JOHANNES KEPLER UNIVERSITY LINZ

Institute of Computational Mathematics

On the Best Uniform Approximation by Low-Rank Matrices

Irina Georgieva
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev St., Bl. 8, 1113 Sofia, Bulgaria

Clemens Hofreither
Institute of Computational Mathematics, Johannes Kepler University
Altenberger Str. 69, 4040 Linz, Austria

NuMa-Report No. 2016-10 December 2016

A–4040 LINZ, Altenbergerstraße 69, Austria

Technical Reports before 1998:

1995

95-1	Hedwig Brandstetter *Was ist neu in Fortran 90?*	March 1995
95-2	G. Haase, B. Heise, M. Kuhn, U. Langer *Adaptive Domain Decomposition Methods for Finite and Boundary Element Equations.*	August 1995
95-3	Joachim Schöberl *An Automatic Mesh Generator Using Geometric Rules for Two and Three Space Dimensions.*	August 1995

1996

96-1	Ferdinand Kickinger *Automatic Mesh Generation for 3D Objects.*	February 1996
96-2	Mario Goppold, Gundolf Haase, Bodo Heise und Michael Kuhn *Preprocessing in BE/FE Domain Decomposition Methods.*	February 1996
96-3	Bodo Heise *A Mixed Variational Formulation for 3D Magnetostatics and its Finite Element Discretisation.*	February 1996
96-4	Bodo Heise und Michael Jung *Robust Parallel Newton-Multilevel Methods.*	February 1996
96-5	Ferdinand Kickinger *Algebraic Multigrid for Discrete Elliptic Second Order Problems.*	February 1996
96-6	Bodo Heise *A Mixed Variational Formulation for 3D Magnetostatics and its Finite Element Discretisation.*	May 1996
96-7	Michael Kuhn *Benchmarking for Boundary Element Methods.*	June 1996

1997

97-1	Bodo Heise, Michael Kuhn and Ulrich Langer *A Mixed Variational Formulation for 3D Magnetostatics in the Space* $H(rot) \cap H(div)$	February 1997
97-2	Joachim Schöberl *Robust Multigrid Preconditioning for Parameter Dependent Problems I: The Stokes-type Case.*	June 1997
97-3	Ferdinand Kickinger, Sergei V. Nepomnyaschikh, Ralf Pfau, Joachim Schöberl *Numerical Estimates of Inequalities in* $H^{\frac{1}{2}}$.	August 1997
97-4	Joachim Schöberl *Programmbeschreibung NAOMI 2D und Algebraic Multigrid.*	September 1997

From 1998 to 2008 technical reports were published by SFB013. Please see
`http://www.sfb013.uni-linz.ac.at/index.php?id=reports`
From 2004 on reports were also published by RICAM. Please see
`http://www.ricam.oeaw.ac.at/publications/list/`
For a complete list of NuMa reports see
`http://www.numa.uni-linz.ac.at/Publications/List/`

On the Best Uniform Approximation by Low-Rank Matrices

Irina Georgieva* Clemens Hofreither†

December 7, 2016

Abstract

We study the problem of best approximation, in the elementwise maximum norm, of a given matrix by another matrix of lower rank. We generalize a recent result by Pinkus that describes the best approximation error in a class of low-rank approximation problems and give an elementary proof for it. Based on this result, we describe the best approximation error and the error matrix in the case of approximation by a matrix of rank one less than the original one. For the case of approximation by matrices with arbitrary rank, we give lower and upper bounds for the best approximation error in terms of certain submatrices of maximal volume. We illustrate our results using 2×2 matrices as examples, for which we also give a simple closed form of the best approximation error.

1 Introduction

We consider the problem of approximating a given matrix as closely as possible by a matrix of the same size, but lower rank. When measuring the approximation error in the spectral or Frobenius norms, a full description of the best approximation and its error is given in terms of the singular value decomposition [10, 3]. In different matrix norms, very little was known about this approximation problem until a recent article by Pinkus [8], where approximation by a class of elementwise norms, and there in particular ℓ_1-like norms, was studied. Pinkus derives expressions for the best approximation error in such norms and, in particular cases, shows that a best approximating matrix matches the original matrix in a number of rows and columns.

In the present paper, we derive analogues of several of Pinkus' results for the case of approximation in the elementwise maximum norm. In the process, we generalize one core result from [8] and prove it using only known basic results on best approximation, whereas the proof of the original result relied heavily on the theory of n-widths. Building on this result, we obtain an expression for the best approximation error of a matrix by another one with rank one less, as well as a characterization of the matrix of best approximation.

For approximation where the difference in ranks is greater than one, we have no closed formula for the best approximation error, but give lower and upper bounds for it involving certain submatrices of maximal volumes, that is, with greatest modulus of their determinants. These results are similar to some given by Babaev [1] in the continuous setting. The relevance of

*Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G. Bonchev St., Bl. 8, Sofia, Bulgaria. irina@math.bas.bg

†Institute of Computational Mathematics, Johannes Kepler University, Altenberger Str. 69, 4040 Linz, Austria. chofreither@numa.uni-linz.ac.at

submatrices of maximal volume to the problem of low-rank approximation was first established by Goreinov and Tyrtyshnikov [4].

The remainder of the paper is structured as follows. In Section 2, we state the low-rank approximation problem and prove a result on the best approximation error which generalizes a result by Pinkus. In Section 3, we focus on the case of approximating a matrix by another matrix of rank one less, where the best approximation error can be described quite closely and we obtain an equioscillation result for the error matrix. In Section 4, we deal with approximations of arbitrary rank and give lower and upper bounds for the best approximation error in terms of certain submatrices of maximal volume. Finally, in Section 5, we illustrate some of our results in the simple case of 2×2-matrices, where we are also able to give a simple closed form for the best approximation error.

2 Approximation with low-rank matrices

2.1 Problem statement

Let $A \in \mathbb{R}^{m \times n}$ and $p, q \in [1, \infty]$. We define the entrywise matrix norm

$$|A|_{p,q} := \left(\sum_{i=1}^{m} \left(\sum_{j=1}^{n} |a_{ij}|^q \right)^{p/q} \right)^{1/p},$$

where as usual $p = \infty$ or $q = \infty$ means the maximum norm in the corresponding direction. The two most common special cases are the entrywise maximum (or Chebyshev) norm and the Frobenius norm,

$$|A|_{\max} := |A|_{\infty,\infty} = \max_{i,j} |a_{ij}|, \qquad |A|_F := |A|_{2,2} = \left(\sum_{i,j} a_{ij}^2 \right)^{1/2}.$$

For vectors, we denote by $\| \cdot \|_p$ the usual ℓ_p-vector norm.

Definition 1. For $p, q \in [1, \infty]$ and $k \in \mathbb{N}_0$, we define the best approximation error

$$E_{p,q}^k(A) := \inf_{\operatorname{rank} G \leq k} |A - G|_{p,q},$$

where G runs over all $m \times n$ matrices of rank at most k.

The only completely solved instance of the above best approximation problem is in the Frobenius norm $| \cdot |_F = | \cdot |_{2,2}$ (and the spectral norm, which however does not fall into our class of elementwise norms) [10, 3]. In this case, given the singular value decomposition

$$A = U \Sigma V^\top, \qquad U \in \mathbb{R}^{m \times K}, \quad \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_K), \quad V \in \mathbb{R}^{n \times K},$$

where $K = \operatorname{rank} A$, both U and V have mutually orthonormal columns, and $\sigma_1 \geq \ldots \geq \sigma_K > 0$ are the singular values of A, the best approximation of rank $k \leq K$ is given by the truncated singular value decomposition

$$A_k = U \operatorname{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) V^\top$$

and the best approximation error is given by

$$|A - A_k|_F^2 = \sigma_{k+1}^2 + \ldots + \sigma_K^2.$$

2.2 Characterization of the best approximation error

The following theorem yields an expression for the best low-rank approximation error in arbitrary elementwise norms. Here and in what follows, for $q \in [1, \infty]$, we denote its Hölder conjugate by q' such that $1/q + 1/q' = 1$.

Theorem 1. *Let $A \in \mathbb{R}^{m\times n}$ with rows $(a_i)_{i=1}^m$ be of rank n and $p, q \in [1, \infty]$. Then the best approximation error by a matrix of rank $k \in \{0, \dots, n\}$ in the $|\cdot|_{p,q}$-norm is given by*

$$E_{p,q}^k(A) = \inf_{U_{n-k}} \left\| \left(\max_{\substack{h\in U_{n-k}\\ \|h\|_{q'}=1}} |h \cdot a_i| \right)_{i=1,\dots,m} \right\|_p ,$$

where the infimum runs over all subspaces $U_{n-k} \subset \mathbb{R}^n$ of dimension $n-k$.

For $k = n-1$, we have

$$E_{p,q}^{n-1}(A) = \min_{h\neq 0} \frac{\|Ah\|_p}{\|h\|_{q'}}. \tag{1}$$

Remark 1. The statement for the case $k = n-1$ of the above theorem is already given by Pinkus in [8, Corollary 2.3]. However, whereas Pinkus arrived at this result via the theory of n-widths, we give below a more direct proof which relies only on a fundamental result from approximation theory and allows us to cover also the case $k < n-1$. However, our approach does not yield the result on n-widths which is also a part of [8, Corollary 2.3].

For the proof of Theorem 1, we make use of the following classical characterization of best approximation by duality.

Theorem 2 ([12, 2])**.** *Let $(X, \|\cdot\|)$ a normed linear space and $U \subset X$ a closed subspace. Then $u \in U$ is a best approximant in U to $x \in X \setminus U$ if and only if there exists a*

$$h \in U^\perp := \{h \in X^* : \langle h,\, u\rangle = 0 \ \ \forall u \in U\}$$

with the properties

$$\|h\|_* = 1,$$
$$\langle h,\, x-u\rangle = \|x-u\|.$$

Furthermore, the best approximation error is given by

$$\inf_{u\in U} \|x-u\| = \sup_{\substack{h\in U^\perp\\ \|h\|_*=1}} \langle h,\, x\rangle.$$

Here, X^* denotes the continuous dual, $\|\cdot\|_*$ the dual norm and $\langle\cdot,\,\cdot\rangle$ the duality product.

Applying this result to the space $(\mathbb{R}^n, \|\cdot\|_p)$, $p \in [1, \infty]$, which has dual space $(\mathbb{R}^n, \|\cdot\|_{p'})$, $1/p + 1/p' = 1$, we immediately obtain the following statements.

Corollary 1. *For $x \in \mathbb{R}^n$ and $U_k \subset \mathbb{R}^n$ a k-dimensional subspace, let $U_k^\perp \subset \mathbb{R}^n$ denote the orthogonal complement to U_k. Then*

$$\inf_{u\in U_k} \|x-u\|_p = \max_{\substack{h\in U_k^\perp\\ \|h\|_{p'}=1}} |h\cdot x|. \tag{2}$$

For $x \in \mathbb{R}^n$ and $U_{n-1} \subset \mathbb{R}^n$ a $(n-1)$-dimensional subspace, let $h \in \mathbb{R}^n$ orthogonal to U_{n-1} with $\|h\|_{p'} = 1$. Then

$$\inf_{u\in U_{n-1}} \|x-u\|_p = |h\cdot x|. \tag{3}$$

By treating the low-rank matrix approximation problem as a simultaneous approximation problem for the rows of the matrix, we obtain the following proof.

Proof of Theorem 1. The problem of rank k approximation is equivalent to finding a k-dimensional subspace $U_k \subset \mathbb{R}^n$ and approximating each row in U_k with error

$$e_i := \inf_{u \in U_k} \|a_i - u\|_q, \qquad i = 1, \dots, m. \tag{4}$$

The total error is then given by

$$\|(e_1, \dots, e_m)\|_p.$$

Thus the first statement follows by using the identity (2) to represent the errors (4).

For $k = n-1$, we can identify each subspace U_{n-1} with a vector $h \in \mathbb{R}^n$, $\|h\|_{q'} = 1$, which is orthogonal to U_{n-1}. Thus, from (3) we obtain

$$E^{n-1}_{p,q}(A) = \min_{\|h\|_{q'}=1} \|(|h \cdot a_i|)_{i=1,\dots,m}\|_p .$$

Since the vector $(h \cdot a_i)_{i=1,\dots,m}$ is nothing but Ah, the second statement follows. $\square$

3 Approximation with rank reduced by one

Due to the simple form of the best approximation error (1) given in the case $k = n-1$ in Theorem 1, much more can be said for approximation of a matrix of rank n by one of rank at most $n-1$.

3.1 Description of the best approximation error

We introduce, for any square matrix B of size n, a quantity $\alpha_{p,q}(B)$ which we will show to be a lower bound for the best approximation error in the $|\cdot|_{p,q}$ norm by a matrix of rank at most $n-1$. As we will see, in certain cases this quantity coincides with the best approximation error.

Definition 2. For any square matrix $B \in \mathbb{R}^{n\times n}$ and any $p, q \in [1, \infty]$, we let

$$\alpha_{p,q}(B) := \begin{cases} \dfrac{1}{|B^{-1}|_{p,q}}, & \det B \neq 0, \\ 0, & \det B = 0. \end{cases}$$

In the following, we let $\langle n \rangle := \{1, \dots, n\}$. For a matrix $A \in \mathbb{R}^{m\times n}$ and vectors of indices $I \subset \langle m \rangle, J \subset \langle n \rangle$, we write $A_{I,J} \in \mathbb{R}^{|I|\times|J|}$ for the submatrix of A formed by taking the rows from the index vector I and the columns from the index vector J. We denote the set of all square submatrices of A with size k by

$$\mathcal{S}_k(A) := \{A_{I,J} : I \subset \langle m \rangle, J \subset \langle n \rangle, |I| = |J| = k\}.$$

The rank of a nonzero matrix can be characterized as

$$\operatorname{rank}(A) = \max\{k \in \mathbb{N} : \exists B_k \in \mathcal{S}_k(A) : \det B_k \neq 0\}.$$

In addition to the entrywise matrix norms $|\cdot|_{p,q}$, we will also make use of the operator norms

$$\|A\|_{p,q} := \max_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_q} = \max_{\|x\|_q=1} \|Ax\|_p.$$

The following simple lemma gives a relation between these two classes of norms. It is not new, but we prove it here for the sake of completeness.

Lemma 1. *For any matrix $B \in \mathbb{R}^{m\times n}$ and $p, q \in [1, \infty]$, we have*

$$\|B\|_{q,p} = \max_{\|x\|_p=1} \|Bx\|_q \le |B|_{q,p'}.$$

Proof. Assume $\|x\|_p = 1$. Using Hölder's inequality, we estimate

$$\|Bx\|_q^q = \sum_{i=1}^m \left| \sum_{j=1}^n B_{ij} x_j \right|^q \le \sum_{i=1}^m \left(\sum_{j=1}^n |B_{ij} x_j| \right)^q \le \sum_{i=1}^m (\|B_{i*}\|_{p'} \|x\|_p)^q = \sum_{i=1}^m \|B_{i*}\|_{p'}^q,$$

where by B_{i*} we mean the i-th row of B. It follows

$$\|Bx\|_q \le \left(\sum_{i=1}^m \|B_{i*}\|_{p'}^q \right)^{1/q} = \left(\sum_{i=1}^m \left(\sum_{j=1}^n |B_{ij}|^{p'} \right)^{q/p'} \right)^{1/q} = |B|_{q,p'}. \qquad \square$$

Using the above lemma, we can now bound the best approximation error from below.

Lemma 2. *For any square matrix $A \in \mathbb{R}^{n\times n}$ and $p, q \in [1, \infty]$, we have*

$$E_{p,q}^{n-1}(A) \ge \alpha_{q',p'}(A). \tag{5}$$

Proof. Due to Theorem 1, we have

$$E_{p,q}^{n-1}(A) = \min_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_{q'}}.$$

If A is singular, the statement is trivial. Hence let A be nonsingular. We have

$$\min_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_{q'}} = 1/\max_{x\neq 0} \frac{\|x\|_{q'}}{\|Ax\|_p} = 1/\max_{y\neq 0} \frac{\|A^{-1}y\|_{q'}}{\|y\|_p} = \frac{1}{\|A^{-1}\|_{q',p}}.$$

With Lemma 1, it follows that

$$\min_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_{q'}} \ge \frac{1}{|A^{-1}|_{q',p'}},$$

which is the desired result. $\square$

Remark 2. For a special case of the above result, Pinkus [8, Theorem 2.5] has proved equality. In particular, he shows that for $A \in \mathbb{R}^{n\times n}$ of full rank, it holds that

$$E_{1,1}^{n-1}(A) = 1/|A^{-1}|_{\infty,\infty} = \alpha_{\infty,\infty}(A)$$

(where the second equality is merely a rewriting in our notation). What is more, he proves that there exists a best rank $n-1$ approximation which agrees with A on $n-1$ rows and $n-1$ columns. Note that this result only holds for the $|\cdot|_{1,1}$-norm and the approximation by a matrix of rank $n-1$. In the case of uniform approximation, an additional condition is necessary for this equality to hold, as we show below.

Definition 3. We say that a matrix $B \in \mathbb{R}^{m\times n}$ has *rank 1 sign pattern* if there exist

$$\sigma_i, \rho_j \in \{-1, 1\}, \qquad i = 1, \dots, m,\ j = 1, \dots, n,$$

such that

$$\sigma_i B_{ij} \rho_j \ge 0 \qquad \forall i = 1, \dots, m,\ j = 1, \dots, n.$$

That is, B can be made nonnegative by flipping the sign of an arbitrary number of rows and columns.

For this class of matrices and the maximum norm, the inequality in Lemma 2 becomes an equality, giving us a characterization of the best approximation error. This is the statement of the following theorem. Here and in the sequel we write $E^k_{\max}$ for $E^k_{\infty,\infty}$.

Theorem 3. *For $A \in \mathbb{R}^{n\times n}$ of full rank whose inverse has rank 1 sign pattern, we have*

$$E^{n-1}_{\max}(A) = \alpha_{1,1}(A).$$

Proof. As in the proof of Lemma 2, we have

$$E^{n-1}_{\max}(A) = 1/\|A^{-1}\|_{1,\infty}.$$

For convenience, denote $B := A^{-1}$. Its norm is given by

$$\|B\|_{1,\infty} = \max_{\|x\|_\infty=1} \|Bx\|_1 = \max_{\|x\|_\infty=1} \sum_{i=1}^n \left| \sum_{j=1}^n B_{ij}x_j \right|. \tag{6}$$

By assumption, we have $\sigma_i, \rho_j \in \{-1, 1\}$ such that, for fixed i, the products $B_{ij}\rho_j$ are either nonnegative or nonpositive for all j. Therefore, the maximum in (6) is clearly attained for the choice $x_j := \rho_j$, and it follows

$$\|B\|_{1,\infty} = \sum_{i=1}^n \sum_{j=1}^n |B_{ij}| = |B|_{1,1} = 1/\alpha_{1,1}(A). \qquad \square$$

Remark 3. By inspection of the proof, it becomes clear that A^{-1} having rank 1 sign pattern is both a sufficient and a necessary condition for (5) to become an equality when using the maximum norm.

3.2 Properties of the error matrix

We now derive some properties of the error matrix between A and its best approximation with lower rank. For this, we first prove a characterization of minimizers of the type of expressions appearing in (1).

Lemma 3. *Let $A \in \mathbb{R}^{n\times n}$ of full rank. Let $p \in \{1, \infty\}$ and $q \in [1, \infty]$. There exists a minimizer $h^* \in \mathbb{R}^n$ for*

$$\min_{h\neq 0} \frac{\|Ah\|_p}{\|h\|_q}$$

with $\|h^*\|_q = 1$ and the following properties.

- *For $p = 1$, it satisfies $Ah^* = \pm ce_j$ for some $j = 1, \dots, n$, where $c > 0$ and e_j denotes the unit vector in the positive direction of the j-th coordinate axis.*
- *For $p = \infty$, it satisfies $Ah^* = c(\pm1, \dots, \pm1)$ for some $c > 0$.*

Proof. Since

$$\frac{\|Ah\|_p}{\|h\|_q} = \left(\frac{\|A^{-1}(Ah)\|_q}{\|Ah\|_p} \right)^{-1},$$

the minimum for $\frac{\|Ah\|_p}{\|h\|_q}$ is attained at $h^* \in \mathbb{R}^n \setminus \{0\}$ if and only if the maximum for $\frac{\|A^{-1}g\|_q}{\|g\|_p}$ is attained at $g^* = Ah^*$. Therefore we can equivalently consider

$$\max_{\|g\|_p\leq 1} \|A^{-1}g\|_q.$$

We note that $\|A^{-1}g\|_q$ is a convex function of g for all $q \in [1, \infty]$. Therefore, its maximum over the bounded, convex polytope $\{\|g\|_p \le 1\}$ is attained at an extreme point of that polytope (see, e.g., [9, Corollary 32.3.4]). Therefore, for $p = 1$, the maximizing argument has the form $g^* = \pm e_j$ for some $j = 1, \dots, n$. For $p = \infty$, it has the form $g^* = (\pm 1, \dots, \pm 1) \in \mathbb{R}^n$. The minimum in the original expression is attained at $h^* = A^{-1}g^*$, i.e.,

$$\min_{h \neq 0} \frac{\|Ah\|_p}{\|h\|_q} = \frac{\|g^*\|_p}{\|A^{-1}g^*\|_q}.$$

In both cases, the constant c is determined by $c = 1/\|A^{-1}g^*\|_q$. □

Using the above lemma, we can now prove certain properties of the error matrices in best low-rank approximation. In particular, for $p = 1$, the error is zero on $n-1$ rows, whereas for $p = q = \infty$, the error matrix equioscillates elementwise, that is, all of its entries have the same modulus.

Theorem 4. *Let $A \in \mathbb{R}^{n \times n}$ of full rank. Let $p \in \{1, \infty\}$ and $q \in [1, \infty]$. There exists a best approximation matrix $G \in \mathbb{R}^{n \times n}$ of rank $n - 1$ satisfying*

$$E_{p,q}^{n-1}(A) = |A - G|_{p,q}$$

with the following properties.

- *If $p = 1$, the matrix G agrees with A on $n - 1$ rows.*
- *If $p = \infty$, the row-wise error is constant, i.e., for the rows a_i and g_i of A and G respectively, it holds*
$$\|a_i - g_i\|_q = E_{\infty,q}^{n-1}(A) \qquad \forall i = 1, \dots, n. \tag{7}$$
If in addition $q = \infty$, then the error matrix equioscillates in the sense that
$$|a_{ij} - g_{ij}| = E_{\infty,\infty}^{n-1}(A) \qquad \forall i, j = 1, \dots, n. \tag{8}$$

Proof. Let $h^* \in \mathbb{R}^n$ be a minimizer as described in Lemma 3 for

$$\frac{\|Ah\|_p}{\|h\|_{q'}}.$$

Let G be the best approximation matrix constructed as in the proof of Theorem 1, where each row g_i is the best approximation to a_i with respect to the norm $\|\cdot\|_q$ in $U_{n-1} := \{x \in \mathbb{R}^n : x \cdot h^* = 0\}$. It was shown there that the vector of row-wise errors $e \in \mathbb{R}^n$ has the components

$$e_i := \|a_i - g_i\|_q = |[Ah^*]_i|, \qquad i = 1, \dots, n.$$

If $p = 1$, due to Lemma 3, the vector Ah^* has exactly one non-zero component. This implies that the error e_i is zero in $n - 1$ rows, which proves the desired statement.

If $p = \infty$, due to Lemma 3, the components of the vector Ah^* all have equal magnitude, which shows that e_i is constant, proving (7). If, in addition, $q = \infty$, recall that each row g_i is chosen as a minimizer of

$$\min_{g_i \in U_{n-1}} \|a_i - g_i\|_\infty.$$

Due to standard results on best uniform approximation in the linear space U_{n-1}, (see, e.g., [12, Chapter II, Theorem 1.3]), the error $a_i - g_i$ equioscillates in the sense that

$$|a_{ij} - g_{ij}| = \|a_i - g_i\|_\infty = E_{\infty,\infty}^{n-1}(A) \qquad \forall i, j = 1, \dots, n.$$ □

Remark 4. The case $p = 1$ of the above theorem was already proved by Pinkus [8, Proposition 2.4] by different means. Even more, for $p = q = 1$, he proved that there exists a best approximation matrix which agrees with A on $n-1$ rows and $n-1$ columns. The result for $p = \infty$ is new.

4 Approximations with arbitrary rank

In the case where the difference in rank between A and its approximation is greater than one, we cannot give an explicit characterization of the best approximation error; instead, we provide lower and upper bounds. To this end, we first recall some results from the theory of skeleton approximations.

4.1 Skeleton approximation

Recall that we denoted $\langle n \rangle := \{1, \dots, n\}$. Given a matrix $A \in \mathbb{R}^{m \times n}$, assume we have row and column indices $I = (i_1, \dots, i_k)$ and $J = (j_1, \dots, j_k)$ such that the submatrix $\hat{A} := A_{I,J}$ situated on rows I and columns J is nonsingular. With the matrices

$$C := A_{\langle m \rangle, J} \in \mathbb{R}^{m \times k}, \qquad R := A_{I, \langle n \rangle} \in \mathbb{R}^{k \times n}$$

containing k columns and rows of A, respectively, we can define a rank k approximation

$$C \hat{A}^{-1} R \in \mathbb{R}^{m \times n}$$

which agrees with A on the k rows I and the k columns J. If A has rank k, then $A = C\hat{A}^{-1}R$. These facts are easily verified by the Schur complement factorization formulae (see, e.g., [7]). A theoretical framework for skeleton and pseudo-skeleton approximation is given in [6].

We define, for $i \in \langle m \rangle$, $j \in \langle n \rangle$, the matrix

$$\mathcal{E}_k(i,j) := \begin{pmatrix} A_{i,j} & A_{i,j_1} & \dots & A_{i,j_k} \\ A_{i_1,j} & A_{i_1,j_1} & \dots & A_{i_1,j_k} \\ \vdots & \vdots & \ddots & \vdots \\ A_{i_k,j} & A_{i_k,j_1} & \dots & A_{i_k,j_k} \end{pmatrix} \in \mathbb{R}^{(k+1)\times(k+1)}$$

which contains $\hat{A}$ in its lower right block. The following representation of the error matrix resulting from skeleton approximation is the discrete version of an analogous identity for functions given by Schneider [11].

Lemma 4. *Let index vectors I and J be given and assume that $A_{I,J}$ is nonsingular. Then the entries of the error matrix for the skeleton approximation*

$$E = A - A_{\langle m \rangle, J} (A_{I,J})^{-1} A_{I, \langle n \rangle}$$

are given by

$$E_{i,j} = \frac{\det \mathcal{E}_k(i,j)}{\det A_{I,J}}, \qquad i = 1, \dots, m, \; j = 1, \dots, n.$$

Proof. By developing the determinant of $\mathcal{E}_k(i,j)$ along the first row and then the resulting determinants along the first column, we obtain

$$\det \mathcal{E}_k(i,j) = A_{ij} \det A_{I,J} - \sum_{\alpha,\beta=1}^{k} (-1)^{\alpha+\beta} \det B(\alpha,\beta) A_{i,j_\beta} A_{i_\alpha, j},$$

where $B(\alpha, \beta) \in \mathbb{R}^{(k-1)\times(k-1)}$ denotes the matrix $A_{I,J}$ with the α-th row and the β-th column removed. Thus we obtain

$$\frac{\det \mathcal{E}_k(i,j)}{\det A_{I,J}} = A_{ij} - \sum_{\alpha,\beta=1}^{k} A_{i,j_\beta} \frac{(-1)^{\alpha+\beta} \det B(\alpha,\beta)}{\det A_{I,J}} A_{i_\alpha,j}.$$

Since the entries of the inverse matrix are given by the well-known cofactor identity

$$[A_{I,J}^{-1}]_{\beta,\alpha} = \frac{(-1)^{\alpha+\beta} \det B(\alpha,\beta)}{\det A_{I,J}},$$

the statement follows. ☐

4.2 Upper and lower bounds on the approximation error

Before we derive approximation error bounds, we need the simple result that submatrices can be approximated at least as well as the whole matrix.

Lemma 5. *Let $A \in \mathbb{R}^{m\times n}$. For any submatrix B of A, we have*

$$E^k_{p,q}(B) \le E^k_{p,q}(A).$$

Proof. Let $G_A \in \mathbb{R}^{m\times n}$ be a matrix of rank at most k which realizes the best approximation error to A. Such a matrix always exists since the set of matrices of rank at most k is closed. By deleting the appropriate rows and columns from G_A, we obtain a matrix G_B which has the same size as B and rank at most k. Clearly, $E^k_{p,q}(B) \le |B - G_B|_{p,q} \le |A - G_A|_{p,q} = E^k_{p,q}(A)$. ☐

The next result generalizes Lemma 2 to the case of approximation by matrices with arbitrary rank.

Lemma 6. *Let $A \in \mathbb{R}^{m\times n}$ and $p, q \in [1,\infty]$. Then*

$$E^{k-1}_{p,q}(A) \ge \max_{B\in\mathcal{S}_k(A)} \alpha_{q',p'}(B).$$

Proof. For any submatrix $B \in \mathcal{S}_k(A)$, using Lemma 2 we have

$$E^{k-1}_{p,q}(B) \ge \alpha_{q',p'}(B).$$

The statement then follows from Lemma 5. ☐

Using Lemma 6, we can now prove a lower and upper bound for the best approximation error with low rank matrices in the maximum norm. The quantity which we use in the lower and upper bounds involves the maximally achievable modulus of determinants of submatrices of certain sizes. Goreinov and Tyrtyshnikov [4] use such "maximal volumes" to derive some error estimates for skeleton approximation of matrices. In the following, we follow their nomenclature and call the modulus of the determinant the "volume" of a matrix.

Let

$$\beta_k(A) := \frac{\max_{B_k\in\mathcal{S}_k(A)} |\det B_k|}{\max_{B_{k-1}\in\mathcal{S}_{k-1}(A)} |\det B_{k-1}|},$$

that is, the quotient of the maximal volumes obtainable in a size k and a size $k-1$ submatrix, respectively. If the denominator is zero, then so is the numerator and we set $\beta_k(A) := 0$ in this case. If $k = 1$, the denominator is considered to be 1 so that $\beta_1(A) = |A|_{\max}$.

The following result is similar to one obtained by Babaev [1] in the continuous setting using his technique of exact annihilators. It is possible to employ this proof technique also in our discrete setting. However, we found that using the approximation result stated in Theorem 1 yields a much shorter proof.

Theorem 5. *Let $A \in \mathbb{R}^{m\times n}$ and $k \le \min\{m,n\}$ an integer. We have the bounds*

$$1/k^2 \beta_k(A) \le E_{\max}^{k-1}(A) \le \beta_k(A).$$

Proof. We first prove the lower bound. Let $B_k \in \mathcal{S}_k(A)$ be of maximal volume. From Lemma 6 we have

$$E_{\max}^{k-1}(A) \ge \alpha_{1,1}(B_k).$$

If B_k is singular, then A has rank at most $k-1$ and the statement is trivial. Otherwise, using the cofactor identity for the entries of B_k^{-1}, we obtain

$$\begin{aligned}\alpha_{1,1}(B_k) = \frac{1}{|B_k^{-1}|_{1,1}} = \frac{|\det B_k|}{\sum_{C\in\mathcal{S}_{k-1}(B_k)} |\det C|} &\ge \frac{|\det B_k|}{k^2 \max_{C\in\mathcal{S}_{k-1}(B_k)} |\det C|} \\ &\ge \frac{|\det B_k|}{k^2 \max_{B_{k-1}\in\mathcal{S}_{k-1}(A)} |\det B_{k-1}|} = 1/k^2\beta_k(A).\end{aligned}$$

To prove the upper bound, we let $B_{k-1} \in \mathcal{S}_{k-1}(A)$ and $B_k \in \mathcal{S}_k(A)$ be of maximal volume. We denote by $I = (i_1, \ldots, i_{k-1})$ and $J = (j_1, \ldots, j_{k-1})$ the rows and columns on which B_{k-1} is located within A, that is, $B_{k-1} = A_{I,J}$. We construct the skeleton approximation

$$A_{k-1} = A_{\langle m\rangle,J}(B_{k-1})^{-1} A_{I,\langle n\rangle} \in \mathbb{R}^{m\times n},$$

which has rank at most $k-1$. Therefore,

$$E_{\max}^{k-1}(A) \le |A - A_{k-1}|_{\max}.$$

Due to Lemma 4, the entries of the error matrix satisfy

$$|A - A_{k-1}|_{ij} = \frac{|\det \mathcal{E}_{k-1}(i,j)|}{|\det B_{k-1}|} \tag{9}$$

Since $\mathcal{E}_{k-1}(i,j)$ is a submatrix of size $k\times k$ of A and B_k is the maximal volume submatrix of that size, we have

$$|\det \mathcal{E}_{k-1}(i,j)| \le |\det B_k|$$

and thus

$$|A - A_{k-1}|_{\max} \le \beta_k(A), \tag{10}$$

which finishes the proof. □

Remark 5. The low-rank approximation A_{k-1} used in the above proof is just the skeleton approximation where the skeleton component is chosen as the submatrix B_{k-1} of maximal volume, cf. [6]. Equation (10) thus also gives an error estimate for this approximation method, and by combining it with the lower bound from Theorem 5 we obtain the quasi-optimality estimate proven in [5],

$$|A - A_{k-1}|_{\max} \le k^2 E_{\max}^{k-1}(A).$$

5 The case of 2×2 matrices

We begin by considering some examples of different rank 1 approximation techniques in the case of 2×2 matrices and observe that common techniques (cross approximation, approximation in the space spanned by one column) do not generally yield best approximations in the maximum norm. We then give a closed formula for the best uniform approximation error in this setting and examine the sharpness of our error estimates.

Example 1. Let

$$A = \begin{pmatrix} 1 & b \\ b & 0 \end{pmatrix}$$

with $b > 0$. We study the maximum entrywise error of various rank 1 approximation strategies.

Approximation matching on one column and one row (Skeleton approximation). By pivoting on the upper left or upper right matrix entry, respectively, we obtain the rank 1 skeleton approximations

$$C_1 = \begin{pmatrix} 1 & b \\ b & b^2 \end{pmatrix}, \qquad C_2 = \begin{pmatrix} 1 & b \\ 0 & 0 \end{pmatrix}$$

which have maximum errors

$$|A - C_1|_{\max} = b^2, \qquad |A - C_2|_{\max} = b.$$

Pinkus [8] shows that in the $|\cdot|_{1,1}$-norm, there always exists a best approximation of this form, i.e., matching in all but one entry. The same does not hold for the $|\cdot|_{\max}$-norm.

Approximation matching on one column. By approximating the right column by a scaled version of the left one or vice versa, we obtain the rank 1 approximations

$$\begin{pmatrix} 1 & \gamma \\ b & \gamma b \end{pmatrix}, \qquad \begin{pmatrix} \gamma b & b \\ 0 & 0 \end{pmatrix}, \qquad \gamma \in \mathbb{R},$$

with best achievable approximation errors using $\gamma = b/(b+1)$ and $\gamma = 1/b$, respectively,

$$\frac{b^2}{b+1} < \min\{b^2, b\} \qquad \text{and} \qquad b.$$

Pinkus [8] shows that in the $|\cdot|_{p,1}$-norm for any $p \in [1, \infty]$, there always exists a best approximation matching on all but one column; our Theorem 4 makes the analogous statement for the rows. No such best approximating matrix exists in the $|\cdot|_{\max}$-norm.

Approximation by a constant matrix. By choosing the best possible constant $d = \max(1, b)/2$, we obtain the error

$$\left| A - \begin{pmatrix} d & d \\ d & d \end{pmatrix} \right|_{\max} = \frac{\max(1, b)}{2}.$$

Approximation by arbitrary rank 1 matrix. With the rank 1 matrix

$$G = \frac{1}{2b+1} \begin{pmatrix} b+1 \\ b \end{pmatrix} \begin{pmatrix} b+1 & b \end{pmatrix},$$

we obtain

$$|A - G|_{\max} = \frac{b^2}{2b+1}$$

which is better than all previous approximations. In fact, we have $\alpha_{1,1}(A) = \frac{b^2}{2b+1}$, and thus it follows from Lemma 2 that G is the best rank 1 approximation in the $|\cdot|_{\max}$-norm. The error matrix

$$A - G = \frac{b^2}{2b+1}\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

equioscillates in the sense of Theorem 4.

We point out that an invertible 2×2 matrix, due to the way its inverse is formed, has a rank 1 sign pattern (in the sense of Definition 3) if and only if its inverse has a rank 1 sign pattern. Thus A satisfies the assumptions of Theorem 3, which is another way to show that $\alpha_{1,1}(A) = \frac{b^2}{2b+1}$ is its best approximation error.

The following theorem gives a closed formula for the best uniform approximation error of a 2×2 matrix by a rank 1 matrix.

Theorem 6. *A nonsingular real matrix* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *has uniform best approximation error*

$$E^1_{\max}(A) = \frac{|\det A|}{\max\{|a+c| + |b+d|, |a-c| + |b-d|\}}$$

by a matrix of rank 1 or less.

Proof. From Theorem 1, we have

$$E^1_{\max}(A) = \min_{h \neq 0} \frac{\|Ah\|_\infty}{\|h\|_1}.$$

As in the proof of Lemma 3, it follows that

$$E^1_{\max}(A) = \min_{g \neq 0} \frac{\|g\|_\infty}{\|A^{-1}g\|_1} = \left(\max_{g\neq 0} \frac{\|A^{-1}g\|_1}{\|g\|_\infty}\right)^{-1}$$

and that the maximum is attained in one of the four corners of $[-1,1]^2$. Since

$$A^{-1} = \frac{1}{\det A}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

the statement follows by considering the four cases

$$\max\left\{\|A^{-1}\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)\|_1, \|A^{-1}\left(\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right)\|_1, \|A^{-1}\left(\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right)\|_1, \|A^{-1}\left(\begin{smallmatrix} -1 \\ -1 \end{smallmatrix}\right)\|_1\right\},$$

of which the first and last as well as the second and third have identical values. $\square$

Since $\alpha_{1,1}(A) = \frac{|\det A|}{|a|+|b|+|c|+|d|}$ as one can compute directly, and recalling that an invertible 2×2 matrix has a rank 1 sign pattern if and only if its inverse does, Theorem 6 illustrates that the condition given in Definition 3 is both necessary and sufficient for the statement of Theorem 3 to hold, i.e., for the best approximation error to be equal to $\alpha_{1,1}(A)$, in the 2×2 setting. The following example illustrates a case where the rank 1 sign pattern condition does not hold and therefore the inequality in Lemma 2 is strict, i.e., the best approximation error is strictly greater than $\alpha_{1,1}(A)$.

Example 2. Let $A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$. Then $\alpha_{1,1}(A) = \frac{1}{2}$. However, Theorem 6 shows that the best rank 1 approximation G has $|A - G|_{\max} = 1$ (realized, for instance, by the zero matrix).

Remark 6. For $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ of full rank, we have

$$E^1_{\max}(A) = \frac{|\det A|}{\max\{|a+c|+|b+d|, |a-c|+|b-d|\}}, \qquad \beta_2(A) = \frac{|\det A|}{\max\{|a|,|b|,|c|,|d|\}}.$$

In Theorem 5, which holds for approximations with arbitrary ranks, we proved the lower and upper bounds

$$\tfrac{1}{4}\beta_2(A) \le E^1_{\max}(A) \le \beta_2(A).$$

Without loss of generality, assume that a is the maximum entry of A in modulus. If the upper bound in the above estimate were to be sharp, we would have

$$\max\{|a+c|+|b+d|, |a-c|+|b-d|\} = |a|.$$

If $c \neq 0$, the above maximum is greater than $|a|$. If $c = 0$, the only way to achieve equality is by setting $b = d = 0$, in which case A does not have full rank.

For the lower bound in Theorem 5 to be sharp, we would require

$$4|a| = \max\{|a+c|+|b+d|, |a-c|+|b-d|\},$$

and by similar arguments, one easily sees that this cannot hold if A has full rank.

This shows that, in the case of 2×2 matrices of full rank, both the lower and the upper bounds in Theorem 5 are never sharp, i.e., we have

$$\tfrac{1}{4}\beta_2(A) < E^1_{\max}(A) < \beta_2(A).$$

Acknowledgments

The authors would like to thank Matúš Benko (JKU Linz) for helpful discussions.

The authors acknowledge the support by Grant DNTS-Austria 01/6 funded by the Bulgarian National Science Fund. The second author was supported by the National Research Network "Geometry + Simulation" (NFN S117, 2012–2016), funded by the Austrian Science Fund (FWF).

References

[1] M.-B. A. Babaev. Best approximation by bilinear forms. *Mathematical Notes*, 46(2):588–596, 1989. doi: 10.1007/BF01137621.

[2] R.A. DeVore and G.G. Lorentz. *Constructive Approximation*, volume 303 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 1993. ISBN 978-3-540-50627-0.

[3] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. ISSN 1860-0980. doi: 10.1007/BF02288367.

[4] S. A. Goreinov and E. E. Tyrtyshnikov. The maximal-volume concept in approximation by low-rank matrices. *Contemporary Mathematics*, 280:47–52, 2001.

[5] S. A. Goreinov and E. E. Tyrtyshnikov. Quasioptimality of skeleton approximation of a matrix in the Chebyshev norm. *Doklady Mathematics*, 83(3):374–375, 2011. ISSN 1531-8362. doi: 10.1134/S1064562411030355.

[6] S. A. Goreinov, E. E. Tyrtyshnikov, and N. L. Zamarashkin. A theory of pseudoskeleton approximations. *Linear Algebra and its Applications*, 261(1):1–21, 1997. doi: 10.1016/S0024-3795(96)00301-1.

[7] R. A. Horn and F. Zhang. Basic properties of the Schur complement. In F. Zhang, editor, *The Schur Complement and Its Applications*, pages 17–46. Springer US, Boston, MA, 2005. ISBN 978-0-387-24273-6. doi: 10.1007/0-387-24273-2_2.

[8] A. Pinkus. On best rank n matrix approximations. *Linear Algebra and its Applications*, 437 (9):2179–2199, 2012. ISSN 0024-3795. doi: 10.1016/j.laa.2012.05.016.

[9] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, N.J., 1970. ISBN 0691015864.

[10] E. Schmidt. Zur Theorie der linearen und nichtlinearen Integralgleichungen. *Mathematische Annalen*, 63(4):433–476, 1907. ISSN 1432-1807. doi: 10.1007/BF01449770.

[11] J. Schneider. Error estimates for two-dimensional cross approximation. *J. Approx. Theory*, 162(9):1685–1700, September 2010. ISSN 0021-9045. doi: 10.1016/j.jat.2010.04.012.

[12] I. Singer. *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, volume 171 of *Grundlehren der mathematischen Wissenschaften.* Springer Berlin Heidelberg, 1970. ISBN 978-3-662-41583-2. doi: 10.1007/978-3-662-41583-2.

Latest Reports in this series

2009 - 2014

[..]

2015

[..]

2015-09	Peter Gangl, Samuel Amstutz and Ulrich Langer *Topology Optimization of Electric Motor Using Topological Derivative for Nonlinear Magnetostatics*	June 2015
2015-10	Clemens Hofreither, Stefan Takacs and Walter Zulehner *A Robust Multigrid Method for Isogeometric Analysis using Boundary Correction*	July 2015
2015-11	Ulrich Langer, Stephen E. Moore and Martin Neumüller *Space-time Isogeometric Analysis of Parabolic Evolution Equations*	September 2015
2015-12	Helmut Gfrerer and Jiří V. Outrata *On Lipschitzian Properties of Implicit Multifunctions*	Dezember 2015

2016

2016-01	Matúš Benko and Helmut Gfrerer *On Estimating the Regular Normal Cone to Constraint Systems and Stationarity Conditions*	May 2016
2016-02	Clemens Hofreither and Stefan Takacs *Robust Multigrid for Isogeometric Analysis Based on Stable Splittings of Spline Spaces*	June 2016
2016-03	Helmut Gfrerer and Boris S. Mordukhovich *Robinson Stability of Parametric Constraint Systems via Variational Analysis*	August 2016
2016-04	Helmut Gfrerer and Jane J. Ye *New Constraint Qualifications for Mathematical Programs with Equilibrium Constraints via Variational Analysis*	August 2016
2016-05	Matúš Benko and Helmut Gfrerer *An SQP Method for Mathematical Programs with Vanishing Constraints with Strong Convergence Properties*	August 2016
2016-06	Peter Gangl and Ulrich Langer *A Local Mesh Modification Strategy for Interface Problems with Application to Shape and Topology Optimization*	September 2016
2016-07	Bernhard Endtmayer and Thomas Wick *A Partition-of-Unity Dual-Weighted Residual Approach for Multi-Objective Goal Functional Error Estimation Applied to Elliptic Problems*	October 2016
2016-08	Matúš Benko and Helmut Gfrerer *New Verifiable Stationarity Concepts for a Class of Mathematical Programs with Disjunctive Constraints*	November 2016
2016-09	Dirk Pauly and Walter Zulehner *On Closed and Exact* $\operatorname{Grad}\operatorname{grad}$- *and* $\operatorname{div}\operatorname{Div}$-*Complexes, Corresponding Compact Embeddings for Tensor Rotations, and a Related Decomposition Result for Biharmonic Problems in 3D*	November 2016
2016-10	Irina Georgieva and Clemens Hofreither *On the Best Uniform Approximation by Low-Rank Matrices*	December 2016

From 1998 to 2008 reports were published by SFB013. Please see
`http://www.sfb013.uni-linz.ac.at/index.php?id=reports`
From 2004 on reports were also published by RICAM. Please see
`http://www.ricam.oeaw.ac.at/publications/list/`
For a complete list of NuMa reports see
`http://www.numa.uni-linz.ac.at/Publications/List/`