Divide and conquer methods for functions of matrices with banded or hierarchical low-rank structure

Alice Cortinovis   Daniel Kressner   Stefano Massei

23 October 2021
Conference on Fast Direct Solvers

Matrix functions

Matrix functions arise in many applications, e.g.:

- Solution of PDEs (exponential, square root, fractional powers, ...)

  $\begin{cases} \frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = A\mathbf{y},\ A \in \mathbb{R}^{n\times n} \\ \mathbf{y}(0) = \mathbf{c},\ \mathbf{c} \in \mathbb{R}^n \end{cases}$ has solution $\mathbf{y}(t) = \exp(At) \cdot \mathbf{c}$.
- Electronic structure calculations (sign function)
- Network analysis (exponential for Estrada index)
- Statistical learning (logarithm, ...)
- Nonlinear matrix equations
- ...

Definition

Given function f and $A = V \cdot \mathrm{diag}(\lambda_1, \dots, \lambda_n) \cdot V^{-1}$,

$$f(A) := V \cdot \mathrm{diag}(f(\lambda_1), \dots, f(\lambda_n)) \cdot V^{-1}.$$

(Generalization to non-diagonalizable matrices possible [Higham'2008])

(Approximate) preservation of structure

Example: Tridiagonal matrix $A = \mathrm{tridiag}(-1, 2, -1)$.

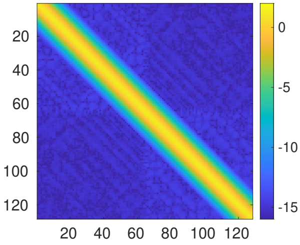


Figure: $f(z) = \exp(z)$. Log of entries of $f(A)$: matrix function $f(A)$ is *approximately* banded.

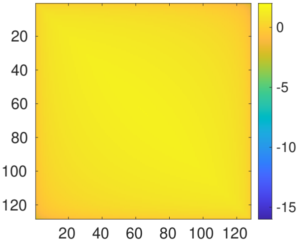


Figure: $f(z) = z^{-1}$. Log of entries of $f(A)$: matrix function $f(A)$ is not (approximately) banded...

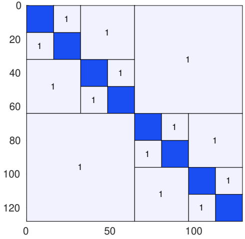


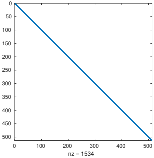
Figure: ... but $f(A) = A^{-1}$ has all off-diagonal blocks of rank 1!

Connection to polynomial/rational approximation

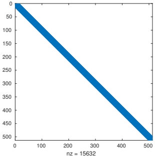
For now, assume A banded.

If f is well approximated by a small-degree polynomial p on spectrum of A then

$$f(A) \approx p(A) = \text{ banded.}$$

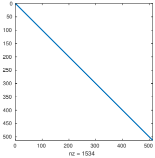


$\rightarrow$

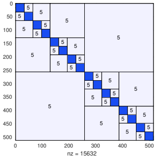


If f is well approximated by a small-degree rational function r on spectrum of A then

$$f(A) \approx r(A) = \text{ HSS matrix.}$$



$\rightarrow$



Generalizations to other formats such as hierarchically semiseparable (HSS) matrices are possible.

Existing methods for computing $f(A)$

- A priori polynomial approximation
  [Benzi/Boito/Razouk'2013], [Goedecker'1999], [Benzi/Razouk'2008], ...
- Iterations + thresholding
  [Németh/Scuseria'2000] (sign), [Bini et al.'2016] (Toeplitz matrices), ...
- A priori rational approximation
  [Gavrilyuk et al.'2002] (exponential), [Kressner/Šušnjara'2017] (spectral projectors), [Beckermann/Bisch/Luce'2021] (Markov functions of Toeplitz matrices), ...
- Iterations in HSS arithmetics + truncation strategies
  [Grasedyck et al.'2003] (sign), $\sqrt{\cdot}$, ...

Goal:

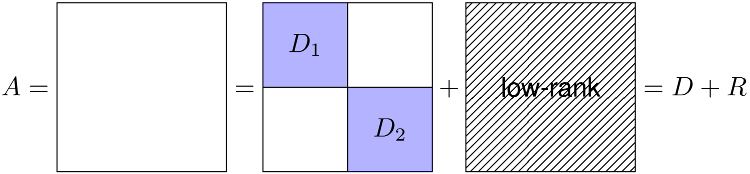
Compute $f(A)$

for matrix $A \in \mathbb{R}^{n \times n}$

with some low-rank structure.

Divide-and-conquer

We consider matrices which can be decomposed (recursively) as



e.g. banded matrices, HODLR/HSS matrices, adjacency matrices of graphs with community structure.

Divide-and-conquer idea for computing $f(A)$:

$$f(A) = \begin{bmatrix} f(D_1) & \\ & f(D_2) \end{bmatrix} + \text{ correction.}$$

Low-rank updates

Nice fact: In many cases $f(D+R) - f(D)$ is approximately low-rank!

Example: Singular value decay of $f(A) - f(D)$ for

- $A = \mathrm{tridiag}(-1, 2, -1)$ of size 256×256,
- $D_1 = D_2 = \mathrm{tridiag}(-1, 2, -1)$ of size 128×128,
- $R = A - D$ has rank 2.

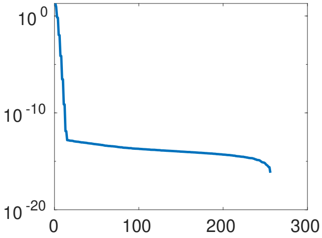


Figure: $f(z) = \exp(z)$.

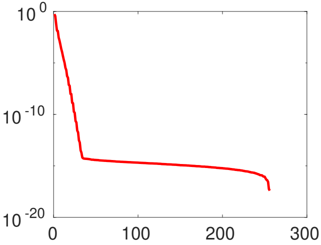


Figure: $f(z) = \sqrt{z}$.

[Beckermann/Kressner/Schweitzer'2018], [Beckermann et al.'2021]

Low-rank updates: Algorithm

Let $R = BJC^T$, with $B, C \in \mathbb{R}^{n\times r}$ and $J \in \mathbb{R}^{r\times r}$.

Choose rank m and approximate $f(D+R) - f(D) \approx U_m X_m(f) V_m^T$, where:

- $U_m \in \mathbb{R}^{n\times mr}$ orthonormal basis of $\mathcal{K}_m(D,B)$ or $q_m(D)^{-1}\mathcal{K}_m(D,B)$;
- $V_m \in \mathbb{R}^{n\times mr}$ orthonormal basis of $\mathcal{K}_m(D^T,C)$ or $q_m(D^T)^{-1}\mathcal{K}_m(D^T,C)$;

Definition

- Polynomial Krylov subspace: $\mathcal{K}_m(D,B) := \mathrm{span}\left[B, DB, D^2B, \dots, D^{m-1}B\right]$.
- Rational Krylov subspace associated with $q(z) = (z-\xi_1)\cdots(z-\xi_m)$ for prescribed poles $\xi = (\xi_1,\dots,\xi_m)^T \in \mathbb{C}^m$:

  $\mathcal{RK}_m(D,B,\xi) := \mathrm{span}\left[q_m(D)^{-1}B, q_m(D)^{-1}DB, q_m(D)^{-1}D^2B, \dots, q_m(D)^{-1}D^{m-1}B\right]$.

$X_m(f) \in \mathbb{R}^{mr\times mr}$ chosen in suitable way according to [Beckermann/Kressner/Schweitzer'2018], [Beckermann et al.'2021]

Divide-and-conquer algorithm

Input: Matrix A with hierarchical low-rank structure, function f
Output: Approximation of $f(A)$, in HSS format

if A is small **then**
    Compute $f(A)$ in “dense” arithmetics, e.g. by Schur-Parlett’s algorithm
else
    Decompose $A = D + R$ = block-diagonal + low-rank
    Compute f(diagonal blocks) recursively
    Add correction $f(D+R) - f(D)$ computed by low-rank updates algorithm
end if

[C./Kressner/Massei’2021]

Convergence of D&C algorithm

Theorem ([C./Kressner/Massei'2021])

Let A be symmetric and let f be a function analytic on an interval $\mathbb{E}$ containing the eigenvalues of A. Suppose that we use rational Krylov subspaces with poles $\xi_1, \ldots, \xi_m$, closed under complex conjugation, for computing updates. Then the output F_A of the D&C algorithm satisfies

$$\|f(A) - F_A\|_2 \leq 4 \cdot \textit{recursion depth} \cdot \min_{r \in \Pi_m / q_m} \|f - r\|_{\mathbb{E}},$$

where $q_m(z) = \prod_{i=1}^m (z - \xi_i)$.

Main ingredient of the proof:
Each low-rank update is exact for rational functions $r \in \Pi_m / q_m$.

Nice fact: Convergence is related to best rational approximation, but does not need to *explicitly find* such rational function.

Numerical experiments (1)

Example from [Ilić/Turner/Simpson'2010]: Sampling from a Gaussian random field requires $A^{-1/2}$ for $A =$ covariance matrix; here A is banded.

- D&C algorithm (extended Krylov subspaces for low-rank updates)
- Denman and Beavers iteration with HSS arithmetic (hm-toolbox [Massei/Robol/Kressner'2020])
- Matlab's `sqrtm`

A		D&C		`sqrtm` (HSS)		Dense	$A^{-\frac{1}{2}}$
Size	Band	Time	Err	Time	Err	Time	HSS rank
512	22	0.05	$2.02 \cdot 10^{-9}$	0.49	$3.44 \cdot 10^{-9}$	**0.02**	14
1,024	20	0.16	$3.45 \cdot 10^{-9}$	1.41	$5.22 \cdot 10^{-9}$	**0.13**	17
2,048	19	**0.37**	$3.76 \cdot 10^{-9}$	3.99	$6.38 \cdot 10^{-9}$	0.95	19
4,096	21	**0.8**	$3.23 \cdot 10^{-9}$	9.05	$5.61 \cdot 10^{-9}$	9.03	19
8,192	22	**2.46**	$3.46 \cdot 10^{-9}$	21.27	$6.61 \cdot 10^{-9}$	70.42	20
16,384	25	**5.7**		48.92			22
32,768	26	**15.12**		102.65			25
65,536	26	**26.25**		209.56			24
131,070	25	**60.44**		417.21			24
262,140	26	**146.97**		918.81			26

Numerical experiments (2)

For A = adjacency matrix of undirected graph, diagonal entries of $\exp(A)$ are *subgraph centralities* and $\frac{1}{n}\mathrm{trace}(\exp(A))$ is the Estrada index.

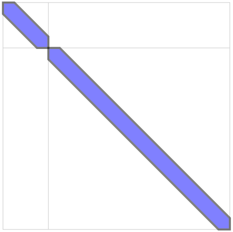
- D&C algorithm (splitting vertices into 2 components using METIS algorithm, polynomial Krylov subspaces for the low-rank updates)
- `mmq` (Gauss quadrature to approximate each diagonal entry of $\exp(A)$, see [Golub/Meurant'2010])
- Matlab's `expm` and `eig`

A	D&C diagonal		mmq diagonal		`expm`	D&C trace		`eig`
Size	Time	Err	Time	Err	Time	Time	Err	Time
2,642	1.01	$6.24 \cdot 10^{-10}$	**0.8**	$1.82 \cdot 10^{-11}$	1.98	**0.14**	$7.71 \cdot 10^{-13}$	0.44
4,941	**2.06**	$1.29 \cdot 10^{-8}$	5.15	$3.39 \cdot 10^{-11}$	16.11	**0.47**	$7.75 \cdot 10^{-11}$	3.61
7,716	**8.01**	$4.03 \cdot 10^{-9}$	24.19	$2.29 \cdot 10^{-10}$	56.59	**3.91**	$1.96 \cdot 10^{-12}$	8.73
10,774	**15.87**	$1.04 \cdot 10^{-8}$	39.42	$3.54 \cdot 10^{-10}$	151.52	**2.98**	$2.69 \cdot 10^{-10}$	21.04
20,055	**38.49**	$2.59 \cdot 10^{-9}$	97.53	$1.4 \cdot 10^{-11}$	929.25	**6.99**	$2.66 \cdot 10^{-13}$	124.34
45,087	**182.19**		603.57			**27.99**		

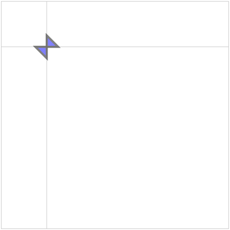
Test matrices from the SuiteSparse Matrix Collection.

Special case: Banded matrices (1)

$$A = D + R =$$



$$+$$

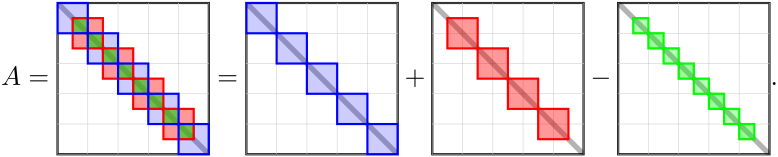


Bases of polynomial Krylov subspaces inherit sparsity and have the form

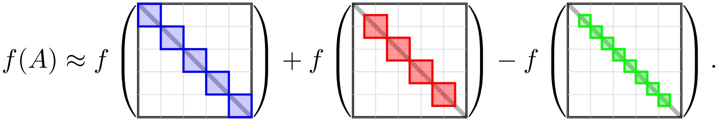
$$U_m = V_m = \begin{bmatrix} 0 \\ I \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Special case: Banded matrices (2)

D&C algorithm simplifies a lot! $\rightsquigarrow$ "block-diagonal splitting algorithm"



Then,



Theorem ([C./Kressner/Massei'2021])

For a matrix A with bandwidth b, block size s, let $m := \lfloor s/2b \rfloor$, then

$$\|f(A) - \mathrm{approx}\|_2 \leq 10 \min_{p \in \Pi_m} \|f - p\|_{\textit{numerical range of } A}.$$

Numerical example for banded matrices

The size of the blocks can be chosen adaptively.

Example: A is tridiagonal, $\texttt{linspace}(2,3,n)$ on the diagonal, -1 on super- and sub-diagonals; $f(A) = \sqrt{A}$.

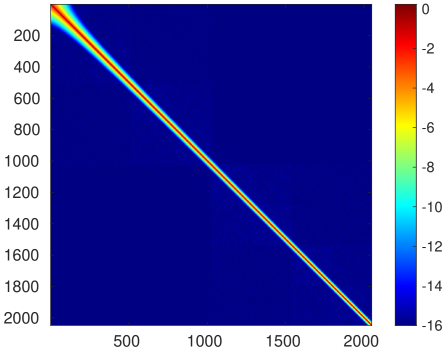


Figure: $\log|f(A)|$

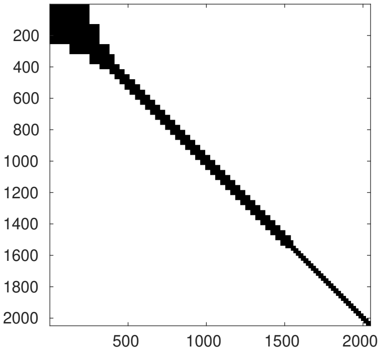


Figure: Sparsity pattern of the output

Computational complexity

Simplified assumption: low-rank updates converge in a fixed number of steps.

- General D&C algorithm: $O(k^2 n \log n)$ for matrix of size n and HSS rank k
- Block diagonal splitting algorithm: $O(nb^2)$ for a matrix of bandwidth b

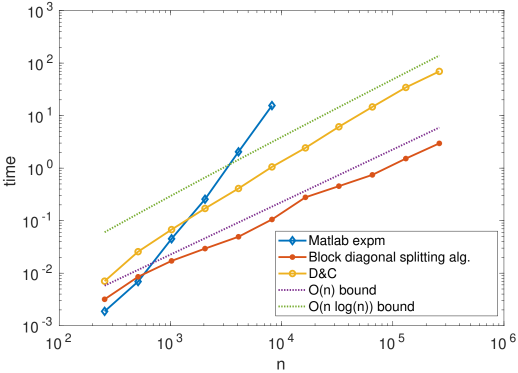


Figure: $A = \mathrm{tridiag}(-1, 2, -1)$, $f = \exp$.

Trace & diagonal of matrix functions

Assume polynomial Krylov subspaces are used for the low-rank updates.

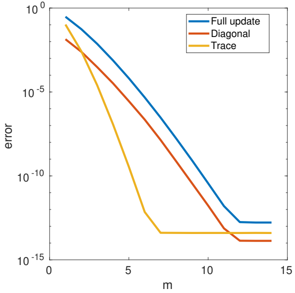


Figure: Convergence of low-rank updates algorithm $f(A) - f(A - R)$, A and R symmetric, $f = \exp$.

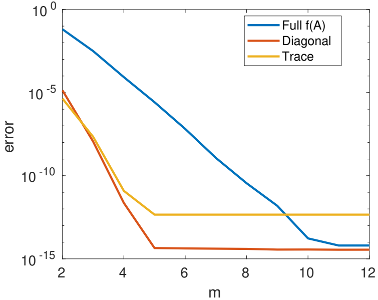


Figure: Convergence of block-diagonal splitting algorithm for the exponential of normalized non-symmetric pentadiagonal matrix A.

Trace and/or diagonal converge faster! (And we proved it)

Summary & conclusions

1. Two new algorithms for approximating matrix functions:
   - A general D&C algorithm for matrices with hierarchical low-rank structure;
   - An algorithm that is specialized to banded matrices.
2. Convergence analysis: Links to best polynomial/rational approximation of f on a suitable region containing eigenvalues of A.
3. Numerical tests: Generally faster than existing methods (for large n), with comparable accuracy.

A. Cortinovis, D. Kressner, S. Massei. Divide and conquer methods for functions of matrices with banded or hierarchical low-rank structure.
https://arxiv.org/abs/2107.04337