on r_0 produces vectors which will have a very strong component in the direction of the dominant eigenvalue of A the angle between the newest vectors in the Krylov subspace will approach zero thus resulting in a very ill conditioned basis. Of course one way to overcome this issue is to explicitly orthogonalise the newly created vector $A^m r_0$ with respect to the preceding basis vectors. If the orthogonalisation is done by using a Gram-Schmidt procedure starting from $v_1 = v/\|v\|$ and we perform on every iteration

$$\hat{v}_{j+1} = Av_j - \sum_{i=1}^{j} h_{j,i} v_i, \tag{6}$$

$$h_{j,i} = \frac{(Av_j, v_i)}{(v_i, v_i)}, \quad i = 1, \ldots, j, \tag{7}$$

$$h_{j+1,j} = \|\hat{v}_{j+1}\|, \tag{8}$$

$$v_{j+1} = \frac{\hat{v}_{j+1}}{h_{j+1,j}}, \tag{9}$$

we arrive at the well known Arnoldi method. It yields an orthonormal basis V_m consisting of the vectors $\{v_j\}_{j=1}^{m+1}$ for the Krylov subspace $\mathcal{K}_m(A, r_0)$ as well as the Arnoldi relation

$$AV_m = V_{m+1}\bar{H}_m, \tag{10}$$

where $\bar{H}_m \in \mathbb{C}^{m+1 \times m}$ an upper Hessenberg matrix of the form:

$$\bar{H}_m = \begin{pmatrix} h_{1,1} & \cdots & \cdots & h_{1,m} \\ h_{2,1} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & h_{m,m-1} & h_{m,m} \\ & & & h_{m+1,m} \end{pmatrix} = \begin{pmatrix} H_m \\ e_m^* \end{pmatrix}.$$

The Arnoldi method is one of the two basic ingredients for an optimal Krylov subspace method, it generates the basis for $\mathcal{K}_m(A, r_0)$. The other ingredient then is how to choose the optimal update

out of the Krylov subspace. Remember we want to have an optimal method and thus to minimise the residual $r_m = b - Ax_m$ in some norm, we are thus looking for

$$\underset{x_m \in x_0 + \mathcal{K}_m(A,r_0)}{\arg\min} \|b - Ax_m\| = \underset{y_m \in \mathcal{K}_m(A,r_0)}{\arg\min} \|r_0 - Ay_m\|. \tag{11}$$

Note that since V_m constitutes an orthonormal basis for our Krylov subspace we have $y_m = V_m z_m$ with $z_m \in \mathbb{C}^n$ so that $Ay_m = AV_m z_m = V_{m+1}\bar{H}_m z_m$. We thus find that[2]

$$\|r_0 - Ay_m\| = \|V_{m+1}V_{m+1}^* r_0 - V_{m+1}\bar{H}_m y_m\| = \|V_{m+1}^* r_0 - \bar{H}_m z_m\| = \|\sqrt{(r_0,r_0)}e_1 - \bar{H}_m z_m\|, \tag{12}$$

and thus

$$\|r_m\| = \min_{z_m \in \mathbb{C}^n} \|\sqrt{(r_0,r_0)}e_1 - \bar{H}_m z_m\|, \qquad z_m = \underset{\hat{z}_m \in \mathbb{C}^n}{\arg\min} \|\sqrt{(r_0,r_0)}e_1 - \bar{H}_m \hat{z}_m\|. \tag{13}$$

This combination of building a Krylov space using the Arnoldi method and minimising the residual in every iteration yields the well-known GMRES algorithm by Saad and Schulz [13].

One can now understand why the computational work tends to become tedious if the iterations grow when performing an optimal Krylov method, we need to perform the Arnoldi process for growing spaces and thus need to compute more and more inner products. Besides this we need to store more and more vectors for our Krylov basis which can for large systems become a memory bottleneck. For this reason one would like to store less vectors and use less orthogonalisations and that is what is meant by short recurrences. If the matrix A is Hermitian, i.e. $A^* = A$, we get the result $H_m^* = H_m$ using the fact that $V_m^* A V_m = H_m$. Now recall that H_m is an upper Hessenberg matrix and in order for $H_m^* = H_m$ to be true we must have that H_m is in fact tridiagonal. In this case we often denote H_m by T_m

$$\bar{T}_m = \begin{pmatrix} t_{1,1} & t_{1,2} & & \\ t_{1,2}^* & \ddots & \ddots & \\ & \ddots & \ddots & t_{m-1,m} \\ & & t_{m-1,m}^* & t_{m,m} \\ & & & t_{m+1,m} \end{pmatrix} = \begin{pmatrix} T_m \\ e_m^* \end{pmatrix}.$$

The great advantage of this is that we now get recursions in the Gram-Schmidt procedure started from $v_1 = v/\|v\|$ of the form

$$\begin{aligned} \hat{v}_{n+1} &= Av_n - t_{n,n}v_n - t_{n,n-1}^* v_{n-1}, \\ t_{n,n} &= \frac{(Av_n, v_n)}{(v_n, v_n)}, \\ t_{n+1,n} &= \|\hat{v}_{n+1}\|, \\ v_{n+1} &= \frac{\hat{v}_{n+1}}{t_{n+1,n}}. \end{aligned}$$

This is also known as the Lanczos process to construct an orthonormal basis for $\mathcal{K}_n(A,v)$. As can be seen the Lanczos process uses so called three-term recurrence relations and only needs to store 3 consecutive vectors. The best known algorithm derived from the Lanczos process and which uses a minimal residual step to calculate x_m in each step is the aforementioned CG method.

3. Faber-Manteuffel theorem

First of all some technicalities about inner products. Given a matrix $A \in \mathbb{C}^{n\times n}$ and $x, y \in \mathbb{C}^n$ we have the Euclidean inner product $\langle \cdot, \cdot \rangle$ defined by $\langle x, y \rangle = y^* x$. If we take a Hermitian positive definite (HPD) matrix $B \in \mathbb{C}^{n\times n}$ then we can define the B-inner product $(\cdot,\cdot)_B$ by $(x,y)_B = y^* B x$. Using this B-inner product we can define the B-adjoint matrix of A as $(A^\dagger x, y)_B = (x, Ay)_B$ which then gives $A^\dagger = B^{-1}A^*B$. This already enlarges the class of matrices for which a short recurrence method exists from the Hermitian matrices to the $B-$Hermitian matrices[3] for which holds $A^\dagger = A$. We can just normally apply CG to this matrices by only changing the Euclidean inner product and

[2]Since V_m is an orthonormal matrix it preserves norms defined by an inner-product, i.e. $\|Vx\| = \|x\|$. For more exotic norms the statement might not be true, however these norms are not often encountered in solving linear systems.

[3]Note that Hermitian matrices are I-Hermitian in this sense where I is the identity matrix.

corresponding norm to the B-inner product and corresponding norm. Now we will focus mainly on the Euclidean inner product but the results can be easily extended to general B-inner products later on.
The Faber-Manteuffel theorem concerns the question whether there exists optimal Krylov methods which use recurrences of the form:

$$v_{i+1} = Av_i + \sum_{j=i-s}^{i} h_{j,i} v_j, \tag{14}$$

which are called $(s+2)$-term recurrences and are thus a generalisation of the 3-term recurrence of the Lanczos method. In their 1984 paper Faber and Manteuffel proved necessary and sufficient conditions for such a s-term recurrence to exist. An important definition in this context is the $(s+2)$-band Hessenberg matrix which is a Hessenberg matrix whose s-th superdiagonal contains at least one non-zero element and all its entries above the s-th diagonal are zero:

$$\bar{H}_n = \begin{pmatrix} h_{1,1} & h_{1,2} & \dots & h_{1,s+1} & 0 & \dots & 0 \\ h_{2,1} & \ddots & \ddots & & h_{2,s+2} & \ddots & \vdots \\ & \ddots & \ddots & \ddots & & \ddots & 0 \\ & & \ddots & \ddots & \ddots & & h_{n-s,n} \\ & & & \ddots & \ddots & \ddots & \vdots \\ & & & & \ddots & \ddots & h_{n-1,n} \\ & & & & & h_{n,n-1} & h_{n,n} \\ & & & & & & h_{n+1,n} \end{pmatrix}.$$

If the Arnoldi process for every given v_1 results in a $(s+2)$-band Hessenberg matrix decomposition $AV_m = V_{m+1}\bar{H}_m$ with at least one of the elements of the s-th superdiagonal being non-zero for at least one given v_1 we say that A admits an optimal $(s+2)$-recurrence. The word optimal is in the sense that there is a starting vector v for which at least in one iteration we need to orthogonalise against $s+1$ earlier vectors but for no vector we need more than $s+1$ vectors to get an orthonormal basis for $\mathcal{K}_m(A, v)$.

Before stating the main theorem we need to define one more matrix property. A matrix A having the property $AA^* = A^*A$, i.e. A commutes with its adjoint, is called normal. If the matrix furthermore satisfies $AA^* = A^*A = I$ then we call the matrix unitary. Some equivalent definitions of normality which will be of use later on are stated in the following lemma.

Lemma 3. *Let $A \in \mathbb{C}^{n\times N}$, then the following properties are equivalent:*

1. *A is normal,*
2. *A has a complete set of orthonormal eigenvectors,*
3. *$A^* = p(A)$ for any polynomial satisfying $p(\lambda_i) = \bar{\lambda}_i$ for all eigenvalues λ_i of A,*
4. *there exists a polynomial p such that $A^* = p(A)$.*

Proof. Suppose A is normal. If we look at the Schur decomposition of A we get an upper triangular matrix R and unitary matrix Q such that $A = QRQ^*$. This implies that $QR^*RQ^* = A^*A = AA^* = QRR^*Q^*$ which yields that R is normal as well. So R is triangular and normal which together imply that it must be diagonal. This can be seen from

$$\|Re_i\|_2^2 = \langle e_i, R^*Re_i\rangle = \langle e_i, RR^*e_i\rangle = \|R^*e_i\|_2^2,$$

where $e_i \in \mathbb{C}^n$ are the standard basis vectors. Starting from $i = 1$ we can show that only diagonal entries of R can be non-zero. As R is diagonal we have $A = QDQ^*$ for a diagonal matrix D, so A is unitarily diagonalisable or equivalently put, has a complete set of orthonormal eigenvectors, (1) $\Rightarrow$ (2).

Suppose $p(\lambda)$ is a polynomial such that $p(\lambda_i) = \bar{\lambda}_i$ for all eigenvalues λ_i of A. Assume that A has a complete set of orthonormal eigenvectors, or equivalently $A = U\Lambda U^*$ with U unitary, containing the orthonormal eigenbasis, and Λ diagonal with on the diagonal the eigenvalues of A. Then we have that $p(\Lambda) = \Lambda^*$ and therefore $Up(\Lambda)U^* = U\Lambda^*U^* = A^*$. Now note that for any

power of Λ we have $(U\Lambda U^*)^n = U\Lambda^n U^*$ since U is unitary. And therefore $q(U\Lambda U^*) = Uq(\Lambda)U^*$ for any polynomial q. Which proves that $p(A) = A^*$ and thus (2) $\Rightarrow$ (3).

Assume that $A^* = p(A)$ for any polynomial satisfying $p(\lambda_i) = \bar{\lambda}_i$ for all eigenvalues λ_i of A. By explicit construction using Lagrange polynomials there exists at least one polynomial (the Lagrange interpolation polynomial on the set $\{(\lambda_i, \bar{\lambda}_i)\}$) such that $p(\lambda_i) = \bar{\lambda}_i$ for all eigenvalues λ_i of A and thus there exists at least one polynomial p such that $A^* = p(A)$, (3) $\Rightarrow$ (4).

If $A^* = p(A)$ then it is trivial to see that A commutes with $p(A)$ and thus A^*. Which proves that A is normal, thus (4) $\Rightarrow$ (1).

□

Therefore an equivalent definition of A being normal is that the adjoint of A is a polynomial in A. This we can use to define a new class of matrices.

Definition 1. *Let $A \in \mathbb{C}^{n\times n}$ such that $AA^* = A^*A$. Let s be the minimal degree of a polynomial p such that $p(A) = A^*$. Then A is a normal(s) matrix.*

The minimal degree of a polynomial p such that $p(A) = A^*$ is called the normal degree $d_n(A)$ of A which is now uniquely defined. A trivial generalisation can be made to B-inner products and adjoints, a matrix is B-normal(s) if and only if $A^\dagger = p(A)$ where p is of minimal degree s.

Now we are ready to state the Faber-Manteuffel theorem in its most general form.

Theorem 1 (Faber-Manteuffel)**.** *Let $A \in \mathbb{C}^{n\times n}$ non-singular with $d_{\min}(A) = d$ and $B \in \mathbb{C}^{n\times n}$ a HPD matrix. Furthermore let $s \in \mathbb{N}$ be such that $s + 2 < d$, then A admits an optimal $(s + 2)$-recurrence for B if and only if A is B-normal(s).*

Remark 1. For the case $s + 2 = d_{\min}(A)$ see [12]. We also note that some subtleties involving the degree of the starting residual r_0 and the influence on the length of the recurrence relation is not treated in the original paper, but is treated in [12].

The proof of this theorem consists of two parts, sufficiency and necessity which will be treated separately. We assume the Euclidean inner product from which the results for B-inner products can be derived.

3.1. **Sufficiency.** An equivalent condition to an optimal $(s + 2)$-recurrence was $\bar{H}_n$ is $(s + 2)$-banded. Therefore we show that A being normal(s) implies that $\bar{H}_n$ is $(s + 2)$-banded.

Lemma 4. *Let $s \in \mathbb{N}$ such that $s + 2 < d$. If $A \in \mathbb{C}^{n\times n}$ is normal(s) then $\bar{H}_n$ is at most $(s + 2)$-banded.*

Proof. For $\bar{H}_n$ to be $(s + 2)$-banded we need $h_{i,j} = 0$ for $j > i + s$. Recall the definition of $h_{i,j} = \frac{(Av_j, v_i)}{(v_i, v_i)} = \frac{(v_j, A^* v_i)}{(v_i, v_i)}$. Since A is normal(s) there exists a polynomial q of exact degree s such that $A^* = q(A)$. Then note that $q(A)v_i \in \mathcal{K}_{i+s}(A, v_1)$ by construction of the Krylov subspace basis. We also know by construction via the Arnoldi procedure that $v_j \perp \mathcal{K}_{j-1}(A, v_1)$. Therefore we have that for $j \geq i + s + 1$ that $\mathcal{K}_{i+s}(A, v_1) \subseteq \mathcal{K}_{j-1}(A, v_1)$ and thus

$$(Av_j, v_i) = (v_j, A^* v_i) = (v_j, q(A)v_i) = 0. \tag{15}$$

As a result we find that for $j > i + s$ (equivalent to $j \geq i + s + 1$) that $h_{i,j} = 0$ and that thus $\bar{H}_n$ has at least all the diagonals above the s-th one completely zero and thus $\bar{H}_n$ is at most $(s + 2)$-banded. □

Therefore we know that there exists an optimal recurrence relation which relates at most $s + 2$ vectors. It can furthermore be proven that if there exists a vector of grade at least $s + 2$ then A admits an optimal $(s + 2)$-recurrence relation and no shorter one.

An interesting question one can now ask is which class of matrices are B-normal(s) for small s which is also answered in the original Faber-Manteuffel paper. First we relate s to the degree of A and thus to the number of distinct eigenvalues.

Lemma 5. *If A is normal(s) and $d = d_{\min}(A)$ then $s \leq d - 1$. Furthermore if $s > 1$ then $d \leq s^2$ and thus A has less then s^2 distinct eigenvalues.*

Proof. By lemma 1 we know that A has exactly d distinct eigenvalues as A is normal and thus diagonalisable. By lemma 3 there exists a Lagrange interpolation polynomial p such that $p(\lambda) = \bar{\lambda}$ for all eigenvalues of A. This polynomial will be of degree $d-1$ at most. Therefore we know that the normal degree of A is less then or equal to the degree of the polynomial p and thus $s \leq d-1$.

Now assume that $s > 1$. The question then is how many $\lambda \in \mathbb{C}$ can satisfy $p(\lambda) = \bar{\lambda}$ for a given polynomial of degree s. We note that $\bar{p}(\bar{\lambda}) = \lambda$ since p is a polynomial and therefore $\bar{p}(p(\lambda)) = \lambda$ which gives us a new polynomial equation of degree at most s^2

$$\bar{p}(p(\lambda)) - \lambda = 0. \tag{16}$$

A well known result of polynomial theory is that an polynomial of degree s^2 can have at most s^2 roots and therefore the equation (16) can have at most s^2 solutions which implies that at most s^2 eigenvalues of A can exist or otherwise put $d \leq s^2$. □

This result clearly destroys much hope of finding a short-recurrence method for most matrices since the number of distinct eigenvalues is bound by s^2 which we want to be small. The only interesting case could be if $s \leq 1$. This yields the following result proved in [6].

Lemma 6. *If A is normal(s) and $s \leq 1$ then either $d_{\min}(A) = 1$ or*

$$A = e^{i\theta}(\alpha I + iB), \tag{17}$$

where $\theta \in [0, 2\pi)$, $\alpha \in \mathbb{R}$ and $B^* = B$.

This lemma basically states that a three-term recurrence relation is only possible when the eigenvalues of the matrix A are all collinear, i.e. they lie on a straight line in the complex plane, or A only has one distinct eigenvalue.

3.2. **Necessity.** The proof of necessity, which needs to prove that $\bar{H}_n$ is normal(s) under the assumption that $\bar{H}_n$ is $(s+2)$ banded, is much harder than sufficiency. The original proof is based on a clever continuity argument and uses topological arguments. It is therefore a lot less intuitive. One of the reasons for the difficulty of the necessity part is that we now want to prove a result on the Hessenberg matrix $\bar{H}_n$ when only information is known about the first $(s+2)$ superdiagonals. Another approach was taken by Faber, Liesen and Tichý in [7] which gives two new and more intuitive proofs. One makes use of a very deep result in linear algebra, namely the cyclic decomposition theorem, and can therefore in essence not be viewed as an intuitive proof. The main difference with the original proof is that it relies solely on linear algebra arguments.

4. More short recurrences; extend the Faber-Manteuffel result

The Faber-Manteuffel theorem simply tells us that a short recurrence 'CG-like' method is very unlikely to exist for a general matrix A. The only matrices of practical interest are those whose eigenvalues lie on a straight line in the complex plane, a rather small subset of the class of all matrices.

However a result by Gragg appearing in 1982 in a Russian paper showed that for more general matrices an efficient method exists to construct an Arnoldi basis and by this opened new ways to explore short recurrences.

4.1. **Isometric Arnoldi process.** For a given inner-product recall that for a unitary matrix $U \in \mathbb{C}^{n\times n}$ we have $U^*U = UU^* = I$ and thus see that every unitary matrix is normal. As a result we might be able to use Theorem 1 to find short-recurrence Krylov methods. However, as we will see soon, U is in general not normal(s) for small s, it will be normal($n-1$) with n the number of eigenvalues of U. Therefore the Faber-Manteuffel theorem will in fact not help us.

Gragg, however, observed that if we apply the Arnoldi process to U we get the Arnoldi relation after the process is finished $UV_n = V_n H_n$ [8]. So we can see that $H_n = V_n^* U V_n$ and this implies that H_n is in fact unitary as well. Therefore we can make a QR-decomposition of H_n with $Q \in \mathbb{C}^{n\times n}$ a unitary matrix and $R \in \mathbb{C}^{n\times n}$ upper triangular such that $H_n = QR$. By using the fact that the product of two unitary matrices is unitary again we find that R in fact is unitary. A matrix which is both unitary and upper triangular must be diagonal as we saw in Lemma 3.

A way to construct the matrix Q is by using Givens rotation matrices $G_j(\gamma_j)$ where $\gamma_j \in \mathbb{C}$ with $|\gamma_j| \leq 1$, $\sigma_j \geq 0$

$$G_j(\gamma_j) = \begin{pmatrix} I_{j-1} & & & \\ & -\gamma_j & \sigma_j & \\ & \sigma_j & \bar{\gamma}_j & \\ & & & I_{n-j-1} \end{pmatrix}, \tag{18}$$

where $\sigma_j^2 + |\gamma_j|^2 = 1$. Also note that Givens matrices are unitary as well. Using this Givens matrices we can construct $Q = G_1(\gamma_1)...G_{n-1}(\gamma_{n-1})$ from which we find an expression for H_n

$$H_n = G_1(\gamma_1)...G_{n-1}(\gamma_{n-1})R, \tag{19}$$

where R is a diagonal unitary matrix. Now using $UV_n = V_nH_n$ and comparing columns we find a relation for $m < n$, namely $UV_m = V_{m+1}\bar{H}_m$ where

$$\bar{H}_m = G_1(\gamma_1)...G_{m-1}(\gamma_{m-1})\bar{G}_m(\gamma_j), \tag{20}$$

with the $G_j(\gamma_j)$ as in (18) and

$$\bar{G}_m(\gamma_m) = \begin{pmatrix} I_{m-1} & & \\ & -\gamma_m & \\ & \sigma_m & \end{pmatrix}. \tag{21}$$

Now we can define a recurrence relation to calculate the new Arnoldi vectors by seeing that

$$V_{m+1}G_1(\gamma_1) = (-\gamma_1 v_1 + \sigma_1 v_2, \sigma_1 v_1 + \bar{\gamma}_1 v_2, v_3, ..., v_{m+1}) = (-\gamma_1 v_1 + \sigma_1 v_2, \hat{v}_2, v_3, ..., v_{m+1}). \tag{22}$$

We can continue in this way by using the definition for $1 < k \leq m$

$$\hat{v}_{k+1} = \sigma_k \hat{v}_k + \bar{\gamma}_k v_{k+1}, \tag{23}$$

and we thus find that

$$V_{m+1}G_1(\gamma_1)..G_m(\gamma_m) = (-\gamma_1 v_1 + \sigma_1 v_2, -\gamma_2 \hat{v}_2 + \sigma_2 v_3, ..., -\gamma_{m-1}\hat{v}_{m-1} + \sigma_{m-1}v_m, \hat{v}_m, v_{m+1}). \tag{24}$$

This provides us with a way to calculate the new vector v_{m+1} using $\bar{G}_m(\gamma_m)$ as

$$v_{m+1} = \sigma_m^{-1}(Uv_m + \gamma_m \hat{v}_m). \tag{25}$$

One ingredient remains to be specified and that is the parameter γ_j. One can check that if we choose $\gamma_j = -\hat{v}_j^* U v_j$ we get an orthonormal basis V_n. The resulting algorithm is called the Isometric Arnoldi process (IAP) and is depicted in algorithm 1.

Input: $m \leq n$, $r_0 \in \mathbb{C}^n$, unitary matrix $U \in \mathbb{C}^{n\times n}$
Result: $V_m, \{\gamma_j\}_{j=1}^m, \{\sigma_j\}_{j=1}^m$
$v_1 = \frac{r_0}{\|r_0\|}$;
$\hat{v_1} = v_1$;
for $j = 1, 2, .., m-1$ **do**
  $u = Uv_j$;
  $\gamma_j = -\hat{v}_j^* u$;
  $\sigma_j = \left(1 - |\gamma_j|^2\right)^{1/2}$;
  $v_{j+1} = \sigma_j^{-1}(u + \gamma_j \hat{v}_j)$;
  $\hat{v}_{j+1} = \sigma_j \hat{v}_j + \bar{\gamma}_j v_{j+1}$;
end
$\gamma_m = -\hat{v}_m^* U v_m$;
$\sigma_m = \left(1 - |\gamma_m|^2\right)^{1/2}$;

Algorithm 1: Isometric Arnoldi process (IAP).

The reason why this process is called Isometric is because of the fact that a general isometric linear operator is an operator which is norm preserving. Now we see that a unitary matrix is of this kind, since for all $x, y \in \mathbb{C}^n$ we have

$$(Ux, Uy) = (x, U^*Uy) = (x, y) \tag{26}$$

and thus U preserves the inner product and the corresponding norm.

As for a computational note, we see that the algorithm only needs to store 2 vectors, namely v_j and $\hat{v}_j$ in order to generate an orthonormal basis. Therefore we can definitely regard this as a short-recurrence relation although it is not of the Lanczos type.

4.2. **More short-recurrences.** Inspired by the work of Gragg one can ask if there are other 'non-Arnoldi' type processes with short-recurrences to generate orthonormal bases. In 1995 Barth and Manteuffel published a paper [2] in which they generalise the result of Gragg, they considered short-recurrences of the type:

$$v_{n+1} = \sum_{i=n-m}^{n} \beta_{i,n} A v_i - \sum_{i=n-l}^{n} \sigma_{i,n} v_i. \tag{27}$$

Recurrences of this type are called (l, m)-recurrences and they lead to an adapted Arnoldi relation of the form

$$AV_k R_k = V_{k+1} \bar{H}_k, \tag{28}$$

where now $R_k \in \mathbb{C}^{k\times k}$ is an upper triangular matrix, of which the bandwidth is determined by m.

We can immediately see that this type of recurrences generalises the earlier result by Faber and Manteuffel since those short-recurrences, which we called $(s+2)$-recurrences, can be written in the form of Equation (27) with $m = 0$ and $l = s$. To see that the IAP is also of the (l, m)-recurrence type we write

$$\sigma_j v_{j+1} = U v_j + \gamma_j \hat{v}_j. \tag{29}$$

We want to get rid of all the $\hat{v}_j$ and therefore use Equation (23) to get

$$\gamma_j^{-1}(\sigma_j v_{j+1} - U v_j) = \hat{v}_j = \sigma_{j-1}\hat{v}_{j-1} + \bar{\gamma}_{j-1} v_j. \tag{30}$$

This can be rewritten to find $\hat{v}_{j-1}$ and thus $\hat{v}_j$

$$\hat{v}_{j-1} = \sigma_{j-1}^{-1}\big(\gamma_j^{-1}(\sigma_j v_{j+1} - U v_j) - \bar{\gamma}_{j-1} v_j\big). \tag{31}$$

Now we can finally get rid of $\hat{v}_j$ in Equation (29) and get

$$\sigma_{j-1} v_j = U v_{j-1} + \gamma_{j-1}\sigma_{j-1}^{-1}\gamma_j^{-1}(\sigma_j v_{j+1} - U v_j - \gamma_j \bar{\gamma}_{j-1} v_j), \tag{32}$$

or written as a (l, m)-recursions

$$v_{j+1} = \gamma_{j-1}^{-1}\gamma_j \sigma_{j-1}\sigma_j^{-1}(\sigma_{j-1} v_j - U v_{j-1}) + \sigma_j^{-1}(U v_j + \gamma_j \bar{\gamma}_{j-1} v_j) \tag{33}$$

$$= \frac{1}{\sigma_j} U v_j - \frac{\gamma_j \sigma_{j-1}}{\gamma_{j-1}\sigma_j} U v_{j-1} + \left(\frac{\gamma_j \sigma_{j-1}^2}{\gamma_{j-1}\sigma_j} + \frac{\gamma_j \bar{\gamma}_{j-1}}{\sigma_j}\right) v_j. \tag{34}$$

So we can conclude that IAP is a (0,1)-recurrence, i.e. $l = 0$ and $m = 1$. Therefore the (l, m)-recurrences are a wider class than the aforementioned ones. The very important question remains whether this (l, m)-recurrences construct an orthonormal basis, and ideally for small l, m. In order to answer this question, at least partially, we need to define a generalisation of the concept normal(s) matrices, the so called normal(l, m) matrices.

Definition 2. *Let $A \in \mathbb{C}^{n\times n}$ for which we have $AA^* = A^*A$. Then A is a normal(l, m) if there exist relatively prime polynomials p, q of degree l, m respectively such that*

$$A^* q(A) = p(A). \tag{35}$$

By relatively prime polynomials we mean that the only common divisor to both polynomials is the constant polynomial 1. This definition can be generalised to B-inner products as well. Furthermore we have the following sufficiency condition proved by Barth and Manteuffel [3]

Theorem 2. *Let $A \in \mathbb{C}^{n\times n}$ non-singular and $B \in \mathbb{C}^{n\times n}$ a HPD matrix. If A is B-normal(l, m) then A admits a (l, m)-recurrence which yields an orthonormal basis.*

In the same manner as for the B-normal(s) matrices one would like to know which class of matrices are normal(l, m) for small l, m. To answer this question we use some results by Liesen [11]. In this work the paper of Barth and Manteuffel is improved in the sense that one ambiguity is removed. This ambiguity is due to the fact that there exists in general no unique (l, m) such that l and m are both smallest which is something that is assumed in [3, Theorem 3.1]. One can

see this for example when considering a unitary matrix $U \in \mathbb{C}^{n\times n}$. Then we know that on the one hand U is normal$(1,0)$ since $U^*U = I$. But on the other hand we have that U is normal$(0, m-1)$ with m the number of eigenvalues since $U^* = p(U)$ with the degree of p being $m-1$ by a result in [11].

Let p, q be relatively prime polynomials of degree l, m respectively. Then the McMillan degree of $r = p/q$ is defined by $\max\{l, m\}$. This can be generalised to a McMillan degree $d_r(A)$ of a matrix A.

Definition 3. *The McMillan degree $d_r(A)$ of a matrix A is the smallest McMillan degree of all rational functions r such that for all eigenvalues λ of A we have that $r(\lambda) = \bar{\lambda}$.*

If A is a diagonalisable matrix with $d_r(A) = t$ then we know that there exists a rational function r of degree t such that $A^* = r(A)$ which we can write as $A^*q(A) = p(A)$ for certain relatively prime polynomials p, q of degree at most t. By Theorem 2 we then know that there exists at least a (t,t)-recurrence relation to compute an orthonormal basis and possibly even a shorter (l,m)-recurrence. It is therefore interesting to know which matrices have a small McMillan degree $d_r(A)$ since for these matrices there exists a short (l,m)-recurrence relation. This is characterised by the following theorem by Liesen [11].

Theorem 3. *Let $A \in \mathbb{C}^{n\times n}$ a diagonalisable matrix with m distinct eigenvalues. Then we have that*

1. *if $m < 4$, then $d_r(A) = 1$ and $d_n(A) \leq m-1$,*
2. *if all the eigenvalues of A are collinear, then $d_r(A) = d_n(A) = 1$,*
3. *if all the eigenvalues of A are concyclic, then $d_r(A) = 1$ and $d_n(A) = m-1$,*
4. *and in all the other cases we find that $d_r(A) \geq m/5 + 1$ and $d_n(A) \geq (m+2)/3$.*

Corollary 2. *Let $A \in \mathbb{C}^{n\times n}$ a diagonalisable matrix with $m \geq 4$ distinct eigenvalues which are neither collinear or concyclic, then we have*

$$1 \leq \frac{d_n(A)}{d_r(A)} \leq 5\frac{m-1}{m+5} < 5. \tag{36}$$

Proof. If A is normal(s) then $p(\lambda) = \bar{\lambda}$ for all eigenvalues of A with degree of p being s. We can rewrite this as a rational equation $r'(\lambda) = \frac{p(\lambda)}{1} = \bar{\lambda}$. Therefore by the minimum property of $d_r(A)$ we know that $d_n(A) = s \leq d_r(A)$. This then provides the first inequality. Furthermore we know by lemma 8 that $d_n(A) \leq m-1$. Then we combine this with $d_r(A) \geq \frac{1}{5}(m+5)$ which yields that $\frac{1}{d_r(A)} \leq 5\frac{1}{m+5}$. So we get the second inequality. The last inequality is a trivial one since $m > 0$. □

First of all we need to clarify what is meant by concyclic. Concyclic means that all the eigenvalues lie on the same circle in the complex plane. This theorem again is not very helpful to us in the sense that it places a rather strict condition on matrices to have small $d_r(A)$ or $d_n(A)$, just as the Faber-Manteuffel theorem did. However, we only considered sufficient conditions for (l,m)-recurrences so far, so there might be a larger class of matrices that have these short recurrences, but it seems that necessary conditions for (l,m)-recurrences have not been found so far.

5. Shifted unitary matrices

Let us now further investigate the case of concyclic eigenvalues. Matrices with eigenvalues on the unit circle are precisely the unitary matrices and therefore all matrices with concyclic eigenvalues are shifted and scaled unitary matrices,

$$A = \zeta I + \rho U, \tag{37}$$

with $I \in \mathbb{C}^{n\times n}$ the unit matrix, $U \in \mathbb{C}^{n\times n}$ unitary and $\zeta, \rho \in \mathbb{C}$. These are matrices with eigenvalues on the circle with center ζ and radius $|\rho|$. This class of matrices is exactly the class that Jagels and Reichel considered in their 1994 paper [10]. Their starting point however was a bit different. They started with the IAP discovered by Gragg and noted that Krylov spaces are shift-invariant as we already stated. Therefore we have the relation

$$\mathcal{K}_m(U, r_0) = \mathcal{K}_m(A, r_0). \tag{38}$$

We see that in order to build a Krylov space for the shifted and scaled unitary matrix A we could choose to simply concentrate on building a Krylov space for U which can be done by the short-recurrence IAP. In order to build an iterative method to solve equations with A as in (37) they used an economical minimal residual step to compute the new approximation, thus resulting in an optimal Krylov method. Economical in the sense that only four additional vectors need to be stored in addition to the current best approximation.

The minimal residual step first makes a QR-factorisation of the matrix $\bar{H}_m$ in Equation (13), i.e. $\bar{H}_m = Q_{m+1}R_m$, where $Q_{m+1} \in \mathbb{C}^{(m+1)\times(m+1)}$ unitary and $R_m \in \mathbb{C}^{(m+1)\times m}$ upper triangular. This can again be done by using Givens matrices and the cost per iteration only involves the calculation of two scalars, σ_m and γ_m.

As the matrix Q_{m+1} is unitary we can use this to rewrite (13) to

$$\|r_m\| = \min_{z_m \in \mathbb{C}^n} \|\sqrt{(r_0, r_0)}Q_{m+1}^* e_1 - R_m z_m\|, \qquad z_m = \arg\min_{\hat{z}_m \in \mathbb{C}^n} \|\sqrt{(r_0, r_0)}Q_{m+1}^* e_1 - R_m z_m\|. \tag{39}$$

This results in solving a triangular system with R_m from the QR-decomposition in order to compute the next iterate. As said before, only four additional vectors need to be stored in order to compute the solution of this triangular system and the matrices R_m and Q_{m+1} do not have to be explicitly created due to the structure of A, see [10] for more details on this.

The resulting algorithm is named Shifted Unitary Minimal Residual (SUMR) method (Algorithm 2) and requires only one matrix vector multiplication with U and six vectors to be stored in total and is equal to GMRES in exact arithmetic. The residual norm is iteratively computed every iteration without any extra cost.

Input: $r_0 \in \mathbb{C}^n$, $\rho, \zeta \in \mathbb{C}$, unitary matrix $U \in \mathbb{C}^{n\times n}$, tolerance ε
Result: approximate solution x
$\delta_0 = \|r_0\|$; $\hat{\varphi}_1 = 1/\delta_0$; $\hat{\tau}_1 = \delta_0/\rho$; $w_{-1} = p_{-1} = v_0 = 0$;
$\varphi_0 = s_0 = \lambda_0 = r_{0,-1} = 0$; $r_{0,0} = \gamma_0 = \sigma_0 = c_0 = 1$;
$v_1 = \tilde{v}_1 = r_0/\delta_0$;
$j = 0$;
while $|\hat{\tau}_{j+1}| > \varepsilon$ **do**
  $u = Uv_j$;
  $\gamma_j = -\tilde{v}_j u$;
  $\sigma_j = ((1 + |\gamma_j|)(1 - |\gamma_j|))^{1/2}$;
  $\alpha_j = -\gamma_j \delta_{j-1}$;
  $r_{j-1,j} = \alpha_j \varphi_{j-1} + s_{j-1}\zeta/\rho$;
  $\hat{r}_{j,j} = \alpha_j \hat{\varphi}_j + \bar{c}_{j-1}\zeta/\rho$;
  $\bar{c}_j = \hat{r}_{j,j}/(|\hat{r}_{j,j}|^2 + |\sigma_j|^2)^{1/2}$; $s_j = -\sigma_j/(|\hat{r}_{j,j}|^2 + |\sigma_j|^2)^{1/2}$;
  $r_{j,j} = -c_j \hat{r}_{j,j} + s_j \sigma_j$;
  $\tau_j = -c_j \hat{\tau}_j$; $\hat{\tau}_{j+1} = s_j \hat{\tau}_j$;
  $\eta_j = \tau_j / r_{j,j}$; $\kappa_{j-1} = r_{j-1,j}/r_{j-1,j-1}$;
  $w_{j-1} = \alpha p_{j-2} - (w_{j-2} - v_{j-1})\kappa_{j-1}$;
  $p_{j-1} = p_{j-1} - w_{j-2} - v_{j-1})\lambda_{j-1}$;
  $x_j = x_{j-1} - (w_{j-1} - v_j)\eta_j$;
  $\delta_j = \delta_{j-1}\sigma_j$;
  $\varphi_j = -c_j \hat{\varphi}_j + s_j \bar{\gamma}_j/\delta_j$; $\lambda_j = \varphi_j / r_{j,j}$;
  $\hat{\varphi}_{j+1} = s_j \hat{\varphi}_j + \bar{c}_j \bar{\gamma}_j/\delta_m$;
  $v_{j+1} = \sigma_j^{-1}(u + \gamma_j \tilde{v}_j)$;
  $\tilde{v}_{j+1} = \sigma_j \tilde{v}_j + \bar{\gamma}_j v_{j+1}$;
  $j = j + 1$;
end
$x = x_j$;

Algorithm 2: Shifted Unitary Minimal Residual (SUMR) by Jagels and Reichel [10].

5.1. **Performance SUMR.** Jagels and Reichel tested their algorithm on two classes of examples which we have tested as well. Note that for testing purposes it is sufficient to test matrices of the

form

$$A = \zeta I + U, \tag{40}$$

since we can rewrite equations $Ax = b$ with A as in (37) to (40) by dividing by ρ. For comparison we use GMRES as well in order to compare SUMR to it, since both are equivalent in exact arithmetic. Jagels and Reichel compared SUMR to a stationary Richardson method as well with shift parameter $\frac{1}{\zeta}$ in order to map all the eigenvalues to the unit circle with center 1. In our examples we implemented the Richardson method via a Local Minimal Residual method (LMR) and by choosing the update parameter as a fixed constant $\frac{1}{\zeta}$ we get the Richardson method. For certain classes of examples we used the LMR method to compare to what happens if the SUMR method starts to degrade.

As starting vector we choose $x_0 = 0 \in \mathbb{C}^n$ and for $b \in \mathbb{C}^n$ a random vector as was done by Jagels and Reichel [10].

5.1.1. *Example 5.1.* As a first example a random matrix $R \in \mathbb{C}^{n\times n}$ for $n = 200$ is taken and then W is the unitary matrix from the QR decomposition of R. Then let

$$\lambda_k = e^{i\pi(k-1)/6}, \quad 1 \leq k \leq 6$$

and θ_k for $7 \leq k \leq 200$ be randomly generated numbers from the interval $(-\pi/4, \pi/4)$ from which we construct $\lambda_k = e^{i\theta_k}$ to get $\Lambda = \mathrm{diag}(\lambda_k) \in \mathbb{C}^{200\times 200}$. Now construct

$$U = W\Lambda W^* \tag{41}$$

as unitary matrix. As shift $\zeta = 1.1$ is chosen, which makes it the second example considered by Jagels and Reichel. This results in a structured eigenvalue distribution as seen in Figure 2a.

Figure 1 shows the results of the iterations. We clearly see something strange happening around the 25$^{\text{th}}$ iteration, SUMR starts to converges significantly slower than GMRES and starts to behave as if it were a Richardson iteration. The reason for this is shown in Figure 3, the norm of the computed Arnoldi vectors start to differ from 1.

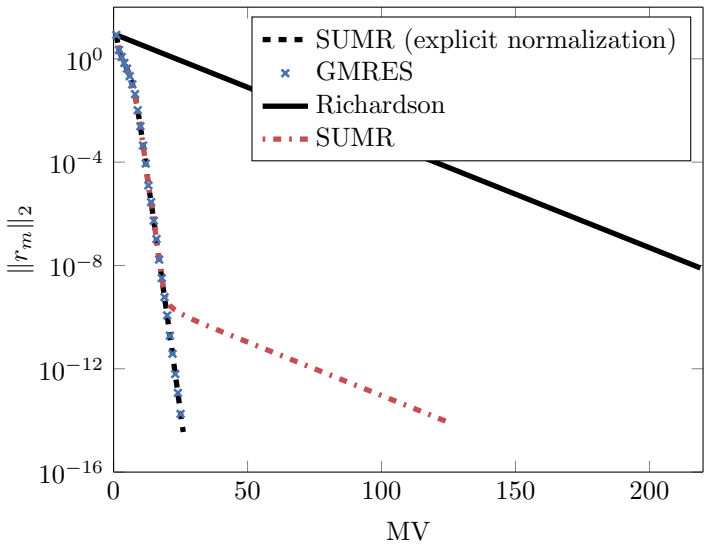
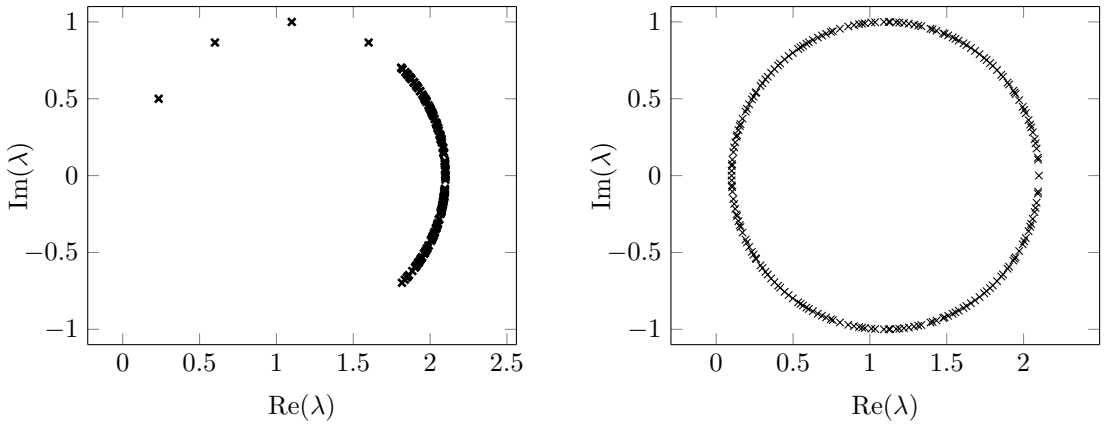


FIGURE 1. Residual norms for Example 5.1 with $\zeta = 1.1$, $\rho = 1$ and $n = 200$ as a function of the number of matrix vector multiplications.

In order to test the hypothesis that this effect causes the lack of convergence we implemented a different version of SUMR in which the basis vectors are explicitly normalised each iteration. The result of this can be more clearly seen in Figure 4 were the orthogonality of the last $m + 1$ vectors of the supposedly orthonormal basis V is plotted for $m = 1, 2, 3, 4, 5$. For the standard SUMR implementation we see that orthogonality is lost as soon as the norm of the vectors starts to differ from 1 whereas the explicitly normalised version clearly only has a very slight decrease in orthonormality with increasing iterations. The explicitly normalised SUMR can be seen to



(A) Eigenvalue distribution for matrix A in Example 5.1 with $\zeta = 1.1$, $\rho = 1$ and $n = 200$.

(B) Eigenvalue distribution for matrix A in Example 5.2 with $\zeta = 1.1$, $\rho = 1$ and $n = 200$.

FIGURE 2

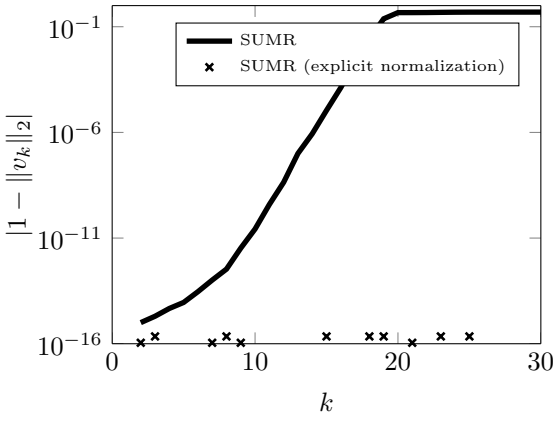


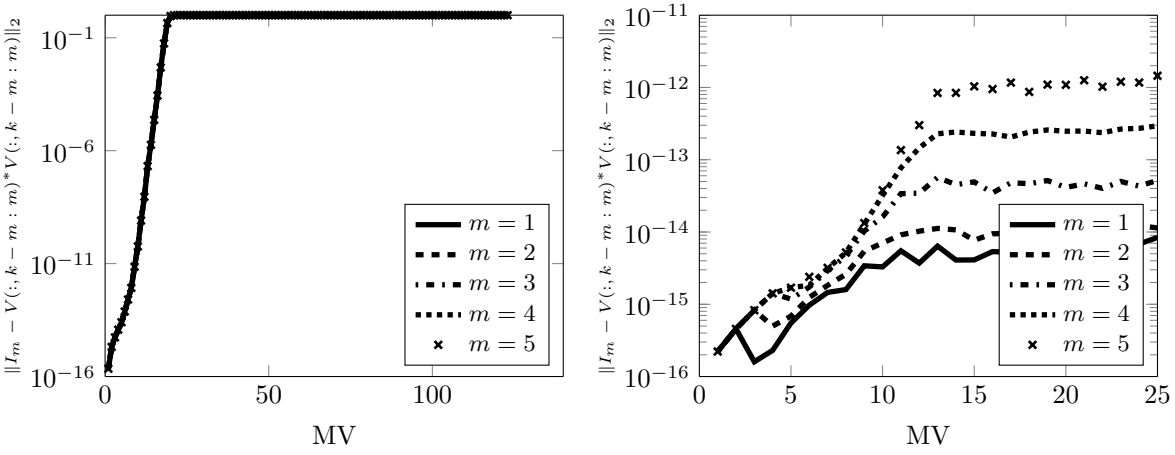
FIGURE 3. Deviation from being normed for the Arnoldi vectors as a function of the iteration k for Example 5.1.

converge at the same pace as GMRES in Figure 1, they seem to be equivalent indeed. We therefore propose the normalised adaptation in Algorithm 3.

5.1.2. *Example 5.2.* As a second example we take a random matrix $W \in \mathbb{C}^{n\times n}$ for $n = 200$ and then let U be the unitary matrix from the QR decomposition of W. As shift we take $\zeta = 1.1$, resulting in a spectrum of randomly distributed eigenvalues along a circle in the complex plane, so without any special structure, as can be seen in 2b.

The convergence is much slower than we saw for example 5.1. Probably this will be due to the fact that no eigenvalues are extremal and therefore the Krylov method won't easily detect eigenvectors. Therefore the matrix will not be deflated and thus no fast convergence is expected, not for SUMR or GMRES. In fact they converge at the roughly the same speed as the Richardson iteration, only to speed up at the moment that the iteration number is around the size of the matrix. This is what we expect from GMRES, and thus SUMR, since the exact solution is in the Krylov space when the dimension of the Krylov space is the size of the matrix.

The original paper by Jagels and Reichel uses this example as well only the convergence seems to differ from our observations. Both our and their results show convergence towards the end only our results clearly show a small speed up at the beginning which does not show up in the paper by Jagels and Reichel. Furthermore convergence is about equal to Richardson for a long time which also is not present in the other paper. A test with the original SUMR (Algorithm 2) does



(A) SUMR without explicit normalisation.

(B) SUMR with explicit normalisation

FIGURE 4. Orthonormality of the last m calculated Arnoldi vectors as a function of the MV's for Example 5.1. Note that the difference in the MV axis between the two methods is caused by the lack of convergence for SUMR without explicit normalisation.

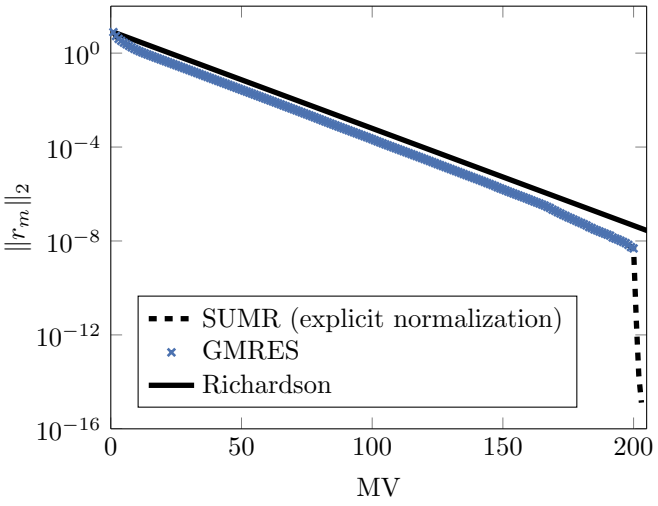


FIGURE 5. Residual norms for Example 5.2 with $\zeta = 1.1$, $\rho = 1$ and $n = 200$ as a function of the number of matrix vector multiplications.

not recover the figures from Jagels and Reichel either, so it is not clear what the reason is for this discrepancy.

5.1.3. *Example 5.3.* As a third example we use a variation on the first example, we want to have a matrix with structure. We pick 50 random numbers in $[0, 2\pi)$ and then choose a small ε to make a cluster of size ε of 20 numbers around the first 50 chosen random numbers. Finally we take the exponential to get small clusters of eigenvalues on the unit circle. A small shift of $\zeta = -0.1$ is introduced resulting in the eigenvalue distribution shown in Figure 6b.

The result, Figure 6, shows that GMRES and SUMR perform equally and that the LMR method (with variable update parameter) does not converge. This is good to know as LMR has very low memory requirements and only uses 2 inner products and 1 matrix vector multiplication and is therefore very fast if it converges. This example, however, shows that it will not converge in general. On the other hand SUMR does converge at the same rate as GMRES due to its minimal residual step.

Input: $r_0 \in \mathbb{C}^n$, $\rho, \zeta \in \mathbb{C}$, unitary matrix $U \in \mathbb{C}^{n\times n}$, m maximum number of iterations, tolerance ε
Result: approximate solution x
$\delta_0 = \|r_0\|$; $\hat{\varphi}_1 = 1/\delta_0$; $\hat{\tau}_1 = \delta_0/\rho$; $w_{-1} = p_{-1} = v_0 = 0$;
$\varphi_0 = s_0 = \lambda_0 = 0$; $r_{0,0} = c_0 = 1$;
$v_1 = \tilde{v}_1 = r_0/\delta_0$;
for $j = 1, ..., m$ **do**
  $u = Uv_j$;
  $\gamma_j = -\tilde{v}_j u$;
  $\sigma_j = ((1 + |\gamma_j|)(1 - |\gamma_j|))^{1/2}$;
  $\alpha_j = -\gamma_j \delta_{j-1}$;
  $r_{j-1,j} = \alpha_j \varphi_{j-1} + s_{j-1}\zeta/\rho$;
  $\hat{r}_{j,j} = \alpha_j \hat{\varphi}_j + \bar{c}_{j-1}\zeta/\rho$;
  $\bar{c}_j = \hat{r}_{j,j}/(|\hat{r}_{j,j}|^2 + |\sigma_j|^2)^{1/2}$; $s_j = -\sigma_j/(|\hat{r}_{j,j}|^2 + |\sigma_j|^2)^{1/2}$;
  $r_{j,j} = -c_j \hat{r}_{j,j} + s_j \sigma_j$;
  $\delta_j = \delta_{j-1}\sigma_j$;
  $\varphi_j = -c_j \hat{\varphi}_j + s_j \bar{\gamma}_j/\delta_j$;
  $\hat{\varphi}_{j+1} = s_j \hat{\varphi}_j + \bar{c}_j \bar{\gamma}_j/\delta_m$;
  $\tau_j = -c_j \hat{\tau}_j$; $\hat{\tau}_{j+1} = s_j \hat{\tau}_j$;
  $\eta_j = \tau_j/r_{j,j}$; $\kappa_{j-1} = r_{j-1,j}/r_{j-1,j-1}$;
  $w_{j-1} = \alpha p_{j-2} - (w_{j-2} - v_{j-1})\kappa_{j-1}$;
  $p_{j-1} = p_{j-1} - w_{j-2} - v_{j-1})\lambda_{j-1}$;
  $x_j = x_{j-1} - (w_{j-1} - v_j)\eta_j$;
  if $|\hat{\tau}_{j+1}| \leq \varepsilon$ **then**
    break;
  end
  $\lambda_j = \varphi_j/r_{j,j}$;
  $v_{j+1} = \sigma_j^{-1}(u + \gamma_j \tilde{v}_j)$;
  $v_{j+1} = v_{j+1}/\|v_{j+1}\|$;
  $\tilde{v}_{j+1} = \sigma_j \tilde{v}_j + \bar{\gamma}_j v_{j+1}$;
  $\tilde{v}_{j+1} = \tilde{v}_{j+1}/\|\tilde{v}_{j+1}\|$;
end
$x = x_j$;

Algorithm 3: Shifted Unitary Minimal Residual (SUMR) with explicit normalisation.

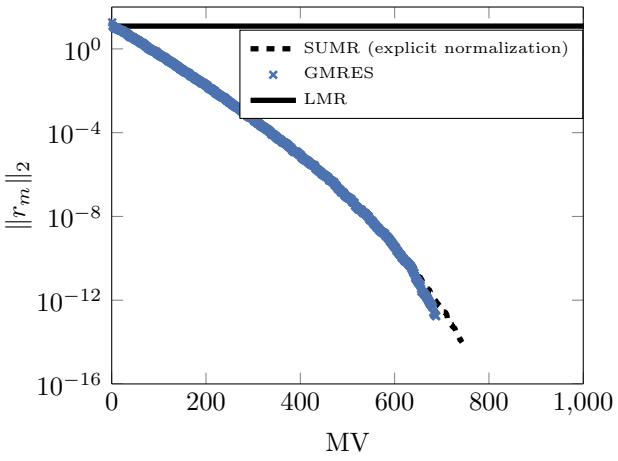
5.2. **Further notes on SUMR.** The computational work for SUMR is very low compared to its equivalent GMRES and it is therefore not surprising that it converges faster and performs better in terms of memory requirements. When we compare SUMR to a method of equal memory storage requirement, for example a restarted GMRES variant, in this case GMRES(6), we conclude that SUMR outperforms this method for every matrix tested.

Some note on preconditioning SUMR can be made as well. In order for SUMR to work we need to be able to apply the IAP which implies that we must have a unitary shifted matrix. This restricts the class of matrices which can be used to precondition A. One could use a unitary matrix $W \in \mathbb{C}^{n\times n}$ and apply a central preconditioning

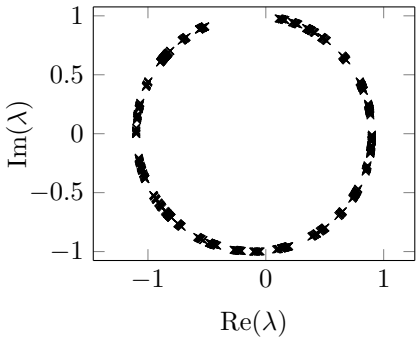
$$WAW^* = \zeta I + \rho WUW^*$$

However this does not help much to improve the spectrum of A since we know that for $M = M_1 M_2$ the matrices $M^{-1}A$ and $M_1^{-1}AM_2^{-1}$ have the same spectrum. If we apply this with $M_1 = W^{-1} = W^*$ and $M_2 = W^{-*} = W$ we find that WAW^* has the same spectrum as A. Therefore preconditioning is not feasible for SUMR.

Current applications of SUMR can be found in Quantum Chromo Dynamics (QCD) in which large sparse linear systems need to be solved with shifted unitary matrices [1].



(A) Residual norms for as a function of the number of matrix vector multiplications.



(B) Eigenvalue distribution for matrix A in Example 5.3.

FIGURE 6. Residual norms and eigenvalues for Example 5.3 with $\zeta = -0.1$, $\rho = 1$ and $n = 1000$.

6. Nearly Hermitian matrices

In their paper [3] Barth and Manteuffel discuss an even further generalisation of normal matrices, namely the so called generalised B-normal(l, m) matrices.

Definition 4. *Let $A \in \mathbb{C}^{n\times n}$. Then A is a generalised-normal(l, m) matrix if there exist relatively prime polynomials p, q of degree l, m respectively such that*

$$A^* q(A) - p(A) = r(A), \tag{42}$$

with r a polynomial such that rank$(r(A)) = \kappa$.

If $\kappa = 0$ then these are exactly the normal(l, m) matrices. This can of course be generalised to the B-inner product. A special subclass of the generalised B-normal(l, m) matrices is the class of nearly Hermitian matrices. For this matrices we have that $q(A) = I$ and $p(A) = A$ such that we have

$$A^* - A = r(A)$$

Remember that we can decompose each matrix A in a Hermitian part $H = \frac{1}{2}(A + A^*)$ and a skew-Hermitian part $S = \frac{1}{2}(A - A^*)$. Now it is interesting to look at the case where $r(A)$ is a low rank matrix, which then implies that A has a low rank skew-Hermitian part S. This is the class of matrices which Beckermann and Reichel considered in their paper [4].

$$A - A^* = \sum_{i=1}^{s} f_i g_i^* = FG^*, \quad f_i, g_i \in \mathbb{C}^n, \tag{43}$$

with $F = [f_1, ..., f_s] \in \mathbb{C}^{n\times s}$ and $G = [g_1, ..., g_s] \in \mathbb{C}^{n\times s}$. So we have a matrix which has a low rank s skew-Hermitian part. If we have an Hermitian matrix we know that there exists an optimal three-term recurrence, the CG method. The question now is, can we use this short-recurrence and combine it with some knowledge of the low rank perturbation to create recurrences which are at least shorter than GMRES. An example of this class of matrices which is often encountered are matrices of the form

$$A = H + \sum_{i=1}^{s/2} f_i g_i^*, \quad f_i, g_i \in \mathbb{C}^n, \tag{44}$$

where $H = H^*$, or in words the case of a Hermitian matrix H which is perturbed by a rank $s/2$ matrix. We can now rewrite (44) to

$$A - A^* = \sum_{i=1}^{s/2} f_i g_i^* - \sum_{i=1}^{s/2} g_i f_i^* \tag{45}$$

and if we then choose $F = [f_1, ..., f_{s/2}, g_1, ..., g_{s/2}]$ and $G = [g_1, ..., g_{s/2}, -f_1, ..., -f_{s/2}]$ we find it to be of the form (43). Beckermann and Reichel showed in their paper how the special structure of the problem can indeed be exploited by finding an Arnoldi process for these nearly Hermitian matrices to generate an orthonormal basis of the Krylov spaces.

6.1. **Arnoldi process for nearly Hermitian matrices.** Assume that we have after k steps of the Arnoldi process the orthonormal basis $V_k = [v_1, ..., v_k] \in \mathbb{C}^{n\times k}$ for $\mathcal{K}_k(A, v_1)$. Now we seek a way to expand our Krylov space to $\mathcal{K}_{k+1}(A, v_1)$. We know that $\Pi_k = V_k V_k^*$ is an orthogonal projection operator on the space spanned by $\mathcal{K}_k(A, v_1)$ and thus $(I-\Pi_k)$ projects on the orthogonal complement of our Krylov space. We know that by our Arnoldi method we have $v_k \perp Av_l$ for all $1 \leq l \leq k-2$ and therefore $A^*v_k \perp v_l$ for all $1 \leq l \leq k-2$. However this does not make A^*v_k directly a candidate to expand our Krylov space as we want this candidate to be inside $\mathcal{K}_{k+1}(A, v_1)\backslash\mathcal{K}_k(A, v_1)$ as well in order to expand our search space. We can however, create a suitable candidate from A^*v_k as the following lemma shows.

Lemma 7. *Assume that* $\dim \mathcal{K}_{k+1}(A, v_1) = k+1$ *and define* $v_k' = A^*v_k + (I - \Pi_k)(A - A^*)v_k$. *Then we have that*

$$v_k' \in \mathcal{K}_{k+1}(A, v_1)\backslash\mathcal{K}_k(A, v_1), \qquad v_k' \perp \mathcal{K}_{k-2}(A, v_1) \tag{46}$$

Proof. First the second statement. We already know that $A^*v_k \perp \mathcal{K}_{k-2}(A, v_1)$. Furthermore $(I-\Pi_k)$ projects on the orthogonal complement of $\mathcal{K}_k(A, v_1)$ and therefore $(I-\Pi_k)(A-A^*)v_k \perp \mathcal{K}_k(A, v_1)$ from which follows that $(I - \Pi_k)(A - A^*)v_k \perp \mathcal{K}_{k-2}(A, v_1)$ as well. This proves the second statement.
We can rewrite v_k' in the following form

$$v_k' = A^*v_k - A^*v_k + Av_k + \Pi_k(A^* - A)v_k = Av_k + \Pi_k(A^* - A)v_k$$

Since $\dim \mathcal{K}_{k+1}(A, v_1) = k+1$ we have that $Av_k \in \mathcal{K}_{k+1}(A, v_1)$ and $Av_k \notin \mathcal{K}_k(A, v_1) \subset \mathcal{K}_{k+1}(A, v_1)$. By the projection property of Π_k we also know that $\Pi_k(A^* - A)v_k \in \mathcal{K}_k(A, v_1)$. Therefore we find that $v_k' \in \mathcal{K}_{k+1}(A, v_1)\backslash\mathcal{K}_k(A, v_1)$. □

This makes v_k' a very good candidate to expand our search space. We only need to orthogonalise v_k' against v_k, v_{k-1} just as is the case in the Lanczos algorithm. One might ask why this is a good choice in finding short recurrences since in order to compute v_k' we seem to need Π_k for which we need to store the whole basis V_k. If we define $\tilde{F}_k = \Pi_k F$ then we find that

$$\tilde{F}_{k+1} = \Pi_{k+1}F = \Pi_k F + v_{k+1}v_{k+1}^* F = \tilde{F}_k + v_{k+1}v_{k+1}^* F \tag{47}$$

so we can iteratively update $\tilde{F}_k$. Now if we look at v_k' we see that we can write it as

$$v_k' = Av_k - \tilde{F}_k G^* v_k \tag{48}$$

So in order to compute v_k' we only need to store v_k, $\tilde{F}_k$ and the action of A, F and G. If we want to compute v_{k+1} which then is orthogonal to v_k, v_{k-1} we need to store these two vectors as well. This makes a total of 2 n-vectors (since as soon as v_{j+1} is known v_{j-1} can be overwritten), one full matrix and two $n \times s$ matrices, the storage requirement is thus $(2s+2)n$ locations in addition to a way to form matrix products with A. If s is small this is a huge improvement over the storage for the Arnoldi process, namely one full matrix A and n times n-vectors, or n^2 locations and the full matrix.

6.2. **A progressive GMRES method.** Using the derived Arnoldi method for nearly Hermitian matrices Beckermann and Reichel derived a minimal residual method using short-recurrences. The key idea again is to solve Equation (13) and then to note that since A has a low rank skew-Hermitian part, H_m has a low rank skew-Hermitian part as well. Then by computing the QR factorisation of $\bar{H}_m$ by Givens matrices one gets their Progressive generalised Minimal RESidual

method (PGMRES). An adapted version of the algorithm to fit better with the preceding section is given in Algorithm 4.

Input: $x_0, b \in \mathbb{C}^n$, $A \in \mathbb{C}^{n\times n}$, $F_1, G_1 \in \mathbb{C}^{n\times s/2}$ such that $A - F_1 G_1^*$ is Hermitian, m maximum number of iterations, tolerance ε
Result: approximate solution x, residual history $\{|\gamma_j|\}_{j=0}^m$
$F = [F_1, G_1]; G = [G_1, -F_1]$;
$r_0 = b - Ax_0; \gamma_0 = \|r_0\|$;
$c_0 = 1$;
$v_1 = r_0/\gamma_0$;
$z_1 = x_0/\gamma_0$;
$\hat{F} = v_1^* F; \hat{G} = v_1^* G$;
$\tilde{F}_1 = v_1 v_1^* F$;
$v_1' = Av_1 - \tilde{F}\hat{G}^*$;
$t_{1,1} = v_1^* v_1'; v_1' = v_1' - t_{1,1} v_1$;
$t_{2,1} = \|v_1'\|; v_2 = v_1'/t_{2,1}$;
$p^* = \hat{F}; W = x_0\hat{F}/\gamma_0$;
$\tau_1 = t_{1,1}^*$;
$c_1 = \tau_1/(|\tau_1|^2 + t_{2,1}^2)^{1/2}; s_1 = t_{2,1}/(|\tau_1|^2 + t_{2,1}^2)^{1/2}$;
$\gamma_1 = -s_1\gamma_0$;
$z_2 = -(v_1 + t_{1,1}^* z_1)/t_{2,1}$;
$x_1 = s_1^2 x_0 + \gamma_1 c_1^* z_2$;
for $j = 2, ..., m$ **do**
  $\hat{F} = v_j^* F; \hat{G} = v_j^* G$;
  $\tilde{F}_j = \tilde{F}_{j-1} + v_j v_j^* F$;
  $v_j' = Av_j - \tilde{F}_j \hat{G}^*$;
  $t_{j-1,j} = v_{j-1}^* v_j'; v_j' = v_j' - t_{j-1,j} v_{j-1}$;
  $t_{j,j} = v_j^* v_j'; v_j' = v_j' - t_{j,j} v_j$;
  $t_{j+1,j} = \|v_j'\|; v_{j+1} = v_j'/t_{j+1,j}$;
  $p^* = -s_{j-1} p^* + c_{j-1}\hat{F}; W = W + z_j\hat{F}$;
  $\tau_j = c_{j-1} t_{j,j} - s_{j-1} c_{j-2} t_{j-1,j} + p^*\hat{G}^*$;
  $c_j = \tau_j/(|\tau_j|^2 + t_{j+1,j}^2)^{1/2}; s_j = t_{j+1,j}/(|\tau_j|^2 + t_{j+1,j}^2)^{1/2}$;
  $\gamma_j = -s_j\gamma_{j-1}$;
  $z_{j+1} = -(v_j + t_{j,j}^* z_j + t_{j-1,j} z_{j-1,j} + W\hat{G}^*)/t_{j+1,j}$;
  $x_j = s_j^2 x_{j-1} + \gamma_j c_j^* z_{j+1}$;
  if $|\gamma_j| \leq \varepsilon$ **then**
    break;
  end
end
$x = x_j$;

Algorithm 4: Progressive generalised Minimal RESidual (PGMRES).

As for the storage requirement, additional to the $(2s+2)n$ vectors from the Arnoldi process we need to store p^*, z_j, z_{j-1} and x_j and thus in total $(2s+6)n$ storage locations for vectors. Besides this one needs to store W which is a $n \times s$ matrix and therefore we would need $(3s+6)n$ storage locations. However, using the structure of F, G we do not need to store G as a matter of fact since we only need $\hat{G}$ and we can compute it from $\hat{F}$. Thus we need to store $(2s+6)$ n-vectors and the matrix A.

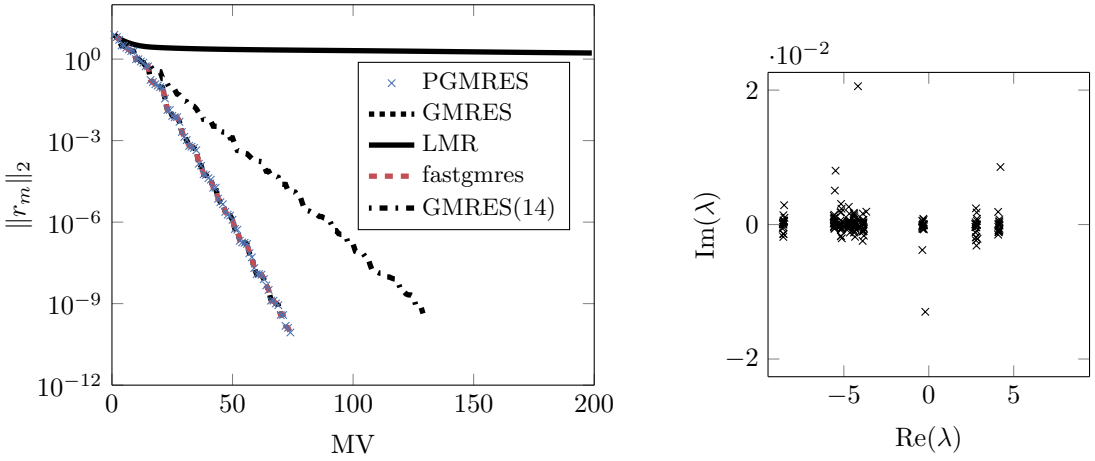
6.3. **Performance PGMRES.** The examples tested by Beckermann and Reichel could not be implemented so we decided to test another class of examples. In order to compare the code written for PGMRES to the original implementation by Beckermann we used their MATLAB code as well

which goes under the name *fastgmres*. We compare the PGMRES code with LMR and GMRES as well in order to see how optimal it behaves. Furthermore we test it against the restarted GMRES variant with equal memory requirements, namely GMRES($2s + 6$) which restarts every $2s + 6$ iterations.

6.3.1. *Example 6.1.* As a variation on EFxample 5.3 we propose a matrix H which has clustered eigenvalues along the real axis in order to be a Hermitian matrix. Then we perturb this matrix by a low rank s random matrix FG^* to get

$$A = H + FG^* \tag{49}$$

First we propose to take $\|FG^*\|_2$ small in norm as this will later on prove to be important[4]. The result is shown in Figure 7. We can see that the convergence of PGMRES is equal to that of GMRES whereas it is faster than GMRES($2s + 6$) and LMR.



(A) Residual norms as a function of the number of matrix vector multiplications.

(B) Eigenvalue distribution for matrix A in Example 6.1.

FIGURE 7. Example 6.1 with $n = 200$, $s = 2$ and $\|FG^*\|_2 = 1$.

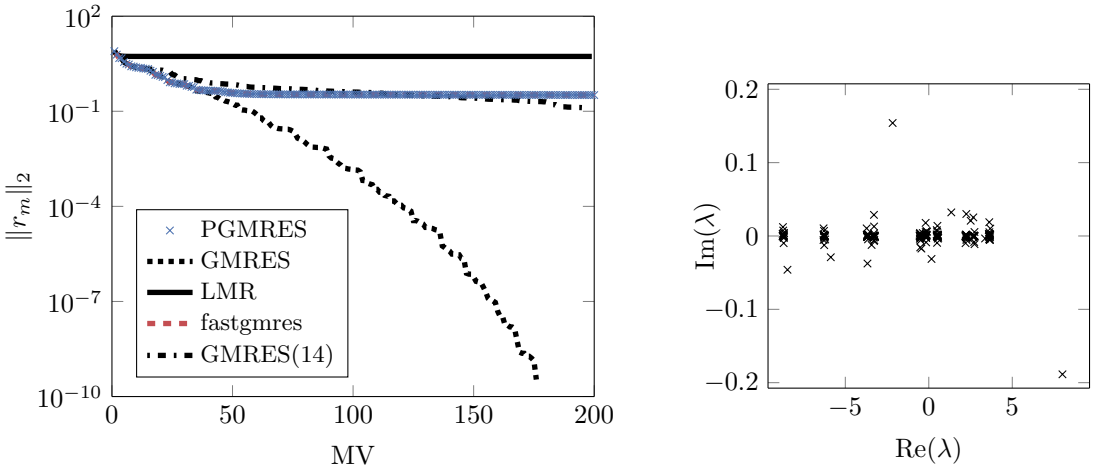
However if we choose $\|FG^*\|_2$ to become of modest size (in the order of $\|H\|_2 \approx 10$) than we observe stagnation in convergence of the PGMRES method as we can see in Figure 8.

This stagnation is probably again due to a rapid loss of orthogonality of the Arnoldi basis for the PGMRES method as can be seen in Figure 9. This effect was observed in 2012 by Embree, Sifuentes, Soodhalter, Szyld and Xue [5]. In their paper they discuss the drawbacks of PGMRES and point out that if the low rank perturbation is not of small norm then there is a stagnation of convergence of the PGMRES method.

6.3.2. *Example 6.2.* They showed this by considering a very simple example which we now will consider. The first ingredient discovered by Embree et al. to stagnate the PGMRES convergence is as we said $\|FG^*\|_2 \gtrsim \|H\|_2$ to create an unstable Arnoldi basis. They furthermore add the following to their example, construct a matrix such that GMRES first converges slowly which has the effect that the PGMRES Arnoldi basis looses its orthogonality. After this slow convergence GMRES should converge fast and as a result of the ill-conditioned basis for PGMRES this method is not able to follow and thus stagnates.

One simple way to achieve this is by specifying $\alpha, \beta, \gamma \in \mathbb{R}_{>0}$ with $\alpha < \beta$ and a $p, m \in \mathbb{N}$ with $p \ll m$ and then construct a matrix with p eigenvalues uniformly space in $[-\beta, -\alpha]$, m eigenvalues uniformly space in $[\alpha, \beta]$ and two more eigenvalues $\pm\gamma i$. We know that we can split this matrix in a Hermitian part H which has all only the real eigenvalues and two times 0 instead of the imaginary eigenvalues. Then we know that since H is normal (Hermitian matrices are normal) we have that $\|H\|_2$ is equal to the spectral radius of H and thus equal to β. To get our full matrix A we perturb our matrix with a rank 2 matrix having eigenvalues $\pm\gamma_i$ and zeros which is thus

[4]We take the 2-norm since we use the standard Euclidean inner product in our tests.



(A) Residual norms as a function of the number of matrix vector multiplications.

(B) Eigenvalue distribution for matrix A in Example 6.1.

FIGURE 8. Adapted Example 6.1 with $n = 200$, $s = 2$ and $\|FG^*\|_2 = 10$.

skew-Hermitian (and thus normal). The norm of this matrix is than equal to γ. The result for $\alpha = 0.125$, $\beta = 1$, $\gamma = 4$, $p = 6$, $m = 192$ can be seen in Figure 10. The condition number of this matrix A is quite low actually, namely $\text{cond}(A) = 32$. And thus even for well-conditioned problems PGMRES stagnates and does not converge.

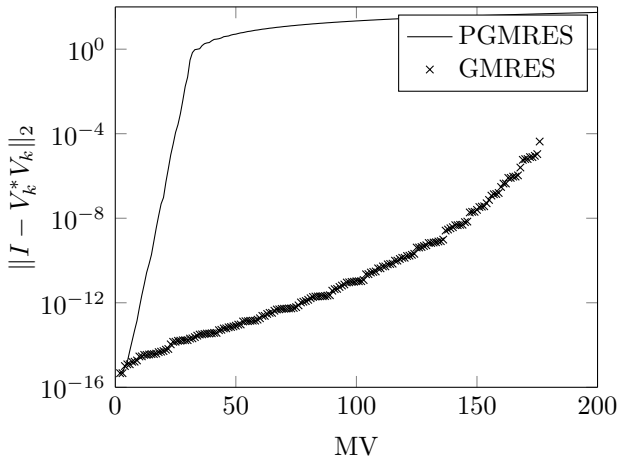
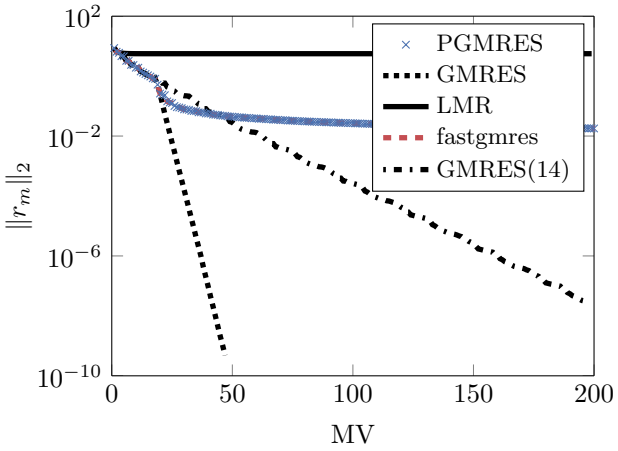


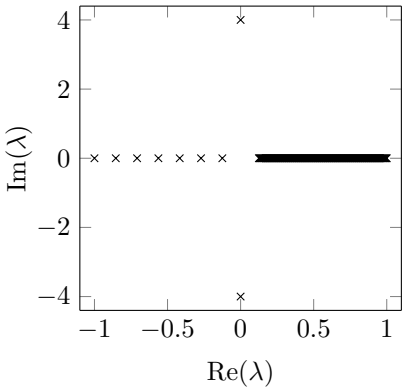
FIGURE 9. Orthogonality comparison of the Arnoldi bases of PGRMES and GMRES in case of $\|FG^*\|_2 = 10$.

6.4. **Lack of convergence PGMRES.** Embree et al. show in their paper [5] how an error in the k-th iterate can be enlarged if the norm of the low rank perturbation is not small. This is due to the fact that matrix vector products Av will get dominated by FG^*v. The angle between the true Arnoldi vectors and the computed Arnoldi vectors will grow due to rounding errors thus resulting in the observed lack of convergence. Therefore we see that for PGMRES we not only need to have a low rank perturbation of a Hermitian matrix but this perturbation has to be of small norm as well. If this holds and GMRES converges sufficiently fast, so that the PGMRES basis will not degrade too much, then PGMRES will converge at roughly the same pace and will therefore outperform GMRES.

Other short-recurrence methods for nearly-Hermitian matrices can be found as well. Barth and Manteuffel proposed a Multiple Recursion Conjugate Gradient algorithm (MRCG) for generalised B-normal(l, m) matrices which should therefore also work for nearly Hermitian matrices. Also in



(A) Residual norms as a function of the number of matrix vector multiplications.



(B) Eigenvalue distribution for matrix A in Example 6.2.

FIGURE 10. Example 6.2 with $\alpha = 0.125$, $\beta = 1$, $\gamma = 4$, $p = 6$, $m = 192$.

their paper Embree et al. propose another short-recurrence method with three term recurrences for nearly-Hermitian matrices which is supposed to avoid many of the instabilities of PGMRES which is called a Schur complement method. However these algorithms are beyond the scope of this report.

7. Conclusion

When solving sparse linear systems by use of optimal Krylov methods there can be some structure of the involved matrix A which can be exploited to speed-up convergence. In particular one could try to minimise the amount of recurrence relations and thus create short-recurrence methods. We showed that for certain classes of matrices this is indeed possible and one of the first results classifying a subset of these matrices was the Faber-Manteuffel theorem [6]. However practically this theorem states that if only if the eigenvalues of A are collinear then there exists a three-term 'Conjugate Gradient-like' recurrence relation. By further generalising this result to the generalised B-normal(l, m) matrices we showed we could get short-recurrences as well. These recurrences are not all of a 'CG-like' short recurrence relation though. This is nicely demonstrated by taking shifted and scaled unitary matrices which are I-normal$(1, 1)$. By use of the Isometric Arnold Process one can construct an orthonormal basis only using three-term recurrences. If incorporated in a minimal residual method one can get the SUMR method. This method is in exact arithmetic equivalent to GMRES, but is computationally much more favourable. The convergence behaviour of SUMR is equal to that of GMRES in all of our computed examples if we incorporate explicit normalisation of the vectors in the process. This normalisation partly clarifies the differences between our results and that of the Jagels and Reichel paper [10]. Therefore SUMR is a very good option for solving shifted unitary matrices.

We also showed that in order to solve a system of nearly Hermitian matrices, which are also generalised B-normal(l, m) matrices, one could apply the PGMRES method proposed by Beckermann and Reichel. This method has low memory requirements and uses short recursions which makes it a computationally cheap method. However, this comes at a price. We showed following Embree et al. [5] that when the perturbation of the Hermitian part is not small in norm then the PGMRES method does not converge and is not the method to use. When GMRES converges at a steady rate and the perturbation is of low rank and small in norm PGMRES becomes a favourable and fast converging method.

References

1. Arnold, G. *et al. Numerical methods for the QCD overlap operator. II: Optimal Krylov subspace methods.* 2005.
2. Barth, T. L. & Manteuffel, T. A. *Conjugate Gradient Algorithms Using Multiple Recursions* in *University of Washington* (1995), 9–14.
3. Barth, T. L. & Manteuffel, T. A. Multiple Recursion Conjugate Gradient Algorithms Part I: Sufficient Conditions. *SIAM J. Matrix Anal. Appl.* **21,** 768–796 (2000).
4. Beckermann, B. & Reichel, L. The Arnoldi Process and GMRES for Nearly Symmetric Matrices. *SIAM J. Matrix Analysis Applications* **30,** 102–120 (2008).
5. Embree, M., Sifuentes, J. A., Soodhalter, K. M., Szyld, D. B. & Xue, F. Short-Term Recurrence Krylov Subspace Methods for Nearly Hermitian Matrices. *SIAM. J. Matrix Anal. and Appl.* **33-2,** 480–500 (2012).
6. Faber, V. & Manteuffel, T. Necessary and sufficient conditions for the existence of a conjugate gradient method. *SIAM J. NUMER. ANAL.* **21,** 352–362 (1984).
7. Faber, V., Liesen, J. & Tichý, P. The Faber-Manteuffel Theorem for Linear Operators. *SIAM J. Numer. Anal.* **46,** 1323–1337 (2008).
8. Gragg, W. B. Positive Definite Toeplitz Matrices, the Arnoldi Process for Isometric Operators, and Gaussian Quadrature on the Unit Circle. *J. Comput. Appl. Math.* **46,** 183–198 (1993).
9. Hestenes, M. R. & Stiefel, E. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards* **49,** 409–436 (1952).
10. Jagels, C. F. & Reichel, L. A fast minimal residual algorithm for shifted unitary matrices. *Numer. Linear Algebra Appl.* **1(6),** 555–570 (1994).
11. Liesen, J. When is the Adjoint of a Matrix a Low Degree Rational Function in the Matrix? *SIAM J. Matrix Anal. Appl.* **29,** 1171–1180 (2007).
12. Liesen, J. & Strakoš, Z. On optimal short recurrences for generating orthogonal Krylov subspace bases. *SIAM Rev.* **50,** 485–503 (2008).
13. Saad, Y. & Schultz, M. H. GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems. *SIAM J. Sci. Stat. Comput.* **7,** 856–869 (1986).
14. Stoll, M. *Lecture notes Linear Algebra II, Course No. 100 222, Spring 2007* (2007).

Appendix A. Code

Matlab code for the algorithms discussed and to generate the figures is freely available for personal use from https://bitbucket.org/CasperBeentjes/short-recurrence-optimal-krylov-solvers.

Mathematical Institute, University of Oxford, Oxford, UK
E-mail address: beentjes@maths.ox.ac.uk