A low-memory Lanczos method with rational Krylov compression for matrix functions

Angelo A. Casulli, Igor Simunec

Abstract

A fundamental problem in numerical linear algebra is the approximation of the action of a matrix function $f(A)$ on a vector $\boldsymbol{b}$, where $A \in \mathbb{C}^{n\times n}$ is a matrix that is typically large and sparse, $\boldsymbol{b} \in \mathbb{C}^n$ is a vector and f is a function defined on the spectrum of A. In this work, we focus on the case of a Hermitian matrix A. We recall that when A is Hermitian, given an eigendecomposition $A = UDU^H$, the matrix function $f(A)$ is defined as $f(A) = Uf(D)U^H$, where $f(D)$ is a diagonal matrix obtained by appliying f to each diagonal entry of D. We refer to [12] for an extensive discussion of matrix functions.

Popular methods for the approximation of $f(A)\boldsymbol{b}$ are polynomial [16, 13, 8, 7, 11] and rational Krylov methods [6, 15, 9, 1, 3]. The former only access A via matrix-vector products, while the latter require the solution of shifted linear systems with A. When the linear systems can be solved efficiently, rational Krylov methods can be more effective than polynomial Krylov methods since they usually require much fewer iterations to converge. However, there are several situations in which rational Krylov methods are not applicable, either because the matrix A is only available implicitly via a function that computes matrix-vector products, or because A is very large and the solution of linear systems is prohibitively expensive.

When A is Hermitian, the core component of a polynomial Krylov method is the Lanczos algorithm [14], which constructs an orthonormal basis $\mathbf{Q}_M = [\boldsymbol{q}_1 \ \dots \ \boldsymbol{q}_M]$ of the polynomial Krylov subspace $\mathcal{K}_M(A, \boldsymbol{b}) = \operatorname{span}\{\boldsymbol{b}, A\boldsymbol{b}, \dots, A^{M-1}\boldsymbol{b}\}$ by exploiting a short-term recurrence. The product $f(A)\boldsymbol{b}$ can then be approximated by the Lanczos approximation

$$\boldsymbol{f}_M := \mathbf{Q}_M f(\mathbf{T}_M)\boldsymbol{e}_1\|\boldsymbol{b}\|_2, \qquad \mathbf{T}_M := \mathbf{Q}_M^H A\mathbf{Q}_M, \tag{1}$$

where $\boldsymbol{e}_1$ denotes the first unit vector. The Lanczos algorithm uses a short-term recurrence in the orthogonalization step, so each new basis vector is orthogonalized only against the last two basis vectors, and only three vectors need to be kept in memory to compute the basis $\mathbf{Q}_M$. Although the basis $\mathbf{Q}_M$ and the projected matrix $\mathbf{T}_M$ can be computed by using the short-term recurrence that only requires storage of the last three basis vectors, forming the approximate solution $\boldsymbol{f}_M$ still requires the full basis $\mathbf{Q}_M$. When the matrix A is very large, there may be a limit on the maximum number of basis vectors that can be stored, so with a straightforward implementation of the Lanczos method there is a limit on the number of iterations of Lanczos that can be performed and hence on the attainable accuracy. In the literature, several strategies have been proposed to deal with low memory issues. See the recent surveys [10, 11] for a comparison of several low-memory methods.

In this presentation we propose a new low-memory algorithm for the approximation of $f(A)\boldsymbol{b}$. Our method combines an outer Lanczos iteration with an inner rational Krylov subspace, which is used to compress the outer Krylov basis whenever it reaches a certain size.

The fundamental insight underlying this method is that, leveraging the results presented in [2], the vector $\boldsymbol{f}_M$ defined in (1) (for simplicity, assuming $\|\boldsymbol{b}\|_2 = 1$) can be approximated by

$$\boldsymbol{f}_M \approx \mathbf{Q}_M \begin{bmatrix} f(T_1)\boldsymbol{e}_1 - U_1 f(U_1^H T_1 U_1)U_1^H \boldsymbol{e}_1 \\ 0 \end{bmatrix} + \mathbf{Q}_M \begin{bmatrix} U_1 & \\ & I \end{bmatrix} f\left(\begin{bmatrix} U_1^H & \\ & I \end{bmatrix} \mathbf{T}_M \begin{bmatrix} U_1 & \\ & I \end{bmatrix} \right) \begin{bmatrix} U_1^H \boldsymbol{e}_1 \\ 0 \end{bmatrix},$$

where T_1 is an $m \times m$ leading principal submatrix of $\mathbf{T}_M$, and U_1 is an orthonormal basis of a rational Krylov subspace generated using the small matrix T_1. One can observe that the first summand of this expression can be computed after m steps of the Lanczos algorithm. Moreover, once the first term has been computed, it is no longer necessary to keep all the first m columns of the matrix $\mathbf{Q}_M$ in memory, since computing the second term only requires the few vectors obtained by multiplying the first m columns of $\boldsymbol{Q}_M$ on the right by the matrix U_1. Finally, the second term can be computed by recursively applying the same procedure.

Similarly to [4], the inner rational Krylov subspace does not involve the matrix A, but only small matrices. This is fundamental, since constructing a basis of the inner subspace does not require the solution of linear systems with A, and hence it is cheap compared to the cost of the outer Lanczos iteration. Theoretical results show that the approximate solutions computed by our algorithm coincide with the ones constructed by the outer Krylov subspace method when f is a rational function, and for a general function they differ by a quantity that depends on the best rational approximant of f with the poles used in the inner rational Krylov subspace.

If the outer Krylov basis is compressed every m iterations and the inner rational Krylov subspace has k poles, our approach requires the storage of approximately $m + k$ vectors. Additionally, due to the basis compression, our approximation involves the computation of matrix functions of size at most $(m+k) \times (m+k)$, so the cost does not grow with the number of iterations. This represents an important advantage with respect to the Lanczos method, since when the number of iterations is very large the evaluation of f on the projected matrix can become quite expensive.

Numerical experiments show that the proposed algorithm is competitive with other low-memory methods based on polynomial Krylov subspaces.

The content of this presentation draws on the findings presented in [5].

References

[1] L. Aceto, D. Bertaccini, F. Durastante, and P. Novati. Rational Krylov methods for functions of matrices with applications to fractional partial differential equations. *J. Comput. Phys.*, 396:470–482, 2019.

[2] Bernhard Beckermann, Alice Cortinovis, Daniel Kressner, and Marcel Schweitzer. Low-rank updates of matrix functions II: rational Krylov methods. *SIAM J. Numer. Anal.*, 59(3):1325–1347, 2021.

[3] Michele Benzi and Igor Simunec. Rational Krylov methods for fractional diffusion problems on graphs. *BIT*, 62(2):357–385, 2022.

[4] Angelo A. Casulli, Daniel Kressner, and Leonardo Robol. Computing functions of symmetric hierarchically semiseparable matrices, 2024.

[5] Angelo A. Casulli and Igor Simunec. A low-memory Lanczos method with rational Krylov compression for matrix functions. *arXiv preprint arXiv:2403.04390*, 2024.

[6] Vladimir Druskin and Leonid Knizhnerman. Extended Krylov subspaces: approximation of the matrix square root and related functions. *SIAM J. Matrix Anal. Appl.*, 19(3):755–771, 1998.

[7] Andreas Frommer, Stefan Güttel, and Marcel Schweitzer. Efficient and stable Arnoldi restarts for matrix functions based on quadrature. *SIAM J. Matrix Anal. Appl.*, 35(2):661–683, 2014.

[8] Andreas Frommer and Valeria Simoncini. Matrix functions. In *Model Order Reduction: Theory, Research Aspects and Applications*, volume 13 of *Math. Ind.*, pages 275–303. Springer, Berlin, 2008.

[9] Stefan Güttel. Rational Krylov approximation of matrix functions: numerical methods and optimal pole selection. *GAMM-Mitt.*, 36(1):8–31, 2013.

[10] Stefan Güttel, Daniel Kressner, and Kathryn Lund. Limited-memory polynomial methods for large-scale matrix functions. *GAMM-Mitt.*, 43(3):e202000019, 19, 2020.

[11] Stefan Güttel and Marcel Schweitzer. A comparison of limited-memory Krylov methods for Stieltjes functions of Hermitian matrices. *SIAM J. Matrix Anal. Appl.*, 42(1):83–107, 2021.

[12] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[13] Marlis Hochbruck and Christian Lubich. On Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 34(5):1911–1925, 1997.

[14] Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Research Nat. Bur. Standards*, 45:255–282, 1950.

[15] I. Moret and P. Novati. RD-rational approximations of the matrix exponential. *BIT*, 44(3):595–615, 2004.

[16] Y. Saad. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 29(1):209–228, 1992.