Approximating the matrix exponential of an advection-diffusion operator using the incomplete orthogonalization method

Antti Koskela

KTH Royal Institute of Technology, Lindstedtvägen 25, 10044 Stockholm, Sweden
akoskela@kth.se

Abstract. In this paper we give first results for the approximation of $\mathrm{e}^{A}b$, i.e. the matrix exponential times a vector, using the *incomplete orthogonalization method.* The benefits compared to the Arnoldi iteration are clear: shorter orthogonalization lengths make the algorithm faster and a large memory saving is also possible. For the case of three term orthogonalization recursions, simple error bounds are derived using the norm and the field of values of the projected operator. In addition, an a posteriori error estimate is given which in numerical examples is shown to work well for the approximation. In the numerical examples we particularly consider the case where the operator A arises from spatial discretization of an advection-diffusion operator.

1 Introduction

An efficient numerical computation of the product $\mathrm{e}^{A}b$ for a matrix $A \in \mathbb{C}^{n\times n}$ and a vector $b \in \mathbb{C}^{n}$ is of importance in several fields of applied mathematics. For various applications and numerical methods, see [3].

One large source of problems of this form comes from the implementation of exponential integrators [5]. These integrators have been shown to be particularly efficient for ODEs coming from a spatial semidiscretization of semilinear PDEs. In this case A is usually sparse and has a large norm and dimension. A widely used approach in this case are *Krylov subspace methods,* see e.g. [4] and [8].

Krylov subspace methods are based on the idea of projecting a matrix $A \in \mathbb{C}^{n\times n}$ and a vector $b \in \mathbb{C}^{n}$ onto a lower dimensional subspace $\mathcal{K}_k(A,b)$ defined by

$$\mathcal{K}_k(A,b) = \operatorname{span}\{b, Ab, A^2b, \ldots, A^{k-1}b\}.$$

The Arnoldi iteration performs a Gram–Schmidt orthogonalization for this subspace and gives an orthonormal matrix $Q_k = [q_1,\ldots,q_k] \in \mathbb{C}^{n\times k}$ which provides a basis of $\mathcal{K}_k(A,b)$, and a Hessenberg matrix $H_k = Q_k^{*}AQ_k \in \mathbb{C}^{k\times k}$,

The work was mostly carried out at the University of Innsbruck, Austria, and was supported by the FWF doctoral program 'Computational Interdisciplinary Modelling' W1227.

which represents the action of A in the subspace $\mathcal{K}_k(A,b)$. If A is Hermitian or skew-Hermitian, H_k will be tridiagonal and we get the Lanczos iteration. Moreover, the recursion

$$AQ_k = Q_k H_k + h_{k+1,k} q_{k+1} e_k^{\mathsf{T}}$$

holds, where $h_{k+1,k}$ denotes the corresponding entry in H_{k+1} and e_k is the kth standard basis vector in $\mathbb{C}^k$.

Using the basis Q_k and the Hessenberg matrix H_k, the product $\mathrm{e}^{A}b$ can be then approximated as (see e.g. [8])

$$\mathrm{e}^{A}b \approx Q_k \mathrm{e}^{H_k} e_1 \|b\|.$$

In case that A is not (skew-)Hermitian the Arnoldi iteration has to be used. The drawback of this approach is that the orthogonalization recursions grow longer which slows down the iteration, and that it needs increasingly memory as k grows. As a remedy for this the *restarted Krylov subspace method* has been proposed [2].

The objective of this paper is to show that the *incomplete orthogonalization method* is a good alternative for approximating the product $\mathrm{e}^{A}b$ for nonnormal matrices A when long orthogonalization recursions should be avoided. The method has been considered before for eigenvalue problems [7] and for solving linear systems [9]. As the numerical experiments and the short analysis of this paper show, it also provides a good alternative for approximating the matrix exponential.

Class of test problems A reasonable example of nonnormal large and sparse matrices is obtained from the spatial discretization of the 1-d advection-diffusion equation

$$\partial_t u = \epsilon \partial_{xx} u + \alpha \partial_x u. \tag{1}$$

Choosing Dirichlet boundary conditions on the interval $[0,1]$ and performing the discretization using central finite differences gives the ordinary differential equation $y' = Ay$, where the operator is the form $A = \epsilon \Delta_n + \alpha \nabla_n \in \mathbb{R}^{n \times n}$ with

$$\Delta_n = \frac{1}{(\Delta x)^2} \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}, \quad \nabla_n = \frac{1}{2\Delta x} \begin{bmatrix} & -1 & & & \\ 1 & & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & & -1 \\ & & & 1 & \end{bmatrix}, \tag{2}$$

where $\Delta x = 1/(n+1)$. We define the grid Péclet number

$$\mathrm{Pe} = \frac{\alpha \Delta x}{2\epsilon}$$

as in the numerical comparisons of [1]. By the Péclet number the nonnormality of A can be controlled.

Throughout the paper, $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product and $\| \cdot \|$ denotes the corresponding norm or its induced matrix norm. The Hermitian part of a matrix A is defined as $A^{\mathrm{H}} = (A^* + A)/2$, and the skew-Hermitian part as $A^{\mathrm{S}} = (A^* - A)/2$.

2 The incomplete orthogonalization method

In the *incomplete orthogonalization method* (IOM) (see e.g. [7]), Aq_i is orthogonalized at step i only against m previous vectors $\{q_{i-m+1}, \dots, q_i\}$ instead of all the previous basis vectors. The coefficients h_{ij} are collected as in the Arnoldi iteration. The incomplete orthogonalization method with orthogonalization length m is denoted as IOM(m) for the rest of the paper. As a result of k steps of IOM(m) we get the matrix

$$Q_{k,m} = \begin{bmatrix} q_1 \dots q_k \end{bmatrix}$$

giving the basis of $\mathcal{K}_k(A, b)$, where $q_1 = b / \|b\|$, and the vectors q_i are orthogonal locally, i.e.,

$$\begin{aligned} &\langle q_i, q_i \rangle = 1, \\ &\langle q_i, q_j \rangle = 0, \quad \text{if} \quad |i - j| \leq m, \quad i \neq j. \end{aligned}$$

The iteration also gives a Hessenberg matrix with an upper bandwidth length m,

$$H_{k,m} := \begin{bmatrix} h_{11} \dots h_{1m} & & & 0 \\ h_{21} \ddots & & \ddots & \\ \ddots & \ddots & & h_{k-m+1,k} \\ & \ddots & \ddots & \vdots \\ 0 & & h_{k,k-1} & h_{kk} \end{bmatrix}, \tag{3}$$

where the nonzero elements are given as

$$h_{ij} = \langle Aq_j, q_i \rangle.$$

We can see that by construction the following relation holds

$$AQ_{k,m} = Q_{k,m} H_{k,m} + h_{k+1,k} q_{k+1} e_k^{\mathsf{T}}. \tag{4}$$

It is easy to verify that if $\dim \mathcal{K}_k(A, b) = k$, then $\mathcal{K}_k(A, b) \subset R(Q_{k,m})$, where $R(Q_{k,m})$ denotes the range of $Q_{k,m}$. However, if it happens that $R(Q_{k+1}) = R(Q_k)$, the subdiagonal element $h_{k,k+1}$ will not necessarily be zero like in the case of the Arnoldi iteration.

2.1 Polynomial approximation property

Using (4) recursively we see that for $0 \leq j \leq k-1$

$$A^j Q_{k,m} = Q_{k,m} H_{k,m}^j + \sum_{i=0}^{j-1} c_i e_{k-i}^{\mathsf{T}}$$

for some vectors c_i. Multiplying this by e_1 from the right side, we see that for $0 \leq j \leq k-1$ it holds

$$A^j b = Q_{k,m} H_{k,m}^j e_1 \|b\|.$$

This results as the following lemma.

Lemma 1. *Let $A \in \mathbb{C}^{n \times n}$ and let $Q_{k,m}$, $H_{k,m}$ be the results of k steps of IOM(m) applied to A with starting vector b. Then for any polynomial p_{k-1} of degree up to $k-1$ the following equality holds:*

$$p_{k-1}(A)b = Q_{k,m} p_{k-1}(H_{k,m}) e_1 \|b\|.$$

This leads us to make the approximation

$$\mathrm{e}^{A} b \approx Q_{k,m} \mathrm{e}^{H_{k,m}} e_1 \|b\|. \tag{5}$$

By Lemma 1, the error ϵ_k of this approximation is given as

$$\epsilon_k = \sum_{\ell=k}^{\infty} \frac{A^\ell}{\ell!} b - Q_{k,m} \sum_{\ell=k}^{\infty} \frac{H_{k,m}^\ell}{\ell!} e_1 \|b\|. \tag{6}$$

3 Bounds for the error

To bound the error (6), we consider bounds using the norm and the field of values of the Hessenberg matrix $H_{k,2}$.

Using the representation (see also the analysis of [8])

$$\sum_{\ell=k}^{\infty} \frac{x^\ell}{\ell!} = x^k \int_0^1 \mathrm{e}^{(1-\theta)x} \frac{\theta^{k-1}}{(k-1)!} \, \mathrm{d}\theta$$

and the bound $\|\mathrm{e}^{A}\| \leq \mathrm{e}^{\mu(A)}$, where $\mu(A)$ is the numerical abscissa of A, i.e., the largest eigenvalue of A^{H} (see [3, Thm. 10.11]), we get the bound

$$\|\epsilon_k\| \leq \frac{\mathrm{e}^{\mu(A)} \|A\|^k + \mathrm{e}^{\mu(H_{k,m})} \|Q_{k,m}\| \|H_{k,m}\|^k}{k!} \|b\|. \tag{7}$$

Note that in the case of the advection-diffusion operator (2), $\mu(A) \leq 0$.

In (7) the norm $\|Q_{k,m}\|$ cannot be bounded, in general. In numerical experiments $\|Q_{k,m}\|$ was found to stay of order 1 for advection-diffusion operators of the form (2) for all values of n, Pe, k and m. A discussion on the effect of the parameter m to the size of $\|Q_{k,m}\|$ can be found in [9].

In the same way as in the analysis of [8], it can be shown that

$$Q_{k,m}\mathrm{e}^{H_{k,m}}e_1\|b\| = p(A)b,$$

where p is the unique polynomial that interpolates the exponential function in the Hermite sense at the eigenvalues of $H_{k,m}$. Then, if the field of values $\mathcal{F}(H_{k,m})$ can be bounded with respect to $\mathcal{F}(A)$, the superlinear convergence of the approximation can be shown in the same way as in the proof of [2, Thm. 4.2] for the case of the restarted Krylov method.

When viewing the incomplete orthogonalization method as an *oblique projection method* [9], also the results [4, Lemma 7 and 8] can be applied.

3.1 Bounds for the field of values and the norm of $H_{k,2}$.

In this subsection we show how to bound the norm and the field of values of the Hessenberg matrix $H_{k,2}$, i.e., for the case of IOM(2). The field of values of a matrix $A \in \mathbb{C}^{n\times n}$ is defined as

$$\mathcal{F}(A) = \{x^*Ax \,:\, x \in \mathbb{C}^n,\, \|x\|_2 = 1\}.$$

We first give the following auxiliary lemma.

Lemma 2. *Let $A \in \mathbb{C}^{n\times n}$ be normal, and let the 0-field of values be defined as*

$$\mathcal{F}_0(A) = \{\langle x, Ay\rangle \,:\, \langle x, y\rangle = 0\,,\, \|x\| = \|y\| = 1\}.$$

Then,

$$\mathcal{F}_0(A) = \{z \in \mathbb{C} \,:\, |z| \leq r\}, \quad \textit{where} \quad r = \frac{1}{2}\max_{\lambda_i,\lambda_j\in\sigma(A)} |\lambda_i - \lambda_j|\,.$$

Proof. Since A is normal, it is unitary similar to a diagonal matrix with the eigenvalues of A on the diagonal, $A = U\Lambda U^*$. Let $x, y \in \mathbb{C}^{n\times n}$ such that $\langle x, y\rangle = 0\,,\, \|x\| = \|y\| = 1$. Then $\langle x, Ay\rangle = \langle U^*x, (\Lambda - cI)U^*y\rangle$ for all $c \in \mathbb{C}$. By choosing c to be the center of the smallest disc containing $\sigma(A)$, and by using the Cauchy–Schwartz inequality, we see that

$$|\langle x, Ay\rangle| \,\leq\, \|\Lambda - cI\| \,\leq\, \frac{1}{2}\max_{\lambda_i,\lambda_j\in\sigma(A)} |\lambda_i - \lambda_j|\,.$$

By choosing $x = u_i/\sqrt{2} + u_j/\sqrt{2}$ and $y = u_i/\sqrt{2} - u_j/\sqrt{2}$, where u_i and u_j are eigenvectors of A corresponding to eigenvalues λ_i and λ_j, we see that

$$\langle x, Ay\rangle = \frac{1}{2}(\lambda_i - \lambda_j).$$

Thus, the inequality above is sharp. Since $\mathcal{F}_0(A)$ is a disc centered at the origin [6], the claim follows. □

Using Lemma 2, we may now obtain a bound for the field of values of $H_{k,2}$.

Theorem 3. *Let $A \in \mathbb{C}^{n\times n}$ and let $H_{k,2}$ be the Hessenberg matrix obtained after k steps of IOM(2) applied to A. Then it holds that*

$$\mathcal{F}(H_{k,2}) \subset \{\, z \in \mathbb{C} \,:\, d(z,\mathcal{F}(A)) \leq \tfrac{1}{2}(\|A^H\| + \|A^S\|)\,\}.$$

Proof. First, we extend $H_{k,2}$ to a matrix $\widetilde{H} \in \mathbb{C}^{(k+2)\times(k+2)}$ by adding zeros such that

$$\widetilde{H} = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & H_{k,2} & \vdots \\ 0 & \dots & 0 \end{bmatrix}$$

and set $q_0 = q_{k+1} = 0$. It clearly holds that $\mathcal{F}(H_{k,2}) \subset \mathcal{F}(\widetilde{H})$.

Let $x = \begin{bmatrix} x_0 \dots x_{k+1} \end{bmatrix}^{\mathsf{T}} \in \mathbb{C}^{k+2}$, $\|x\| = 1$. Then, by inspecting (3), we see that

$$\begin{aligned} x^* \widetilde{H} x = {} & \frac{1}{2} \sum_{i=0}^{k} \begin{bmatrix} x_i \\ x_{i+1} \end{bmatrix}^* \begin{bmatrix} \langle Aq_i, q_i \rangle & \langle Aq_{i+1}, q_i \rangle \\ \langle Aq_i, q_{i+1} \rangle & \langle Aq_{i+1}, q_{i+1} \rangle \end{bmatrix} \begin{bmatrix} x_i \\ x_{i+1} \end{bmatrix} \\ & + \frac{1}{2} \sum_{i=1}^{k-1} \begin{bmatrix} x_i \\ x_{i+1} \end{bmatrix}^* \begin{bmatrix} 0 & \langle Aq_{i+1}, q_i \rangle \\ \langle Aq_i, q_{i+1} \rangle & 0 \end{bmatrix} \begin{bmatrix} x_i \\ x_{i+1} \end{bmatrix}. \end{aligned} \tag{8}$$

Due to the local orthogonality of the basis vectors $\{q_i\}$ and the convexity of $\mathcal{F}(A)$, we see that the first term of (8) is in $\mathcal{F}(A)$. For the second term, we split $A = A^{\mathrm{H}} + A^{\mathrm{S}}$ and use the Lemma 2 to see that

$$\left| \begin{bmatrix} x_i \\ x_{i+1} \end{bmatrix}^* \begin{bmatrix} 0 & \langle Aq_{i+1}, q_i \rangle \\ \langle Aq_i, q_{i+1} \rangle & 0 \end{bmatrix} \begin{bmatrix} x_i \\ x_{i+1} \end{bmatrix} \right| \leq |x_i|\,|x_{i+1}|\,(\|A^{\mathrm{H}}\| + \|A^{\mathrm{S}}\|).$$

By the inequality $\sum_{i=1}^{k-1} |x_i|\,|x_{i+1}| \leq \sum_{i=1}^{k} |x_i|^2$, the claim follows. □

Using Lemma 2 we now obtain also a bound for $\|H_{k,2}\|$.

Theorem 4. *Let $A \in \mathbb{C}^{n\times n}$ and let $H_{k,2}$ be the Hessenberg matrix obtained after k steps of IOM(2) applied to A. Then it holds that*

$$\|H_{k,2}\| \leq r(A) + \tfrac{1}{2}(\|A^H\| + \|A^S\|),$$

where $r(A) = \max\limits_{z \in \mathcal{F}(A)} |z|$.

Proof. Let $x \in \mathbb{C}^n$, $\|x\| = 1$. Then for $1 < i < k$ it holds that

$$(H_{k,2}x)_i = x_i \langle Aq_i, q_i \rangle + \langle Aq_i,\, (x_{i-1}q_{i-1} + x_{i+1}q_{i+1}) \rangle.$$

By using the triangle inequality, splitting $A = A^{\mathrm{H}} + A^{\mathrm{S}}$, local orthogonality of the vectors $\{q_i\}$ and Lemma 2, the claim follows. □

Although we consider in the analysis of $\mathcal{F}(H_{k,m})$ and $\|H_{k,m}\|$, and also in the numerical comparisons only the case $m = 2$, we note that in numerical experiments the approximation (5) was found to improve for increasing m.

4 A posteriori error estimate

An a posteriori error estimate follows from the relation (4) and can be derived in the same way as the estimate for the Arnoldi iteration, see [8, Thm. 5.1].

Theorem 5. *The error produced by the incomplete orthogonalization method of* $\mathrm{e}^{A}b$ *satisfies the expansion*

$$\mathrm{e}^{A}b - Q_{k,m}\exp(H_{k,m})e_1 = h_{k+1,k}\sum_{\ell=1}^{\infty} e_k^{\mathsf{T}}\varphi_\ell(H_{k,m})e_1 A^{\ell-1}q_{k+1},$$

where $\varphi_\ell(z) = \sum_{k=0}^{\infty}\frac{z^k}{(k+\ell)!}$. In numerical experiments we estimate the error using the norm of the first term, i.e. by using the estimate

$$\|\epsilon_k\| \approx h_{k+1,k}\left|e_k^{\mathsf{T}}\varphi_1(H_{k,m})e_1\right|, \tag{9}$$

which can be obtained with small computational cost by computing the exponential of

$$\widetilde{H}_m = \begin{bmatrix} H_{k,m} & e_1 \\ 0 & 0 \end{bmatrix}, \quad \text{since} \quad \mathrm{e}^{\widetilde{H}_m} = \begin{bmatrix} \mathrm{e}^{H_{k,m}} & \varphi_1(H_{k,m})e_1 \\ 0 & 1 \end{bmatrix}.$$

5 Numerical examples

For the first example, we take $A = \epsilon\Delta_n + \alpha\nabla_n \in \mathbb{R}^{n\times n}$, where Δ_n and ∇_n are as in (2). The vector b is taken as a discretization of the function $u_0(x) = 16((1-x)x)^2$, $x \in [0,1]$. We set $n = 400$ and $\epsilon = 1$, and consider the cases of a weak advection and a strong advection. We approximate the product $\mathrm{e}^{hA}b$ using IOM(2) and compare it with the standard Arnoldi iteration and the restarted Krylov method with restarting interval 3. We also compute the estimate (9) for IOM. Figure 1 shows the convergence of the three methods.

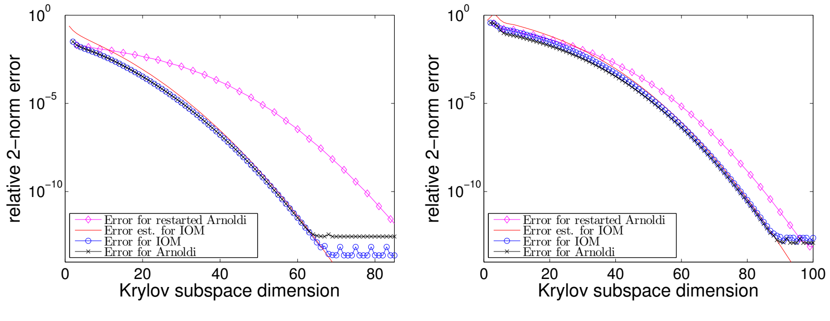

Fig. 1. Left: $h = 3\cdot 10^{-4}$, $\mathrm{Pe} = 6.2\cdot 10^{-3}$. Right: $h = 2\cdot 10^{-4}$, $\mathrm{Pe} = 10.0$.

In the second example, A, n and ϵ are as above, and b is taken randomly. We compare the methods using larger h for the cases of a weak advection and a mild advection. Figure 2 shows the convergence of the three methods.

The differences in the computational costs come mainly from the differences in the lengths of the orthogonalization recursions, the Arnoldi iteration taking $\mathcal{O}(k^2)$ and the other two methods $\mathcal{O}(k)$ inner products. In these numerical examples, the Arnoldi iteration was for $k = 50$ about 4 times slower and for $k = 100$ about 8 times slower than the other two methods.

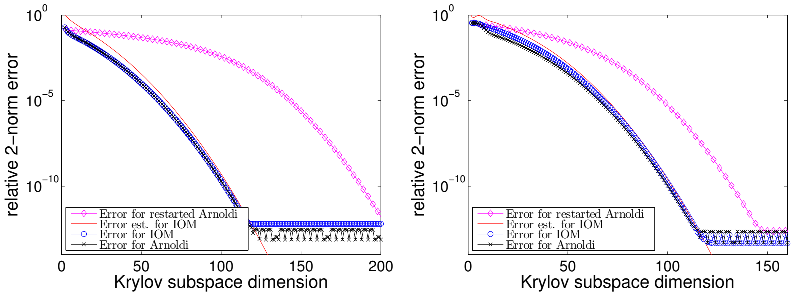

Fig. 2. Left: $h = 1 \cdot 10^{-3}$, Pe $= 6.2 \cdot 10^{-3}$. Right: $h = 6 \cdot 10^{-4}$, Pe $= 1.3 \cdot 10^{-1}$.

References

1. M. Caliari, P. Kandolf, A. Ostermann, and S. Rainer , *Comparison of software for computing the action of the matrix exponential*, BIT (2013), pp. 1–16.
2. M. Eiermann and O.G. Ernst , *A restarted Krylov subspace method for the evaluation of matrix functions*, SIAM J. Numer. Anal., 44 (2005), pp. 2481–2504.
3. N.J. Higham, *Functions of matrices: theory and computation*, SIAM, 2008.
4. M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.
5. M. Hochbruck and A. Ostermann, *Exponential integrators*, Acta Numer. 19 (2010), pp. 209–286.
6. R.A. Horn and C.R. Johnson, *Topics in matrix analysis*, Cambridge University Press (1991).
7. Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebr. Appl. 34 (1980), pp. 269–295.
8. Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal. 29 (1992), pp. 209–228.
9. Y. Saad, *The Lanczos Biorthogonalization Algorithm and Other Oblique Projection Methods for Solving Large Unsymmetric Systems*, SIAM Journal on Numerical Analysis 19 (1982), pp. 485–506.