On maximum volume submatrices and cross approximation

Alice Cortinovis Daniel Kressner Stefano Massei

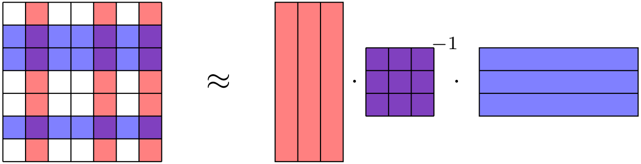
ANCHP
Institute of Mathematics
École Polytechnique Fédérale de Lausanne

Low-Rank Optimization and Applications
Leipzig, 04 April 2019

EPFL

Low-rank approximation from rows and columns of A

Idea [Goreinov/Tyrtyshnikov/Zamarashkin'1997, Bebendorf'2000, many others...]:
Use selected rows and columns of A to build "cross approximation"



- Boundary element method [Bebendorf'2000]
- General tool for assembling $\mathcal{H}$-matrices, $\mathcal{H}^2$-matrices, ... [Hackbusch'2015]
- Uncertainty quantification [Harbrecht/Peters/ Schneider'2012]
- Kernel-based learning [Bach/Jordan'2005] and spectral clustering [Fowlkes/Belongie/Chung/Malik'2004] (Nyström method)
- Extension to low-rank tensor approximation [Oseledets/Tyrtyshnikov'2010, Ballani/Grasedyck/Kluge'2013, Savostyanov'2014, ...]

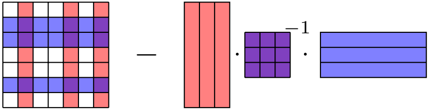
Overview

1. Maxvol submatrices
   - Connection to low-rank approximation problem
   - NP hardness
2. Maxvol submatrices for structured matrices
   - symmetric positive semidefinite matrices;
   - diagonally dominant matrices.
3. Error bounds for cross approximation
   - general matrices;
   - symmetric positive semidefinite matrices;
   - (doubly) diagonally dominant matrices.
4. Cross approximation for functions

Maxvol submatrices

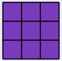
Quasi-optimal choice of rows and columns

How close is error of cross approximation

error = 

to best rank-m approximation error $\min_{B,C \in \mathbb{R}^{n\times m}} \|A - BC^T\|_2 = \sigma_{m+1}(A)$?

Theorem ([Goreinov/Tyrtyshnikov'2001])

If  *is $m \times m$ submatrix of maximum volume (maximum absolute value of the determinant), then*

$$\|\text{error}\|_{\max} \leq (m+1)\sigma_{m+1}(A),$$

where $\|B\|_{\max} = \max_{i,j=1,\dots,n} |b_{ij}|$.

Proof idea by [Goreinov/Tyrtyshnikov'2001]. Assume w.l.o.g. $n = m + 1$. Then

$$\|A^{-1}\|_2 = \frac{1}{\sigma_{m+1}(A)}.$$

On the other hand,

$$A^{-1} = \frac{1}{\det A} \mathsf{adj}(A)$$

implies that maximum element of A^{-1} is at entry $(m+1, m+1)$. In turn,

$$\|\mathrm{error}\|_{\max} = \left|\left(A^{-1}\right)_{m+1,m+1}\right| = \|A^{-1}\|_{\max}.$$

Hence,

$$\sigma_{m+1}(A)^{-1} = \|A^{-1}\|_2 \le (m+1)\|A^{-1}\|_{\max} = (m+1)|(A^{-1})_{m+1,m+1}|.$$

Remark.

- Maxvol approximation is actually optimal in some weird sense. Consider mixed norm

$$\|B\|_{\infty\to 1} := \sup \|Bx\|_1/\|x\|_\infty, \quad \|B\|_{1\to\infty} := \sup \|Bx\|_\infty/\|x\|_1 = \|B\|_{\max}.$$

  and approximation numbers

$$\beta_{k+1}(A) := \min\{\|E\|_{\infty\to 1} : \mathsf{rank}(A+E) \le k\}, \qquad k = 0,\dots,n-1.$$

  Then $\beta_n(A) = \|A^{-1}\|_{\max}^{-1}$ see [Higham'2002]. An extension of the proof shows

$$\|\mathrm{error}\|_{\max} = \beta_{m+1}(A).$$

  By norm equivalence,

$$\|\mathrm{error}\|_{\max} \le (m+1)^2 \min\{\|E\|_{\max} : \mathsf{rank}(A+E) \le m\}.$$

  and the constant is tight. Recovers result by [Goreinov/Tyrtyshnikov'2011].

(Approximate) maxvol is NP hard

- Papadimitriou'1984 reduced NP-complete SAT via a subgraph selection problem to maxvol for 0/1 matrices $\rightsquigarrow$ maxvol is NP hard.
- Di Summa et al.'2014 showed that it is NP-hard to approximate maxvol within factor that does not grow at least exponentially with m.

Remark. Due to nature of determinants, second result has little implication for low-rank approximation. Consider

$$A = \begin{bmatrix} I_m & 0 \\ 0 & B_m \end{bmatrix}, \quad B_m = \text{tridiag}[\frac{1}{2}, 1, -\frac{1}{2}]$$

Note that

$$|\det(I_m)| = 1, \qquad |\det(B_m)| \sim \left(\frac{1+\sqrt{2}}{2}\right)^m.$$

Nevertheless, the first m rows/columns constitute an excellent choice for cross approximation:

$$\|\text{error}\|_{\max} = \|B_m\|_{\max} = 1 = \sigma_{m+1}(B).$$

Is it NP hard to find cross approximation with polynomial constant?

Maxvol submatrices for structured matrices

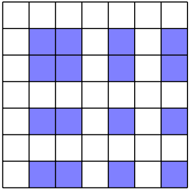
Maxvol submatrix for SPSD

Let A be symmetric positive semidefinite (SPSD).

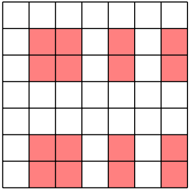
Obvious: Element of maximum absolute value of A on the diagonal.

Less obvious: Submatrix of maximum volume of A can always be chosen to be principal.

Principal submatrix



Not principal

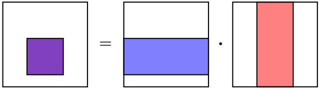


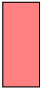
Maxvol submatrix for SPSD

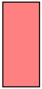
Theorem ([Cortinovis/K./Massei'2019])

If A is symmetric positive semidefinite then the maximum volume $m \times m$ submatrix is attained by a principal submatrix.

Volume of a (rectangular) matrix $=$ product of its singular values.

1. Cholesky decomposition $A = C^T \cdot C$: 
2. Inequality for product of singular values of a product [Horn/Johnson'1991]:

   Volume of  $\leq$ Volume of  $\cdot$ Volume of 

3. Relate volume of  and  to volume of principal submatrices and conclude.

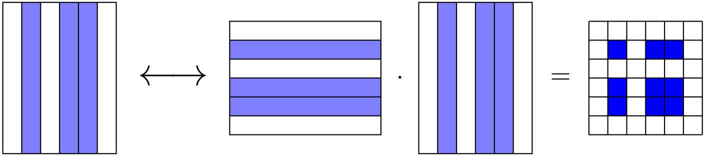
□

Consequences

1. One-to-one correspondence with column selection problem:

   Given $n \times k$ matrix B, find $n \times m$ submatrix of maximum volume.

   Also arises in low-rank approximation problems [Deshpande/Rademacher/Vempala/Wang'2006] and rank-revealing LU factorizations [Pan'2000].



2. Column selection problem is NP-hard [Çivril/Magdon-Ismail'2009] $\rightsquigarrow$ maximum volume submatrix problem is NP-hard even when restricted to symmetric positive semidefinite matrices.

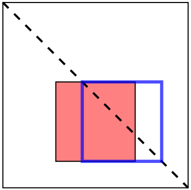
Maxvol submatrix for diagonally dominant

Definition

$A \in \mathbb{R}^{n \times n}$ is (row) diagonally dominant if $\sum_{j=1,\dots,n;\, j\neq i} |a_{ij}| \leq |a_{ii}|$ for all $i = 1, \dots, n$.

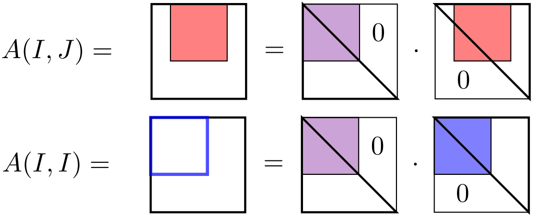
Theorem ([Cortinovis/K./Massei'2019])

If A is diagonally dominant then the maxvol $m \times m$ submatrix is attained by a principal submatrix.



Proof idea.

1. Reduction to upper triangular case via LU factorization.



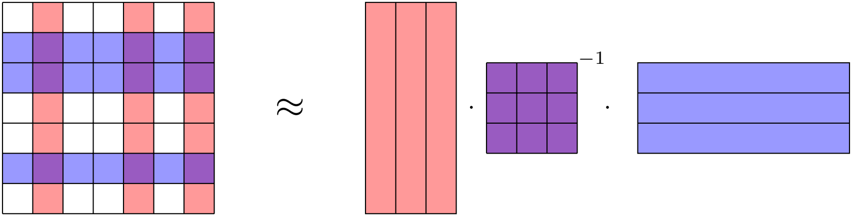
2. Factor U inherits diagonal dominance from A! For unit upper triangular diagonally dominant matrix: Each submatrix has $|\det| \leq 1$. $\square$

Known special case: for $m = n - 1$, the result of the theorem is covered in the proof of Theorem 2.5.12 in [Horn/Johnson'1991].

Cross approximation

Cross approximation algorithm with full pivoting

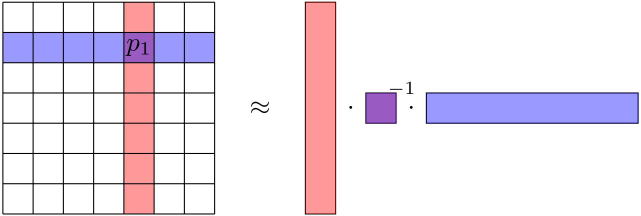
Aim: Finding quasi-optimal row/column indices.



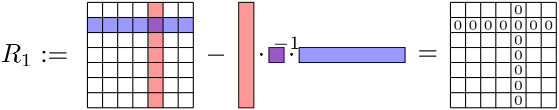
Cross approximation algorithm with full pivoting

ACA = Adaptive Cross Approximation [Bebendorf'2000, Carvajal/Chapman/Geddes'2005, ...]
Greedy algorithm for volume maximization.

First step: Select element of maximum absolute value, p_1 (first pivot).

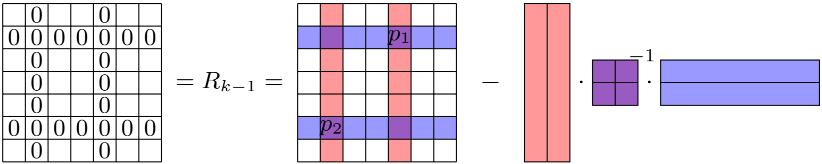


Denote



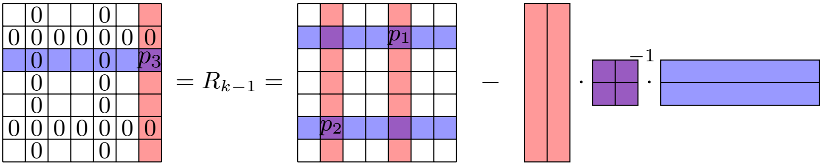
Cross approximation algorithm with full pivoting

Situation after $k-1$ steps:



Cross approximation algorithm with full pivoting

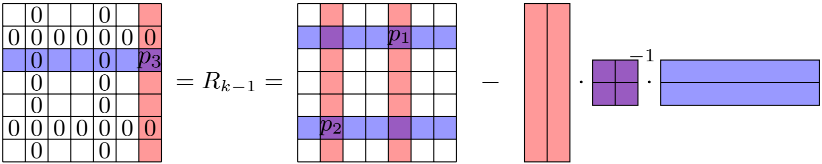
Situation after $k-1$ steps:



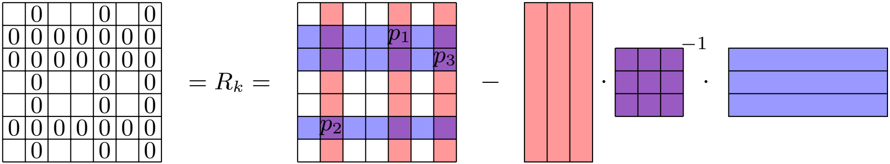
At k-th step: Choose element of maximum absolute value in remainder R_{k-1} (k-th pivot p_k).

Cross approximation algorithm with full pivoting

Situation after $k-1$ steps:



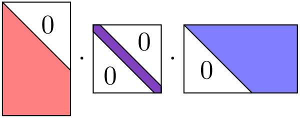
At k-th step: Choose element of maximum absolute value in remainder R_{k-1} (k-th pivot p_k).



Equivalence to Gaussian elimination and LDU factorization

- Cross approximation = greedy algorithm for volume maximization.
- Equivalent to Gaussian elimination with complete pivoting [Bebendorf'2000]. In particular, no need to compute inverses at each step!
- (Up to permutations) we obtain an incomplete LDU factorization

$$A - R_m = L_m D_m U_m =$$



where

- $D_m = \text{diag}(p_1, \ldots, p_m)$ contains pivot elements;
- L_m and U_m have ones on the diagonal.

Error of cross approximation

Analysis of error $\|R_m\|_{\max}$ obtained after m steps of cross approximation: By performing one additional step

$$\|R_m\|_{\max} = \|\text{Schur complement of } A_{11} \text{ in } A\|_{\max} = |p_{m+1}|.$$

Ideally $|p_{m+1}|$ is close to $\sigma_{m+1}(A)$.

Wlog, restrict to $(m+1) \times (m+1)$ matrices and consider factorization $A = LDU$:

$$\frac{1}{\sigma_{m+1}(A)} = \|A^{-1}\|_2 \leq \|U^{-1}\|_2 \|D^{-1}\|_2 \|L^{-1}\|_2.$$

What can go wrong?

1. Intermediate pivots can be $\ll |p_{m+1}| \rightsquigarrow \|D^{-1}\|_2 \not\approx \frac{1}{|p_{m+1}|}$.
2. $\|L^{-1}\|_2$ and $\|U^{-1}\|_2$ can be large.

Closely related to but **not** covered by numerical linear algebra literature on error analysis of (rank-revealing) LU decompositions.

Error bounds for the general case

1. Consider growth factor $= \rho_k := \sup\left\{\frac{\|R_k\|_{\max}}{\|A\|_{\max}} \mid \text{rank}(A) \geq k\right\}$, playing prominent role in error analysis of LU factorization.
   [Wilkinson'1961] proved for complete pivoting

$$\rho_k \leq 2\sqrt{k+1}(k+1)^{\ln(k+1)/4}.$$

   Note that bound is not tight; usually $\rho_k = O(1)$. Obtain

$$\|D^{-1}\|_2 \leq \frac{\rho_m}{|p_{m+1}|}.$$

2. L and U have ones on the diagonal and all other entries have absolute value ≤ 1 because of full pivoting $\rightsquigarrow \|L^{-1}\|_2 \leq 2^m$ and $\|U^{-1}\|_2 \leq 2^m$; see [Higham'1987].

Theorem ([Cortinovis/K./Massei'2019])

After m steps, error of cross approximation satisfies

$$\|R_m\|_{\max} \leq 4^m \rho_m \cdot \sigma_{m+1}(A).$$

Symmetric positive semidefinite case

Benefits for SPSD matrices:

1. Huge: Diagonal pivoting is sufficient (SPSD preserved by Schur compl)
2. Minor: Pivots do not grow $\rightsquigarrow$ ρ_m replaced by 1 in the theorem.

Corollary

If A is symmetric positive semidefinite then

$$\|R_m\|_{\max} \le 4^m \cdot \sigma_{m+1}(A).$$

This matches a result of [Harbrecht/Peters/Schneider'2012].

Bound is tight for SPSD. Kahan'1966:

$$U = L^T = \begin{bmatrix} 1 & -1 & \cdots & -1 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & 1 \end{bmatrix} \in \mathbb{R}^{(m+1)\times(m+1)}$$

$$\|A^{-1}\|_2 = \|U^{-1}\|_2^2 \sim 4^m \quad \rightsquigarrow \quad \sigma_{m+1}(A) \sim 4^{-m}$$

On the other hand, $|p_{m+1}| = 1$.

Diagonally dominant case

Benefits for diagonally dominant matrices:

1. Diagonal pivoting is sufficient (diagonal dominance preserved by Schur compl)
2. Small growth factor: $\|R_k\|_{\max} \leq 2\|A\|_{\max}$ for every k. [Wilkinson'1961]
3. In the LDU factorization, U is diagonally dominant. Hence, $\|U^{-1}\|_2 \leq m$. [Peña'2004]

Corollary

If A is diagonally dominant then $\|R_m\|_{\max} \leq (m+1) \cdot 2^{m+1} \cdot \sigma_{m+1}(A)$.

4. If A is doubly diagonally dominant (that is, A and A^T are diag. dom.) then also L^T is diagonally dominant.

Corollary

If A is doubly diagonally dominant then $\|R_m\|_{\max} \leq 2(m+1)^2 \cdot \sigma_{m+1}(A)$.

Result relevant in spectral clustering based on cross approximation.

Diagonally dominant case: Tightness of bounds

- Diagonally dominant case: example with

$$\frac{\|R_m\|_{\max}}{\sigma_{m+1}(A)} = \Theta(m^2).$$

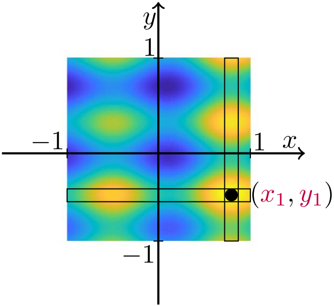
  Related to studies on stability of LDU factorizations [e.g. Demmel/ Koev'2004, Dopico/Koev'2011, Barreras/Peña'2012/13].

- Doubly diagonally dominant case: example with

$$\frac{\|R_m\|_{\max}}{\sigma_{m+1}(A)} = \Theta(m).$$

Extension to functions

Consider bivariate function $f : [-1,1] \times [-1,1] \to \mathbb{R}$ and choose point (x_1, y_1) of maximum absolute value.



"Rank-1 approximation" of f is separable function

$$f_1(x,y) = f(x, y_1) \cdot \frac{1}{f(x_1, y_1)} \cdot f(x_1, y).$$

Next steps $\rightsquigarrow$ analogous to matrix algorithm.
[Bebendorf'2000, Carvajal/Chapman/Geddes'2005, Townsend/Trefethen'2015]

Error bounds for functional approximation

Definition

Bernstein ellipse $\mathcal{E}_r$ = ellipse with foci ± 1 and sum of semiaxes r.

Theorem ([Cortinovis/K./Massei'2019])

Assume that $f(\cdot, y)$ admits an analytic extension $\tilde{f}$ to the Bernstein ellipse $\mathcal{E}_{r_0}$ for each $y \in [-1, 1]$. Choose $1 < r < r_0$. Denote

$$M := \sup_{\eta \in \partial \mathcal{E}_r,\, \xi \in [-1,1]} |\tilde{f}(\eta, \xi)|.$$

Then the error after m steps satisfies

$$\|\mathrm{error}_m\|_{\max} \leq \frac{2M\rho_m}{1 - 1/r} \cdot \left(\frac{r}{4}\right)^{-m}.$$

Idea of proof: error bound for cross approximation for general matrices applied to matrix interpolating the function in suitable points + standard polynomial approximation arguments.

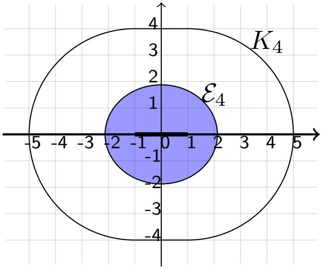
Comparison to existing convergence results

ρ_m has subexponential growth $\rightsquigarrow$ Algorithm converges linearly with rate $\frac{4}{r}$ for $r > 4$.

Previous convergence results for complete pivoting [Townsend/Trefethen'2015]: they need the functions $f(\cdot, y)$ to have an analytic extension in

$$K_r := \{\text{points at distance} \le r \text{ from the segment } [-1, 1]\}$$

for linear convergence with rate $\frac{4}{r}$.



Conclusions

New results:

1. Maxvol submatrix of symmetric positive semidefinite or diagonally dominant matrices attained by principal submatrix.
2. Error analysis of cross approximation (both for matrix and function case).

Major open problem:

3. Cross approximation with complete pivoting always works well in practice. Find appropriate framework that captures this!

Next step:

4. Tensors!

Techreport available from `anchp.epfl.ch`:

- A. Cortinovis, DK, and S. Massei. On maximum volume submatrices and cross approximation for symmetric semidefinite and diagonally dominant matrices. February 2019.