Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

New Estimates for the Recursive Low-Rank Truncation

(revised version: September 2014)

by

Wolfgang Hackbusch

Preprint no.: 34 2014

New Estimates for the Recursive Low-Rank Truncation

Wolfgang Hackbusch
Max-Planck-Institut *Mathematik in den Naturwissenschaften*
Inselstr. 22, D-04103 Leipzig

Abstract

The best approximation of a matrix by a low-rank matrix can be obtained by the singular value decomposition. For large-sized matrices this approach is too costly. Instead one may use a block decomposition. Approximating the small submatrices by low-rank matrices and agglomerating them into a new, coarser block decomposition, one obtains a recursive method. The required computation work is $O(rnm)$ if r is the desired rank and $n \times m$ is the size of the matrix. The paper discusses the errors $A - B$ and $M - A$ where A is the result of the recursive truncation applied to M, while B is the best approximation.

AMS Subject Classifications: 65Fxx, 65F30, 15A18, 15A45
Key words: low-rank approximation, singular value decomposition, error estimate, hierarchical matrix

1 Introduction

We consider large-scale matrices $M \in \mathbb{R}^{n\times m}$ which we want to approximate by low-rank matrices. This means that we want to determine a matrix A of rank $r \ll \min\{n, m\}$ such that $\|M - A\| / \|M\|$ is small. A matrix of rank $\le r$ can be represented in the product form $A = XY^\mathsf{T}$ with $X \in \mathbb{R}^{n\times r}$ and $Y \in \mathbb{R}^{m\times r}$. We recall that a matrix A has the rank r, if $\mathrm{range}(A) = \mathrm{range}(X)$ and $\mathrm{range}(A^\mathsf{T}) = \mathrm{range}(Y)$ have the dimension r. The representation of A by XY^T allows us to reduce the storage size nm of the general matrix A to $(n + m)\, r$ for X and Y. Also operations involving the matrix A are much cheaper to perform using the representation XY^T.

In the optimal case, the matrix has fast decaying singular values σ_k. For instance, the technique of hierarchical matrices is based on the fact that suitable submatrices of a boundary element matrix or of the inverse of a finite element matrix are of this kind (cf. [4], [3], [1]). In these cases, the matrix M from above has to be replaced by a particular (non-principal) submatrix.

It is well known that the best low-rank matrix can be determined by the singular value decomposition (SVD, cf. [9]). Let $M = \sum_k \sigma_k u_k v_k^\mathsf{T}$ be the singular value decomposition characterised by the singular values $\sigma_1 \ge \sigma_2 \ge \ldots$ and orthonormal systems $\{u_k\} \subset \mathbb{R}^n$ and $\{v_k\} \subset \mathbb{R}^m$. Then the 'SVD truncation' $B := \mathrm{SVD}_r(M) := \sum_{k=1}^r \sigma_k u_k v_k^\mathsf{T}$ is the best solution with respect to the spectral and Frobenius norm. Throughout this article, we use the Frobenius norm $\|M\| := \sqrt{\sum_{i,j} M_{ij}^2}$ as matrix norm. Orthogonality $X \perp Y$ of matrices is understood with respect to the Frobenius scalar product $\langle X, Y\rangle = \sum_{i,j} X_{ij} Y_{ij}$. The remainder of the SVD truncation is $\|M - B\| = \sqrt{\sum_{k>r} \sigma_k^2}$. If, for instance, σ_k decays like $\exp(-\omega k)$, the accuracy $\|M - B\| / \|M\| \approx \varepsilon$ can be obtained by the choice $r \approx \frac{1}{2\omega} \log(\varepsilon^{-2} + 1) \approx \frac{1}{\omega} \log \frac{1}{\varepsilon}$. However, the difficulty is the fact that the computation of the singular value decomposition is very expensive. The cost is cubic in the size of the matrix. Furthermore, most of the SVD data are contained in the remainder $M - B$ and not in the desired part B.

For many numerical purposes it is not necessary to use the best approximation $B := \mathrm{SVD}_r(M)$ of M. Quasi-optimal approximations are also welcome. These are rank-r matrices A characterised by

$$\|A - B\| \le C \, \|M - B\| \quad \text{and} \quad \|M - A\| \le C' \, \|M - B\| \,. \tag{1.1}$$

Here it is essential that the constants C and C' in (1.1) do not depend on the matrix M.

An obvious approach is the approximation of the smallest r eigenvalues (the squared singular values) and corresponding eigenvectors u_k of MM^T. This yields $A := \sum_{k=1}^r \tilde\sigma_k \tilde u_k \tilde v_k^\mathsf{T}$ with $\tilde v_k := M^\mathsf{T} \tilde u_k / \tilde\sigma_k$ as a possible

rank-r approximation of M. The estimation of $\|M - A\|$ can be based on the error estimates of $\tilde{\sigma}_k$ and $\tilde{u}_k$. Refined error estimates for the symmetric eigenvalue problem can, e.g., be found in Kandler–Schröder [7]. Unfortunately, these estimates always involve some kind of gap condition. Note that the later worst-case example in (4.5) has a vanishing gap. Therefore, it is an open question how to get an universal bound for C' in (1.1) with A obtained by Krylov methods.

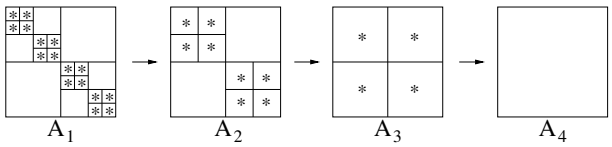
In this article we describe a cheap method for which we can prove quasi-optimality (1.1). Different versions of the algorithm yield different constants; however, in practice, the results are much better (often almost as good as the exact SVD truncation; see the numerical examples in Section 9). The algorithm uses a 'recursive truncation' and is an obvious divide and conquer method[1] (precise definition in Section 3):

(A) Assume first that the $n \times n$ matrix $\square$ has the block structure $\boxplus$ with submatrices of rank $\leq r$. Consequently, $\square$ has a rank bounded by $4r$. Its truncation to rank r can be determined by $\mathcal{O}(nr^2 + r^3)$ operations.

(B) Assume a $2^L \times 2^L$ block structure of the target matrix M with $L > 1$ and blocks of rank $\leq r$ (e.g., $\boxplus$ for $L = 2$). This matrix can be considered as a $2^{L-1} \times 2^{L-1}$ block matrix with all blocks subdivided into $\boxplus$. Use Step (A) to approximate each block $\boxplus$ by a submatrix of rank $\leq r$. The result is a $2^{L-1} \times 2^{L-1}$ block matrix M with blocks of rank $\leq r$. Hence, we can apply again (B) if $L - 1 > 1$, or (A) if $L - 1 = 1$.

For the error analysis we consider any splitting $M = B - \Delta$ such that $\operatorname{rank}(B) \leq r$ ($B := \mathrm{SVD}_r(M)$ is a possible choice). Let A of rank $\leq r$ be the result of the mentioned recursive truncation. We shall prove an estimate of the form $\|A - B\| \leq C \|\Delta\|$. This implies that $\|M - A\| \leq (1 + C) \|\Delta\|$, i.e., inequality (1.1) holds with $C' := 1 + C$. The decisive factor C will be characterised. If the rank-r truncation by singular value decomposition is successful, i.e., if $\|\Delta\|$ is small, also A is a good approximation of B and M.

Primarily, these methods are implemented in the hierarchical matrix technique (cf. [3, Alg. 2.11]). The multiplication $Z := XY$ of two hierarchical matrices leads to an intermediate result Z. The hierarchical format requires that certain submatrices $Z|_b$ must be represented as a rank-r matrix (the exact definition of the restriction $Z|_b$ to the block b is given in (2.1)). However, the intermediate result $Z|_b$ is subdivided into further submatrices (cf. [4, §7.4], [3, §2.1.6]). For instance, $A_1 := Z|_b$ may be substructured as depicted below:



Replacing the 2×2 blocks indicated by stars with rank-r matrices, we obtain A_2 and analogously A_3. Finally, $A := A_4$ is a global rank-r matrix. Also the recompression technique leads to a similar situation (cf. [4, §6.7.2]).

The same procedure can be applied to a regular block subdivision as indicated above. The computational cost of the algorithm as well as the error analysis depends on the kind of the tree describing the recursive subdivision (cf. §3.1). In the general case,[2] the computational work is of the order $\mathcal{O}(rnm)$.

Although the bound C in (1.1) depends exponentially on the depth of the block decomposition tree (cf. (3.4a,b)), the observed behaviour is much better. In fact, one can argue that high amplification factors are not probable (cf. Corollary 5.12). In the special case of unidirectional block decompositions also the theoretical bound C is much better behaved.

The basic linear algebra tools are (1) the QR decomposition (possibly with pivoting) and (2) the SVD algorithm. The latter algorithm is only applied to matrices of the size $\mathcal{O}(r)$. The QR computation is needed to split general (small) matrices into the product XY^{T} where X and Y consist of r columns. To bound the cost of the described algorithms explicitly, we assume that the QR decomposition of an $n \times m$ matrix ($m \leq n$) costs $4nm^2$ operations, while a singular value decomposition of an $n \times n$ matrix requires $21n^3$ operations. These numbers are taken from [2, §5.2.9 and §5.4.5]. In particular the last number should not be misunderstood as a strict bound but as an empirical value.[3]

[1] The algorithm is completely different from spectral divide and conquer methods, as they are analysed, e.g., in [8] and the literature cited therein.

[2] The applications to hierarchical matrices are cheaper because of their special structures.

[3] In principle, the cost must depend on the underlying machine precision. The constants also depend on the implemented version of the algorithms etc. We introduce these explicit numbers to be able to indicate the cost of the later algorithm quantitatively. Otherwise, only statements using $\mathcal{O}(\ldots)$ are possible.

Another recursive truncation is used for the rank-r truncation of a sum $M = \sum_{k=0}^{L} M_k$ of matrices M_k of rank $\le r$. The recursion $S_0 := M_0$ and $S_k := \mathrm{SVD}_r(S_{k-1} + M_k)$ $(1 \le k \le L)$ results in $A = S_L$, which is taken as an approximation of $B = \mathrm{SVD}_r(M)$ (cf. [4, §2.6.3]). Here, the well-known cancellation effect can occur. Already for $L = 1$, one can construct matrices M_0 and M_1 such that $M = M_0 + M_1 = \mathrm{SVD}_r(M) = B$, while $\mathrm{SVD}_r(M_0) = -\mathrm{SVD}_r(M_1)$ so that $A = 0$. In this case, inequality (1.1) cannot hold because of $M - B = O$: $\|A - B\| = \|M\| \not\le C\,\|\Delta\|$. However, in the present case, cancellation is excluded, since we consider the agglomeration $A, B \mapsto [A\ B]$ of matrices. Formally, the matrices are extended by zero: $A \mapsto [A\ O]$ and $B \mapsto [O\ B]$ (this does not change the rank) and the extended matrices are added: $[A\ B] = [A\ O] + [O\ B]$. Cancellation is prevented by the fact that the latter terms are orthogonal with respect to the Frobenius scalar product.

In Section 2 we introduce the notations, the representation of rank-r matrices in the product form XY^{T}, and the singular value decomposition.

The recursive truncation algorithm is described in Section 3. The algorithm in §3.2 uses a partition tree introduced in §3.1. We illustrate the block decomposition by four typical model examples.

The analysis in Section 5 applies to general block decompositions. We discuss the error estimate in the general case and for the mentioned model examples. The proof is based on an estimate stated in Theorem 4.5. Although this result is sharp, we show in Section 6 that the worst case is rather improbable. In particular, for applications which are typical for hierarchical matrices, we show much better estimates under the condition that the error $M - \mathrm{SVD}_r(M)$ is equally distributed over the matrix entries.

A particular block decomposition is the unidirectional splitting studied in Section 7. In this case, quite different tools can be used for an estimate of the error. The bound is given by the square root of the depth of the tree (see Theorem 7.2). Note that the depth of the tree is $\mathcal{O}(\log n)$, where n is the size of the matrix. Section 8 discusses the combination of two unidirectional splittings (column-wise and row-wise). The previous result still allows satisfactory error bounds. The numerical examples in Section 9 show that the practical results are by far better than the theoretical bounds.

2 Basic Statements

2.1 Notations

If S is any set, $\#S$ denotes the cardinality of this set.

If I and J are finite index sets, $\mathbb{R}^I$ is the set of vectors $(v_i)_{i\in I}$ with $v_i \in \mathbb{R}$. Similar for $\mathbb{R}^J$. We consider real vector spaces, but note without further comment that the algorithms extend straightforwardly to the complex case.

If $\tau \subset I$ is a subset, the restriction $v|_\tau \in \mathbb{R}^\tau$ denotes the vector $(v_i)_{i\in\tau}$.

The vector space $\mathbb{R}^{I\times J}$ of matrices consists of $(M_{ij})_{(i,j)\in I\times J}$, where $M_{ij} \in \mathbb{R}$. The set $b = \tau \times \sigma$ with $\tau \subset I$ and $\sigma \subset J$ is called a block (in $I \times J$). The restriction of a matrix M to b yields the matrix block

$$M|_b := (M_{ij})_{(i,j)\in\tau\times\sigma} \in \mathbb{R}^{\tau\times\sigma}. \tag{2.1}$$

Vice versa, the extension of $M \in \mathbb{R}^{\tau\times\sigma}$ to the larger index set $I \times J$ is defined by $M|^{I\times J}$ with the entries

$$(M|^{I\times J})_{ij} = M_{ij} \text{ for } (i,j) \in \tau\times\sigma \text{ and } (M|^{I\times J})_{ij} = 0 \text{ otherwise.}$$

$\mathcal{R}(r, I, J)$ denotes the subset of matrices of a rank not exceeding r:

$$\mathcal{R}(r, I, J) := \left\{ M \in \mathbb{R}^{I\times J} : \mathrm{rank}(M) \le r \right\}.$$

The agglomeration $A, B \mapsto [A\ B]$ of $A \in \mathbb{R}^{\tau\times\sigma'}$ and $B \in \mathbb{R}^{\tau\times\sigma''}$ into $[A\ B] \in \mathbb{R}^{\tau\times\sigma}$ for the disjoint union $\sigma = \sigma' \dot\cup \sigma''$ is formally defined by

$$[A\ B] := A|^{\tau\times\sigma} + B|^{\tau\times\sigma}. \tag{2.2a}$$

Similarly, we have agglomerations of the types

$$A, B \mapsto \begin{bmatrix} A \\ B \end{bmatrix} \in \mathbb{R}^{\tau\times\sigma} \quad \text{for } A \in \mathbb{R}^{\tau'\times\sigma},\ B \in \mathbb{R}^{\tau''\times\sigma},\ \tau = \tau' \dot\cup \tau'' \tag{2.2b}$$

and

$$A_{11}, A_{12}, A_{21}, A_{22} \mapsto \begin{bmatrix} A_{12} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad \text{for } A_{ij} \in \mathbb{R}^{\tau_i \times \sigma_j},\ \sigma = \sigma_1 \dot\cup \sigma_2,\ \tau = \tau_1 \dot\cup \tau_2. \tag{2.2c}$$

For all cases, we use the respective notations $\mathrm{Aggl}\{A, B\}$ and $\mathrm{Aggl}\{A_{11}, A_{12}, A_{21}, A_{22}\}$. The context will indicate the underlying case (2.2a–c).

2.2 r-Term Representation

Whenever a matrix M belongs to $\mathcal{R}(r, \tau, \sigma)$, we use the representation

$$M = XY^{\mathsf{T}} \text{ with } X \in \mathbb{R}^{\tau \times \rho},\ Y \in \mathbb{R}^{\sigma \times \rho},\ \rho = \{1, \ldots, r\}. \tag{2.3}$$

If $x^{(k)}, y^{(k)}$ are the k-th columns of X and Y, the equivalent equation $M = \sum_{k=1}^{r} x^{(k)} y^{(k)\mathsf{T}}$ explains the name 'r-term representation'. $M \in \mathcal{R}(r, \tau, \sigma)$ may have a rank r' lower than r. Then, formally, we can add $r - r'$ zero terms to again obtain $\sum_{k=1}^{r} x^{(k)} y^{(k)\mathsf{T}}$.

The sum of $M_1 = X_1 Y_1^{\mathsf{T}}$ and $M_2 = X_2 Y_2^{\mathsf{T}}$ can be described without computational cost. Write $X_1 \in \mathbb{R}^{\tau \times \rho_1}$ and $Y_1 \in \mathbb{R}^{\sigma \times \rho_1}$ with $\rho_1 = \{1, \ldots, r\}$ as in (2.3), but use the disjoint set $\rho_2 = \{r+1, \ldots, 2r\}$ for $X_2 \in \mathbb{R}^{\tau \times \rho_2}$ and $Y_2 \in \mathbb{R}^{\sigma \times \rho_2}$. Then

$$M_1 + M_2 = M = [X_1\ X_2][Y_1\ Y_2]^{\mathsf{T}} \quad \text{with } X \in \mathbb{R}^{\tau \times \rho},\ Y \in \mathbb{R}^{\sigma \times \rho},\ \rho = \{1, \ldots, 2r\}$$

involves the agglomerated factors $X = [X_1\ X_2]$ and $Y = [Y_1\ Y_2]$, for which the enlarged ranks are at most $2r$.

Remark 2.1 *Let $\#\tau = \#\rho = r$. If a matrix $M \in \mathbb{R}^{\tau \times \rho}$ is given as full matrix, M belongs to $\mathcal{R}(r, \tau, \rho)$ and the QR decomposition $M = QR$ yields the r-term representation (2.3). The related computational cost is $4r^3$ (cf. [2, §5.2.9]).*

2.3 Singular Value Decomposition

The singular value decomposition of a matrix M with $\mu := \mathrm{rank}(M)$ is the μ-term representation

$$M = \sum_{k=1}^{\mu} \sigma_k u_k v_k^{\mathsf{T}} \quad \text{with } \{u_k\} \text{ and } \{v_k\} \text{ orthonormal, } \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_\mu > 0. \tag{2.4}$$

For $r \leq \mu$ the matrix $M \in \mathbb{R}^{I \times J}$ can be split into

$$M = A + E \quad \text{with } A = \mathrm{SVD}_r(M) := \sum_{k=1}^{r} \sigma_k u_k v_k^{\mathsf{T}},$$

while $E = \sum_{k>r} \sigma_k u_k v_k^{\mathsf{T}}$ is the remainder. $\mathrm{SVD}_r(M)$ is called the *rank-r truncation* of M. $\mathrm{SVD}_r(M)$ is optimal in the sense that $\|M - \mathrm{SVD}_r(M)\| = \min\{\|M - B\| : B \in \mathbb{R}^{I \times J}, \mathrm{rank}(B) \leq r\}$. Note that $\mathrm{SVD}_r(M)$ is uniquely defined if and only if $\sigma_r \neq \sigma_{r+1}$. If $\sigma_r = \sigma_{r+1}$, $\mathrm{SVD}_r(M)$ denotes one of the possible solutions selected by the particular software.

A typical SVD application is the truncation of a matrix M given in the format $M = XY^{\mathsf{T}}$ with $X \in \mathbb{R}^{n \times s}$ and $Y \in \mathbb{R}^{m \times s}$, $s > r$, to rank r. Concerning the stability of the representation $M = XY^{\mathsf{T}}$, note that the following constructions (QR, SVD) lead to factors which are either orthogonal[4] up to scaling or at least triangular.

Algorithm 2.2 *Assume $M = XY^{\mathsf{T}} \in \mathbb{R}^{n \times m}$ with $X \in \mathbb{R}^{n \times s}$ and $Y \in \mathbb{R}^{m \times s}$, $\min\{n, m\} \geq s > r$.*
(a) Determine the QR decompositions $X = Q_X R_X$ and $Y = Q_Y R_Y$ ($Q_X \in \mathbb{R}^{n \times s}$, $Q_Y \in \mathbb{R}^{m \times s}$, $R_X, R_Y \in \mathbb{R}^{s \times s}$; cost: $4ns^2 + 4ms^2$).
(b) Compute the product $P := R_X R_Y^{\mathsf{T}} \in \mathbb{R}^{s \times s}$ (cost: $\frac{1}{3}s(2s^2 + 1)$).
(c) Determine $\mathrm{SVD}_r(P) = U \Sigma V^{\mathsf{T}}$ ($U, V \in \mathbb{R}^{s \times r}$ orthogonal, $\Sigma = \mathrm{diag}\{\sigma_1, \ldots, \sigma_r\}$; cost: $21s^3$, cf. [2,

[4] We use the term 'orthogonal matrix' also for rectangular matrices whose columns are pairwise orthonormal.

§5.4.5]).
Then, $\mathrm{SVD}_r(M) = Q_X U \Sigma V^\mathsf{T} Q_Y^\mathsf{T}$ *is the desired SVD truncation. To obtain again a representation of the form* $\mathrm{SVD}_r(M) = X' Y'^\mathsf{T}$*, set* $X' := Q_X U\Sigma$ *and* $Y' := Q_Y V$ *(cost:* $(1+2n+2m)rs - (n+m)\,r$*). The total cost is*

$$2\,(n+m)\,(2s+r)\,s + \frac{65}{3}s^3 + (s-m-n)\,r + \frac{1}{3}s. \tag{2.5a}$$

Sharper bounds can be obtained if we exploit the special block structures of the matrices X, Y, and of their QR factors.

Corollary 2.3 *Let* M *be the agglomeration of matrices of rank* $\le r$*. Depending of the agglomeration procedure, the following bounds hold (compare Footnote 3):*

$$\textit{truncation cost} \le \left\{ \begin{array}{ll} (20n+6m)\,r^2 + \frac{511}{3}r^3 + \left(\frac{2}{3} + 2r - n - m\right) r & \textit{for case (2.2a),} \\ (6n+20m)\,r^2 + \frac{511}{3}r^3 + \left(\frac{2}{3} + 2r - n - m\right) r & \textit{for case (2.2b),} \\ 20\,(n+m)\,r^2 + \frac{4064}{3}r^3 + \left(\frac{4}{3} + 2r - n - m\right) r & \textit{for case (2.2c).} \end{array} \right\} \tag{2.5b}$$

Proof. (i) Consider the case (2.2a), i.e., $M = [M_1\ M_2]$ with $M_i = X_i Y_i^\mathsf{T}$, $Y_i \in \mathbb{R}^{m_i \times r}$, $m = m_1 + m_2$. Extension by zero yields

$$[M_1\ O] = X_1 \begin{bmatrix} Y_1 \\ O \end{bmatrix}^\mathsf{T},\ [O\ M_2] = X_2 \begin{bmatrix} O \\ Y_2 \end{bmatrix}^\mathsf{T},\ M = XY^\mathsf{T} \quad \text{with } Y = \begin{bmatrix} Y_1 & O \\ O & Y_2 \end{bmatrix}.$$

Since $X = [X_1\ X_2]$ has no special structure, the QR decomposition of $X = Q_X R_X$ costs $4n\,(2r)^2$ operations. However, the QR decomposition of Y can be reduced to the QR decompositions $Y_i = Q_i R_i$ $(i = 1, 2)$, requiring $4mr^2$ operations. $Y = Q_Y R_Y$ holds with $Q_Y = \begin{bmatrix} Q_1 & O \\ O & Q_2 \end{bmatrix}$ and $R_Y = \begin{bmatrix} R_1 & O \\ O & R_2 \end{bmatrix}$. Hence, the cost of part (a) is $4n\,(2r)^2 + 4mr^2 = (16n + 4m)\,r^2$.

The multiplication $R_X R_Y^\mathsf{T}$ is cheaper than in the general case because of the upper right zero block and requires less than $\frac{7}{3}r^3 + \frac{2}{3}r$ operations.

The SVD cost is $21\,(2r)^3$. The multiplication $Q_X U\Sigma$ needs $(1+2n)2r^2 - nr$ operations, while $Q_Y V$ is cheaper because of the block diagonal form of Q_Y: $2mr^2 - mr$ operations. The sum of all terms yields the first line in (2.5b).

(ii) Case (2.2b) coincides with case (2.2a) applied to the transpose M^T. This proves the second line in (2.5b).

(iii) Case (2.2c) leads to $X = \begin{bmatrix} X_1 & O \\ O & X_2 \end{bmatrix}$, $X_i \in \mathbb{R}^{n_i \times 2r}$, $Y = \begin{bmatrix} Y_1 & O \\ O & Y_2 \end{bmatrix}$, $Y_i \in \mathbb{R}^{m_i \times 2r}$ with $n_1 + n_2 = n$, $m_1 + m_2 = m$. Correspondingly, the matrices Q_X, R_X are of the form $Q_X = \begin{bmatrix} Q_1 & O \\ O & Q_2 \end{bmatrix}$, $Q_i \in \mathbb{R}^{n_i \times 2r}$, $R_X = \begin{bmatrix} R_1 & O \\ O & R_2 \end{bmatrix}$, $R_i \in \mathbb{R}^{2r \times 2r}$, and similar for Q_Y, R_Y. The arising costs are:
(a) $4\,(n+m)\,(2r)^2$ for the QR decompositions,
(b) $2 * \frac{1}{3}(2r)(2(2r)^2 + 1)$ for the multiplication $R_X R_Y^\mathsf{T}$,
(c) $21(4r)^3$ for the singular value decomposition, $(4n+2)\,r^2 - nr$ for the multiplication in $Q_X U\Sigma$, and $(4r-1)\,mr$ for the multiplication $Q_Y V$. In total, we obtain the number in the third line of (2.5b). ∎

3 Recursive Truncation Algorithm

3.1 Partition Tree

The opposite of the agglomeration is the partition of the matrix (or of a matrix block) into submatrices. Formally, this recursive partition is described by a tree T. The root of the tree is the index pair $I \times J$. For the partition of $I \times J$ we consider three possibilities.

(A) Partition of J ($\square \to \square\!\square$). Let $J = \sigma_1 \dot\cup \sigma_2$ be a disjoint union. This induces the partition of $I \times J$ into the two blocks $b_i := I \times \sigma_i$ $(i = 1, 2)$.

(B) Partition of I (). Let $I = \tau_1 \dot\cup \tau_2$ be a disjoint union. This induces the partition of $I \times J$ into the two blocks $b_i := \tau_i \times J$ $(i = 1, 2)$.

(C) Partition of I and J (). Let $I = \tau_1 \dot\cup \tau_2$ and $J = \sigma_1 \dot\cup \sigma_2$ as above. This induces the partition of $I \times J$ into the four blocks $b_{ij} := \tau_i \times \sigma_j$ $(i, j = 1, 2)$.

We define the set of sons of $I \times J$ by $S(I \times J) := \{b_1, b_2\}$ (cases A and B) and $S(I \times J) := \{b_{11}, b_{12}, b_{21}, b_{22}\}$ (case C), respectively.

The obtained blocks can be split again into subblocks according to the rules A–C. If a block b is not partitioned further (i.e., $S(b) = \emptyset$), b is called a leaf of the tree. The set of all leaves yields the final block partition

$$P := \{b \in T : S(b) = \emptyset\}. \tag{3.1}$$

Remark 3.1 *P satisfies $\bigcup_{b\in P} b = I \times J$ (disjoint union).*

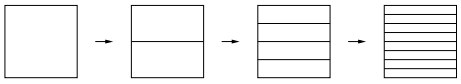
Remark 3.2 *If we want to truncate the matrix $M \in \mathbb{R}^{I\times J}$ to rank r, blocks $b = \tau \times \sigma$ with $\min\{\#\tau, \#\sigma\} \le r$ should not be subdivided.*

Each node $b \in T$ of the tree has a level number. Its recursive definition is $\operatorname{level}(I \times J) := 0$ and, $\operatorname{level}(b') = \operatorname{level}(b) + 1$ for $b' \in S(b)$. The value $\operatorname{level}(b)$ can be interpreted as the length of the path from the root to b. We define $\operatorname{depth}(T) := \max_{b\in T} \operatorname{level}(b)$ and

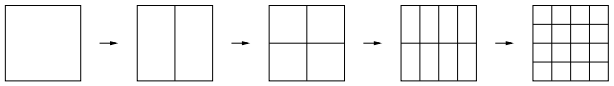
$$T^{(\ell)} := \{b \in T : \operatorname{level}(b) = \ell\} \qquad \text{for } 0 \le \ell \le \operatorname{depth}(T).$$

The following examples will be used later as model cases.

Example 3.3 *Let $M \in \mathbb{R}^{I\times J}$ with $\#I = 2^L r$, where $r, L \in \mathbb{N}$. Apply the partition rule B L times, i.e., any $b := \tau \times J \in T$ is split into $b_1 := \tau_1 \times J$ and $b_2 := \tau_2 \times J$. The size of the index subset τ in $b = \tau \times J \in T^{(\ell)}$ is $\#\tau = 2^{L-\ell} r$ (regular splitting). L is the depth of the tree; i.e., the leaves are blocks $b = \tau \times J$ with $\#\tau = r$ (cf. Remark 3.2). The following picture illustrates the case of $L = 3$:*

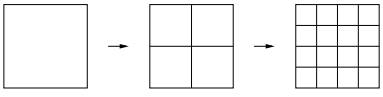


Example 3.4 *Let $M \in \mathbb{R}^{I\times J}$ with $\#I = \#J = 2^p r$, where $r, p \in \mathbb{N}$. Apply $L := 2p$ partition steps according to the rules A, B, A, B,..., A, B in this order. The size of the index subsets τ, σ in $b = \tau \times \sigma \in T^{(\ell)}$ is $\#\tau = \#\sigma = 2^{p-\ell/2} r$ for even ℓ, while it is $\#\tau = 2^{p-(\ell-1)/2} r$ and $\#\sigma = 2^{p-(\ell+1)/2} r$ for odd ℓ. L is the depth of the tree; i.e., the leaves are blocks $b = \tau \times \sigma$ with $\#\tau = \#\sigma = r$ (cf. Remark 3.2). The following picture illustrates the case of $L = 4$:*



The same partition as in Example 3.4 is also obtained by the next tree.

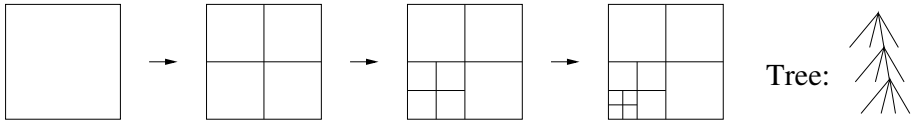
Example 3.5 *Let $M \in \mathbb{R}^{I\times J}$ with $\#I = \#J = 2^L r$, where $r, L \in \mathbb{N}$. Apply the partition rule C L times. The size of the index subsets τ, σ in $b = \tau \times \sigma \in T^{(\ell)}$ is $\#\tau = \#\sigma = 2^{L-\ell} r$. L is the depth of the tree; i.e., the leaves are blocks $b = \tau \times \sigma$ with $\#\tau = \#\sigma = r$. The following picture illustrates the case of $L = 2$:*



The previous examples lead to regular balanced trees. Note that $L = \mathcal{O}(\log \#I/r)$.

In the context of the hierarchical matrix technique, for instance the following partition is of interest.

Example 3.6 *Let $M \in \mathbb{R}^{I\times J}$ with $\#I = \#J = 2^L n_0$, where $n_0, L \in \mathbb{N}$. Apply the partition rule C L times, but only to one son of $S(b)$ as indicated in the following illustration:*



L is the depth of the tree. Now the partition P consists of blocks $b \in T^{(\ell)} \cap P$ of different sizes $2^{L-\ell} n_0 \times 2^{L-\ell} n_0$ ($\ell = 1, \ldots, L$). It is assumed that all submatrices $M|_b$ ($b \in P$) are represented in the form $M|_b = X_b Y_b^{\mathsf{T}}$.

3.2 Algorithm

Assume that a matrix $M \in \mathbb{R}^{I\times J}$ and a tree T with the partition P (cf. (3.1)) is given. $r \in \mathbb{N}$ is the desired rank for the truncation. The next remark defines A_b for the blocks $b \in P$.

Remark 3.7 *Concerning the submatrices $M|_b$ for $b = \tau \times \sigma \in P$ we can distinguish three cases:*

(i) $M|_b$ is already given by the product $M|_b = X_b Y_b^{\mathsf{T}}$ with $X_b \in \mathbb{R}^{\tau\times\rho}$ and $Y_b \in \mathbb{R}^{\sigma\times\rho}$ for $\rho = \{1, \ldots, r\}$ (i.e., $M|_b \in \mathcal{R}(r, \tau, \rho)$). Set $A_b := M|_b$.

(ii) $M|_b$ is given as full matrix with $\min\{\#\tau, \#\sigma\} \le r$. Then $M|_b \in \mathcal{R}(r, \tau, \rho)$ holds and the representation $M|_b = X_b Y_b^{\mathsf{T}}$ from case (1) can be determined by a QR decomposition (cost $\le 4r^2 \max\{\#\tau, \#\sigma\}$). Set $A_b := M|_b$.

(iii) $M|_b$ is given as full matrix and $\min\{\#\tau, \#\sigma\} > r$. Then set $A_b := \mathrm{SVD}_r(M|_b)$. The SVD cost is $21 \min\{\#\tau, \#\sigma\}^2 \max\{\#\tau, \#\sigma\}$. The product $A_b = X_b Y_b^{\mathsf{T}}$ is a side product of the singular value decomposition.

As a result of this preparation, we have a representation $A_b = X_b Y_b^{\mathsf{T}} \in \mathcal{R}(r, \tau, \rho)$ approximating $M|_b$ for all $b \in P$. Now the following algorithm is applied.

$$\begin{aligned}
&\textbf{for } \ell := \mathrm{depth}(T) - 1 \textbf{ downto } 0 \textbf{ do for all } b \in T^{(\ell)} \textbf{ do}\\
&\textbf{if } b \notin P \textbf{ then}\\
&\textbf{begin } \hat{A}_b := \mathrm{Aggl}\{A_{b'} : b' \in S(b)\};\\
&\qquad A_b := \mathrm{SVD}_r(\hat{A}_b)\\
&\textbf{end};
\end{aligned} \tag{3.2}$$

The truncation $A_b := \mathrm{SVD}_r(\hat{A}_b)$ includes the computation of the representation $A_b = X_b Y_b^{\mathsf{T}}$. The algorithm terminates with $\ell = 0$. The only block in $T^{(0)}$ is $b = I \times J$, i.e., the truncation

$$A = A_{I\times J} \tag{3.3a}$$

of M is computed. In the following, we shall discuss the difference between A and the optimal SVD truncation

$$B := \mathrm{SVD}_r(M). \tag{3.3b}$$

Note that the truncation of the submatrices always uses the destination rank r. Using smaller ranks for truncating smaller-sized matrices is not appropriate, since then we cannot prove the desired estimate (1.1) with a bound C' independent of the matrix. A counter-example is the matrix $M \in \mathbb{R}^{I\times J}$ with support in b, i.e., $M = (M|_b)^{I\times J}$.

In §5 we shall prove the estimate

$$\|A - B\| \le q^{1+L} \|M - B\| \qquad \text{for } A, B \text{ in (3.3a,b)}, \tag{3.4a}$$

where $q := \frac{1+\sqrt{5}}{2}$ is derived from Theorem 4.5, while $L = \mathrm{depth}(T)$ is the depth of the underlying partition tree. This estimate corresponds to (1.1) with $C = q^{1+L}$. More detailed results for the model problems are presented in §5.2.

For the particular unidirectional partition defined in §7 the better estimate

$$\|A - B\| \le \left(1 + \sqrt{L+1}\right) \|M - B\| \tag{3.4b}$$

will be stated in Theorem 7.2.

We do not claim that the inequalities (3.4a,b) are sharp. Optimal bounds are unknown. Next we give some examples. Let I_r be the unit matrix in $\mathbb{R}^{r\times r}$ and define $E_n \in \mathbb{R}^{n\times n}$ by the entries $E_{n,ij} := 1$.

Example 3.8 *(a) A possible realisation of algorithm (3.2) with $r = 1$ for the tensor product $M := I_2 \otimes E_2 \in \mathbb{R}^{4\times 4}$ is*

$$M = \left[\begin{array}{cc|cc} 1&0&1&0\\ 0&1&0&1\\ \hline 1&0&1&0\\ 0&1&0&1 \end{array}\right] \mapsto \left[\begin{array}{cc|cc} 1&0&0&0\\ 0&0&0&1\\ \hline 0&0&1&0\\ 0&1&0&0 \end{array}\right] \mapsto \left[\begin{array}{cc|cc} 1&0&0&0\\ 0&0&0&1\\ \hline 0&0&0&0\\ 0&0&0&0 \end{array}\right] \mapsto \left[\begin{array}{cccc} 1&0&0&0\\ 0&0&0&0\\ 0&0&0&0\\ 0&0&0&0 \end{array}\right] = A.$$

Note that the local SVD truncations are non-unique. Therefore one can construct A such that the entry 1 appears at any *position (i, j) with $i + j$ even. Embedding M into a larger matrix and choosing a similar truncation strategy, $M := I_2 \otimes E_n \in \mathbb{R}^{2n\times 2n}$ may yield the result A with $A_{11} = 1$ and $A_{ij} = 0$ otherwise. Together with $B := \begin{bmatrix} 0&0\\ 0&1 \end{bmatrix} \otimes E_n$, the norms $\|A-B\|^2 = 1+n^2$, $\|M - A\|^2 = 2n^2-1$, and $\|M - B\|^2 = n^2$ show that $\|A - B\|/\|M - B\| > 1$ and $\|M - A\|/\|M - B\| = 2 - n^{-2}$.*

(b) Similar (but more involved) constructions for larger r can be applied to $M := I_{r+1} \otimes E_n$ and lead to $\|A - B\|^2 = C_r + rn^2$, $\|M - A\|^2 = (r + 1)n^2 - C_r$, and $\|M - B\|^2 = n^2$. Therefore, the ratio $\|M - A\|/\|M - B\|$ approaches $\sqrt{r+1}$. This shows that (3.4a,b) cannot hold with a constant independent of r.

4 A Perturbation Estimate

An essential tool for the later analysis of algorithm (3.2) is Theorem 4.5. A weaker result stated in Grasedyck–Hackbusch [3] is based on the following trivial estimate (4.3), which can be regarded as a perturbation analysis. In the case of $M = A$ with $A \in \mathcal{R}(r, I, J)$, the SVD result $B = \mathrm{SVD}_r(M)$ coincides with A, i.e., $\|A - B\| = 0$. Now we perturb A by Δ and ask for the norm of $B - A = \mathrm{SVD}_r(A - \Delta) - A$. The following assumption $\mu > r$ avoids the trivial case of $M = A$.

Lemma 4.1 *Assume $M \in \mathbb{R}^{I\times J}$ with $\mu := \mathrm{rank}(M) > r$. Let*

$$M = A - \Delta \qquad \text{with } A \in \mathcal{R}(r, I, J) \tag{4.1a}$$

be an arbitrary splitting, while the optimal SVD splitting is given by

$$M = B + E \qquad \text{with } B = \mathrm{SVD}_r(M). \tag{4.1b}$$

(a) Then the remainder E in (4.1b) can be estimated by

$$\|E\| \le \|\Delta\|. \tag{4.2}$$

(b) Furthermore, the matrices A and B differ by

$$\|A - B\| = \|\Delta + E\| \le 2\,\|\Delta\|. \tag{4.3}$$

Proof. Part (a) follows from the best approximation property of the SVD truncation. Part (b) is an easy consequence of (a). ∎

Estimate (4.3) is too pessimistic as proved next. We introduce the following supremum:

$$q_{I\times J} := q := \sup_{\substack{M=A-\Delta=B+E\in\mathbb{R}^{I\times J} \text{ according}\\ \text{to (4.1a,b) with } \Delta\neq 0}} \frac{\|A - B\|}{\|\Delta\|}. \tag{4.4}$$

Note that the supremum is taken over all pairs $M, A \in \mathbb{R}^{I\times J}$, while B and E result from (4.1b). This states that the estimate $\|A - B\| \le q\,\|\Delta\|$ holds for all matrices $M \in \mathbb{R}^{I\times J}$ independently of their (spectral) properties.

The later results will show that $q_{I\times J}$ does not depend on $I \times J$ provided that $\#I, \#J \ge 2$. Therefore, we write q instead of $q_{I\times J}$. Inequality (4.3) states that $q \le 2$. Next we prove that the supremum is a maximum.

Lemma 4.2 *There are matrices M, A, B satisfying (4.1a,b) and $\frac{\|A-B\|}{\|\Delta\|} = q$ with q defined in (4.4).*

Proof. Let $I \times J$ be fixed. First we assume that $\|A-B\|/\|\Delta\| = q - \varepsilon$ holds for all matrices $M, A \in \mathbb{R}^{I\times J}$ with $\varepsilon = \varepsilon(M,A) > 0$. Since inequality (4.3) is scaling invariant and Δ does not vanish, we restrict ourselves to matrix pairs with $\|\Delta\| = 1$. There is a sequence $M_k, A_k \in \mathbb{R}^{I\times J}$ with $\varepsilon(M_k, A_k) \to 0$. By compactness, there is a subsequence such that $\Delta_k := A_k - M_k \to \Delta$ with $\|\Delta\| = 1$. The matrices M_k, A_k may diverge, but their divergent parts must coincide. Consider $B_k = \mathrm{SVD}_r(M_k) = \sum_{\nu=1}^{r} \sigma_{\nu,k} u_{\nu,k} v_{\nu,k}^{\mathsf{T}}$. The diverging part $B_k^{\mathrm{div}} := \sum^{*} \sigma_{\nu,k} u_{\nu,k} v_{\nu,k}^{\mathsf{T}}$ sums over those ν with $\sup_k \sigma_{\nu,k} = \infty$. Set $D_k := \sum^{*} (\sigma_{\nu,k} - 2)\, u_{\nu,k} v_{\nu,k}^{\mathsf{T}}$ and define $\hat{M}_k := M_k - D_k$ and $\hat{A}_k := A_k - D_k$. By construction, $\hat{M}_k$ and $\hat{A}_k$ are bounded and $\hat{B}_k := \mathrm{SVD}_r(\hat{M}_k) = \mathrm{SVD}_r(M_k) - D_k$, i.e., E_k and Δ_k are not changed by this construction (note that $\|\Delta_k\| = 1$ implies $\sigma_{\nu,k} \le 1$ for all singular values of $E_k := M_k - B_k$). Choosing a second subsequence, the limits M, A, B, E of $\hat{M}_k, \hat{A}_k, \hat{B}_k, E_k$ exist and prove $\|\Delta + E\| = q\,\|\Delta\|$ for the pair M, A. ∎

Corollary 4.3 *Estimate (4.3) is not sharp; i.e., $q < 2$.*

Proof. Assuming $q = 2$, Lemma 4.2 states the existence of a pair (M, A) such that $\|A-B\| = 2\,\|\Delta\|$ and $\Delta \neq 0$. Because (4.2), $\|\Delta + E\| = \|\Delta\| + \|E\|$ and $\|E\| = \|\Delta\|$ hold. The first equality implies that Δ and E are linearly dependent, since $\|\cdot\|$ is a Hilbert norm. The second equation shows that $\Delta = E$. We conclude that $M = A - E = B + E$ and $A = B + 2E$ with $E \neq 0$. The structures of B and E in (4.1b) imply that $\mathrm{rank}(A) = \mathrm{rank}(B + 2E) = \mu > r$, in contradiction to $A \in \mathcal{R}(r, I, J)$. ∎

On the other hand, q is larger than one, as proved by the following example.

Example 4.4 *Let $M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $r = 1$. Possible realisations of (4.1a,b) are given by*

$$A = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1+\sqrt{5}}{2} \end{bmatrix},\ \Delta = \begin{bmatrix} -1 & 0 \\ 0 & \frac{-1+\sqrt{5}}{2} \end{bmatrix},\ B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},\ E = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then $\|A-B\| = \left\| \begin{bmatrix} 1 & 0 \\ 0 & -\frac{1+\sqrt{5}}{2} \end{bmatrix} \right\| = \sqrt{\frac{5+\sqrt{5}}{2}}$ and $\|\Delta\| = \sqrt{\frac{5-\sqrt{5}}{2}}$ prove $q \ge \frac{1+\sqrt{5}}{2} = 1.618\ldots$.

In this example, the SVD truncation B is not unique. However, uniqueness holds for $M = \begin{bmatrix} 1 & 0 \\ 0 & 1-\varepsilon \end{bmatrix}$, $A = \begin{bmatrix} 0 & 0 \\ 0 & (1+\sqrt{5})/2 - \varepsilon \end{bmatrix}$ with $0 < \varepsilon < 1$. Then $q \ge (1+\sqrt{5})/2 - \mathcal{O}(\varepsilon)$ and the definition $q := \sup\{\ldots\}$ in (4.4) imply again that $q \ge (1+\sqrt{5})/2$.

The above example turns out to be the worst case. The next theorem states that Lemma 4.1 holds with $\frac{1+\sqrt{5}}{2}\,\|\Delta\|$ in (4.3) instead of $2\,\|\Delta\|$ and that this estimate is sharp.

Theorem 4.5 *The quantity in (4.4) is $q = \frac{1+\sqrt{5}}{2}$.*

Proof. (i) In $\mathbb{R}^I$ ($\mathbb{R}^J$) we use the orthogonal basis obtained by extension of $\{u_k\}$ ($\{v_k\}$) from (2.4). Then the matrix representations of M, B, and E with respect to these bases are diagonal: $M = \mathrm{diag}\{\sigma_1, \sigma_2, \ldots\}$, $B = \mathrm{diag}\{\sigma_1, \ldots, \sigma_r, 0, \ldots\}$, $E = \mathrm{diag}\{0, \ldots, 0, \sigma_{r+1}, \ldots\}$. Split A into the diagonal part $A_0 := \mathrm{diag}\{A_{11}, A_{22}, \ldots\}$ and $A_\perp := A - A_0$. Note that $A_\perp$ is perpendicular to any diagonal matrix, in particular to M, B, and E. Let $\Delta = \Delta_0 + \Delta_\perp$ be the analogous splitting. One concludes that $A_\perp = \Delta_\perp$ and that

$$\left(\frac{\|A-B\|}{\|\Delta\|}\right)^2 = \frac{\|A_0 - B\|^2 + \|A_\perp\|^2}{\|\Delta_0\|^2 + \|\Delta_\perp\|^2} = \frac{\|A_0 - B\|^2 + \|\Delta_\perp\|^2}{\|\Delta_0\|^2 + \|\Delta_\perp\|^2} \le \left(\max\left\{\frac{\|A_0 - B\|}{\|\Delta_0\|}, 1\right\}\right)^2.$$

Since we want to maximise $\|A-B\|/\|\Delta\|$, we consider only the case $\|A_0 - B\|/\|\Delta_0\| > 1$. The existence of such matrices is given by Example 4.4. Then the unique maximum is attained for $\|\Delta_\perp\| = 0$, i.e., the critical choice of A has to be diagonal.

(ii) A diagonal matrix A with the additional property $A \in \mathcal{R}(r, I, J)$ contains at most r diagonal entries; i.e., there is an index subset I' with[5]

$$\#I' = r, \qquad A = \sum_{i \in I'} \lambda_i u_i v_i^{\mathsf{T}}.$$

Split I' into $I'_1 := I' \cap \{1, \ldots, r\}$ and $I'_2 := I' \cap \{r+1, \ldots, \mu\}$. Similarly, the complement $I'' := \{1, \ldots, \mu\} \backslash I'$ is split into I''_1 and I''_2. Hence, the unions in $I'_1 \cup I'_2 = I'$, $I''_1 \cup I''_2 = I''$, $I'_1 \cup I''_1 = \{1, \ldots, r\}$, $I'_2 \cup I''_2 = \{r+1, \ldots, \mu\}$ are disjoint.

A corresponds to the perturbation $\Delta = A - M = \sum_{i \in I'} (\lambda_i - \sigma_i)\, u_i v_i^{\mathsf{T}} - \sum_{i \in I''} \sigma_i u_i v_i^{\mathsf{T}}$. The quantity considered above is equal to

$$\begin{aligned}\left(\frac{\|A-B\|}{\|\Delta\|}\right)^2 &= \frac{\sum_{i \in I'_1} (\lambda_i - \sigma_i)^2 + \sum_{i \in I''_1} \sigma_i^2 + \sum_{i \in I'_2} \lambda_i^2}{\sum_{i \in I'} (\lambda_i - \sigma_i)^2 + \sum_{i \in I''} \sigma_i^2} \\ &= \frac{\left[\sum_{i \in I'_1} (\sigma_i - \lambda_i)^2 + \sum_{i \in I''_1} \sigma_i^2\right] + \sum_{i \in I'_2} \lambda_i^2}{\left[\sum_{i \in I'_1} (\lambda_i - \sigma_i)^2 + \sum_{i \in I''_1} \sigma_i^2\right] + \sum_{i \in I'_2} (\lambda_i - \sigma_i)^2 + \sum_{i \in I''_2} \sigma_i^2}.\end{aligned}$$

The expression is strictly maximised by shifting all indices of I'_1 into the part I'_2 and choosing λ_i ($i \in I'_2$) such that $(\lambda_i - \sigma_i)^2 < \sigma_i^2$. As a result $I'_1 = \emptyset$ holds. Then the values σ_i^2 for $i \in I''_2$ only appear in the term $\sum_{i \in I''_2} \sigma_i^2$ of the denominator. Hence the choice $\sigma_i = 0$ for $i \in I''_2$ increases the expression again, which is now of the simpler form

$$\left(\frac{\|A-B\|}{\|\Delta\|}\right)^2 = \frac{\sum_{i \in I''_1} \sigma_i^2 + \sum_{i \in I'_2} \lambda_i^2}{\sum_{i \in I''_1} \sigma_i^2 + \sum_{i \in I'_2} (\lambda_i - \sigma_i)^2} \quad \text{with } \#I''_1 = \#I'_2 = r.$$

The next maximisation step concerns the singular values σ_i for $i \in I''_1 = \{1, \ldots, r\}$, while the values σ_i for $i \in I'_2$ are fixed. Set $\sigma := \sigma_{r+1}$. The expression $\|A-B\| / \|\Delta\|$ increases if the sum $\sum_{i \in I''_1} \sigma_i^2$ decreases. Its minimal value is $r\sigma$ for the constant choice $\sigma_k = \sigma$ for $1 \le k \le r$, since the σ_i are ordered by size. For $i \in I'_2$ replace σ_i by $\sigma \ge \sigma_i$ and λ_i by $\lambda_i + \sigma - \sigma_i$. This increases the value again and yields

$$\left(\frac{\|A-B\|}{\|\Delta\|}\right)^2 = \frac{r\sigma + \sum_{i=r+1}^{2r} \lambda_i^2}{r\sigma + \sum_{i=r+1}^{2r} (\lambda_i - \sigma)^2}.$$

Optimisation with respect to λ_i yields $\lambda := \frac{1+\sqrt{5}}{2}\sigma$ for all i. Insertion yields $\frac{\|A-B\|}{\|\Delta\|} \le \frac{1+\sqrt{5}}{2}$. This shows that $q = \frac{1+\sqrt{5}}{2}$ is the maximum of $\|A-B\| / \|\Delta\|$. ∎

Since the previous steps are strict maximisations, the matrices M and A attaining this bound must be of the form

$$M = \sigma\, I_{2r}\ ,\ A = \begin{bmatrix} 0 & 0 \\ 0 & \lambda I_r \end{bmatrix}\ ,\ \Delta = \begin{bmatrix} -\sigma I_r & 0 \\ 0 & (\lambda - \sigma) I_r \end{bmatrix}\ ,\ B = \begin{bmatrix} \sigma I_r & 0 \\ 0 & 0 \end{bmatrix}\ , \tag{4.5}$$

with $\sigma \neq 0$ and $\lambda = \frac{1+\sqrt{5}}{2}\sigma$, where I_r denotes the identity matrix of size $r \times r$.

Let $A \in \mathcal{R}(r, I, J)$ be any rank-r approximation of M. A simple modification yields a possibly better approximation: $M = \hat{A} - \hat{\Delta}$ with $\hat{A} := \omega A$, $\omega := \langle A, M \rangle / \|A\|^2$, and $\hat{\Delta} := \omega A - M$. The choice of ω ensures that $\hat{A} \perp \hat{\Delta}$. $\hat{A}$ can be considered as ΠM, where Π is the orthogonal projection onto $\mathrm{span}\{A\}$.

In Proposition 5.1 we shall show for particular cases that $A \perp \Delta$ holds for the unmodified matrices.

Remark 4.6 *Under the additional condition $A \perp \Delta$, the maximal value q in (4.4) is $q = \sqrt{2}$. This maximum is taken for the matrices from (4.5) with λ replaced by σ.*

5 First Approach

The following estimations can be used for all block decompositions, while the second approach from Section 7 applies only to the unidirectional splitting as, e.g., in Example 3.3.

[5] The possible case of $\mathrm{rank}(A) < r$ is included since we may choose $\lambda_i = 0$.

5.1 Comparison with $B|_b$

For $M \in \mathbb{R}^{I\times J}$ we consider the optimal SVD splitting

$$M = B - \Delta \qquad \text{with } B := \mathrm{SVD}_r(M) \tag{5.1}$$

(for theoretical purpose only). The following statements about the restrictions $B|_b$ and $\Delta|_b$ to the blocks of the tree T are trivial, but important:

$$\|\Delta\|^2 = \sum_{b\in P} \|\Delta|_b\|^2\,, \qquad \|\Delta|_b\|^2 = \sum_{b'\in S(b)} \|\Delta|_{b'}\|^2 \text{ for } b \in T\backslash P. \tag{5.2}$$

The property $B \in \mathcal{R}(r, I, J)$ implies

$$B|_b \in \mathcal{R}(r, \tau, \sigma) \qquad \text{for } b = \tau\times\sigma \in T. \tag{5.3}$$

The agglomeration is trivial:

$$B|_b = \mathrm{Aggl}\{B|_{b'} : b' \in S(b)\}, \qquad \Delta|_b = \mathrm{Aggl}\{\Delta|_{b'} : b' \in S(b)\}. \tag{5.4}$$

Although B is the optimal rank-r matrix, this statement does not hold for the restriction $B|_b$ ($b \neq I\times J$). Its remainder is $\Delta|_b$:

$$M|_b = B|_b - \Delta|_b. \tag{5.5}$$

Nevertheless, for the analysis, we consider the distance of A_b from $B|_b$:

$$A_b = B|_b - F_b, \tag{5.6}$$

and try to estimate $F_b := B|_b - A_b$. We recall that A_b ($b \in P$) are the matrices determined by algorithm (3.2).

Proposition 5.1 *The definition of B and Δ implies $B \perp \Delta$. In general, $B|_b \perp \Delta|_b$ does not hold for the blocks of $T\backslash\{I\times J\}$. However, if b is of the form $\tau\times J$ or $I\times\sigma$, $B|_b \perp \Delta|_b$ is valid.*

Proof. The block $b := I\times\{j\}$ ($j \in J$) is a column. One easily verifies that the columns $B|_b$, $\Delta|_b$ are perpendicular. This proves $B|_b \perp \Delta|_b$ for $b = I\times\sigma$. Similar for $b = \tau\times J$. ∎

For leaves $b \in P$ we have $A_b := \mathrm{SVD}_r(M|_b)$ (if $M|_b \in \mathcal{R}(r, b)$, $A_b = M|_b$ holds). Theorem 4.5 proves that

$$\|F_b\| = \|A_b - B|_b\| \le q\,\|\Delta|_b\| \qquad \text{for } b \in P \quad (q = \tfrac{1+\sqrt{5}}{2}) \tag{5.7}$$

If $b \notin P$, the algorithm computes $\hat{A}_b := \mathrm{Aggl}\{A_{b'} : b' \in S(b)\}$ and $A_b := \mathrm{SVD}_r(\hat{A}_b)$. Statement (5.4) implies that

$$\hat{A}_b = B|_b - \hat{E}_b \quad \text{with } \hat{E}_b := \mathrm{Aggl}\{F_{b'} : b' \in S(b)\};$$

therefore,

$$\|\hat{E}_b\|^2 = \sum_{b'\in S(b)} \|F_{b'}\|^2\,.$$

Let E_b be the SVD remainder in

$$\hat{A}_b = A_b + E_b \quad \text{with } A_b := \mathrm{SVD}_r(\hat{A}_b). \tag{5.8}$$

This implies that (5.6) holds with

$$F_b := \hat{E}_b + E_b.$$

Again, Theorem 4.5 states that

$$\|F_b\| = \|B|_b - A_b\| \le q\,\|\hat{E}_b\| = q\sqrt{\sum_{b'\in S(b)} \|F_{b'}\|^2}$$

with $q = \frac{1+\sqrt{5}}{2}$. This proves the following lemma.

Lemma 5.2 *The norm of F_b satisfies the recursive inequality*

$$\|F_b\|^2 \le q^2 \sum_{b' \in S(b)} \|F_{b'}\|^2 \qquad \text{for } b \in T\backslash P, \tag{5.9}$$

while $\|F_b\| \le q\|\Delta|_b\|$ for $b \in P$. The final result $A := A_{I\times J}$ satisfies $\|A - B\| = \|F_{I\times J}\|$ for $B = \mathrm{SVD}_r(M)$.

Corollary 5.3 *(a) Let χ be the matrix defined by $\chi_{ij} = q^{\ell+1}$ with $\ell := \mathrm{level}(b)$ and $b \in P$ such that $(i,j) \in b$. Then*[6]

$$\|A - B\| \le \|\chi \circ \Delta\|$$

holds with Δ from (5.1). An equivalent estimate is $\|A - B\|^2 \le \sum_{b\in P} q^{2(1+\mathrm{level}(b))}\|\Delta|_b\|^2$.

(b) An upper bound is $\|A - B\| \le q^{1+L}\|\Delta\|$ with $L = \mathrm{depth}(T)$.

Proof. (a) The inductive hypothesis is $\|F_b\|^2 \le \sum_{b'\in P, b'\subset b} q^{2(1+\mathrm{level}(b)-\mathrm{level}(b'))}\|\Delta|_{b'}\|^2$. The induction starts at the leaves. Here, the statement follows from (5.7), since the only block $b' \in P$ with $b' \subset b$ is $b' = b$.

Let $b \in T\backslash P$ and assume that the hypothesis holds for the sons $b' \in S(b)$. The statement for $\|F_b\|^2$ follows from (5.9) and $\mathrm{level}(b') = \mathrm{level}(b) + 1$.

(b) For part (b) use $\mathrm{level}(b) \le L$ and $\sum_{b\in P} \|\Delta|_b\|^2 = \|\Delta\|^2$. ∎

5.2 Application to the Model Problems

We discuss the previous estimate for Examples 3.3 to 3.6. Furthermore, we describe the computational work of the recursive truncation.

Proposition 5.4 *The recursive truncation (3.2) applied to Example 3.3 yields a result $A := A_{I\times J}$ satisfying*

$$\|A - B\| \le q^{L+1}\,\|\Delta\| \tag{5.10}$$

with $L = \mathrm{depth}(T) = \log_2(\#I/r)$ and B and Δ from (5.1).

Remark 5.5 *The preparation at the leaves together with the algorithm (3.2) requires a computational work of*

$$(24r-1)\,nm + r^2\left((6L + \tfrac{511}{3})n - 20m - \tfrac{511}{3}r\right) + \mathcal{O}(mr+n), \text{ where } n := \#I,\ m := \#J.$$

Proof. The first part follows from $\mathrm{level}(b) = L$ for all $b \in P$.

According to Remark (ii), each leaf requires $4r^2\#J$ operations. Since there are 2^L leaves, the total work is $4r\#I\#J$ (note that $\#I = 2^L r$).

The agglomeration and singular value decomposition yielding A_b for $b \in T^{(\ell)}$ cost $\left(6\cdot 2^{L-\ell}r + 20\#J\right)r^2 + \frac{511}{3}r^3 + \left(\frac{2}{3} + 2r - 2^{L-\ell}r - \#J\right)r$ operations (see Corollary 2.3, case (2.2b) with $n = 2^{L-\ell}r$ and $m = \#J$). The number of blocks in $T^{(\ell)}$ is 2^ℓ. Summation of $0 \le \ell \le L-1$ yields $20rnm - 20mr^2 + 6r^2Ln + \frac{511}{3}r^2(n-r) + \frac{2}{3}n - \frac{2}{3}r + 2rn + mr - 2r^2 - nm - Lrn$. ∎

While the inequality in Theorem 4.5 is sharp, we do not claim that (5.10) is sharp. To prove that the bound in (5.10) can be obtained, one has to find an example where in each step of the recursive truncation the matrices are of the form (4.5).

Proposition 5.6 *The recursive truncation (3.2) applied to Example 3.4 yields a result $A := A_{I\times J}$ satisfying (5.10) with $L = \mathrm{depth}(T) = \log_2(\#I\#J/r^2)$.*

Remark 5.7 *The preparation at the leaves together with the algorithm (3.2) requires a computational work of*

$$\left(232 + \tfrac{1}{3}\right)rn^2 + \left(\tfrac{2}{3r} - 5\right)n^2, \text{ where } n := \#I = \#J = 2^p r.$$

[6] $\chi \circ \delta$ is the Hadamard product (entry-wise product) defined by $(\chi\circ\delta)_{ij} = \chi_{ij}\delta_{ij}$.

Proof. According to Remark (ii), the cost at the leaves is $4r^3$. Multiplication by the number 2^L of leaves yields $4r\#I\#J$.

Let ℓ be even. The agglomeration and singular value decomposition for $b \in T^{(\ell)}$ require $26 \cdot 2^{(L-\ell)/2} r^3 + \frac{511}{3} r^3 + \left(\frac{2}{3} + 2r - 2^{1+(L-\ell)/2} r\right) r$ operations (see Corollary 2.3, case (2.2a) with $n = m = 2^{(L-\ell)/2} r$). Note that $\#T^{(\ell)} = 2^\ell$. Summation over all even levels $\ell = 0, 2, \ldots, L-2$ yields $\frac{745}{9} n^2 r - \frac{10}{3} n^2 - 26nr^2 + 2nr + \frac{2}{9} n^2/r - \frac{511}{9} r^3 - \frac{2}{3} r^2 - \frac{2}{9} r$ with $n := \#I = \#J$.

For odd ℓ, the cost connected to $b \in T^{(\ell)}$ is $16 \cdot 2^{(L-\ell+1)/2} r^3 + \frac{511}{3} r^3 + \left(\frac{2}{3} + 2r - 3 \cdot 2^{(L-\ell-1)/2} r\right) r$ (see Corollary 2.3, case (2.2b) with $n = 2^{(L-\ell+1)/2} r$, $m = 2^{(L-\ell-1)/2} r$). Again $\#T^{(\ell)} = 2^\ell$ holds. Summation over all odd levels $\ell = 1, 3, \ldots, L-1$ yields $\frac{1310}{9} rn^2 - \frac{5}{3} n^2 + \frac{4}{9} n^2/r - 32r^2 n + 3rn - \frac{1022}{9} r^3 - \frac{4}{3} r^2 - \frac{4}{9} r$.

Together with the first part, we obtain the sum $\frac{697}{3} n^2 r - 58nr^2 - \frac{2}{3} r + \frac{2}{3} \frac{n^2}{r} + 5nr - 5n^2 - 2r^2 - \frac{511}{3} r^3$. ∎

The larger factor in front of rn^2 is caused by the larger number of blocks in T: $\#P = 2\,(n/r)^2 - 1$ for Example 3.4 compared with $\#P = 2\,(n/r) - 1$ for Example 3.3. For each block a singular value decomposition is required.

Proposition 5.8 *The recursive truncation (3.2) applied to Example 3.5 yields a result $A := A_{I \times J}$ satisfying (5.10) with $L = \operatorname{depth}(T) = \log_2(n/r)$, $n = \#I = \#J$.*

Remark 5.9 *The preparation at the leaves together with algorithm (3.2) requires a computational work of less than*

$$\left(495 + \tfrac{5}{9}\right) rn^2$$

operations.

Proof. Again, the cost at the leaves is $4r^3$. Multiplication with the number 4^L of leaves yields $4rn^2$.

The agglomeration and singular value decomposition for $b \in T^{(\ell)}$ require $40 \cdot 2^{L-\ell} r^3 + \frac{4064}{3} r^3 + \left(\frac{4}{3} + 2r - 2^{L-\ell+1} r\right) r$ operations (see Corollary 2.3, case (2.2c) with $n = m = 2^{L-\ell} r$). Note that $\#T^{(\ell)} = 4^\ell$. Summation over all even levels $0 \le \ell \le L-1$ yields $\left(40 + \frac{4064}{9}\right) rn^2 - 2n^2 - \frac{4}{9} r + \frac{4}{9} n^2/r + 2rn - 40r^2 n + \frac{2}{3} n^2 - \frac{2}{3} r^2 - \frac{4064}{9} r^3$. ∎

Here, the larger factor of rn^2 is caused by the fact that the agglomeration of four blocks quadruples the rank.

Example 3.6 is quite different since much less blocks are involved.

Proposition 5.10 *The recursive truncation (3.2) applied to Example 3.6 yields a result $A := A_{I \times J}$ satisfying (5.10) with $L = \operatorname{depth}(T) = \log_2(n/n_0)$, $n = \#I = \#J$.*

Remark 5.11 *The algorithm (3.2) requires a computational cost of*

$$80r^2 n + \tfrac{4064}{3} Lr^3 + \mathcal{O}(r^2)$$

operations.

Proof. In the worst case, the remainder has its support in the blocks of $P \cap T^{(L)}$. Then again we obtain estimate (5.10).

Since there is only one block in $P \cap T^{(\ell)}$ for each $0 \le \ell \le L-1$, the work is the sum of $40 \cdot 2^{L-\ell} r^2 n_0 + \frac{4064}{3} r^3 + \left(\frac{4}{3} + 2r - 2^{L-\ell+1} n_0\right) r$ (see Corollary 2.3, case (2.2c) with $n = m = 2^{L-\ell} n_0$) over $0 \le \ell \le L-1$, which is equal to $2Lr^2 + \frac{4064}{3} Lr^3 + \frac{4}{3} Lr + 80r^2 n - 4rn + 4rn_0 - 80r^2 n_0$. ∎

As mentioned in the proof, the estimate (5.10) with the amplification factor q^{L+1} can appear only if the support of the remainder Δ is concentrated in the four tiny subblocks of the level L. Obviously, this distribution of Δ is not very probable. Instead we may assume that Δ is equally distributed. Then $\|\Delta|_b\|^2 = \|\Delta\|^2 \#b/\#(I \times J) = 4^{-\ell} \|\Delta\|^2$ is the expectation value for $b \in P^{(\ell)} := P \cap T^{(\ell)}$. Since $\#P^{(\ell)} = 3$ for $0 \le \ell \le L-1$ and $\#P^{(L)} = 4$, Corollary 5.3 yields $\|A - B\|^2 \le \sum_{b \in P} q^{2(1+\operatorname{level}(b))} \|\Delta|_b\|^2 = \sum_{\ell=1}^{L-1} q^{2(\ell+1)} \|\Delta\|^2 + q^{2(L+1)} \cdot 4 \cdot 4^{-L} \|\Delta\|^2 < \frac{33\sqrt{5}+75}{10} \|\Delta\|^2$. This proves the following result.

Corollary 5.12 *In the case of Example 3.6 with an equally distributed Δ, the following inequality holds independently of the depth L of the tree:*

$$\|A - B\| \le c \|\Delta\| \qquad \text{with } c = \sqrt{\tfrac{33\sqrt{5}+75}{10}} = 3.857\ldots .$$

6 Discussion of Theorem 4.5

One observes from (4.5) that M must be of a very particular form to reach the bound q. Below we shall argue that the estimate by q is sharp but not probable.

The first singular values of (4.5) coincide. However, if $M = \sigma I_\mu$ with increasing rank μ, the ratio $q_\mu := \max \|A - B\| / \|\Delta\|$ tends to zero as $\mu \to \infty$. The following table shows q_μ for $r = 1$ and varying μ:

μ	2	3	4	10	100
q_μ	$(1+\sqrt{5})/2$	1.3066	1.1976	1.0599	1.0051

Usually, one expects that the singular values are decaying. In the following experiments, we consider sets of singular values $\Sigma = \{\sigma_1 > \sigma_2 > \ldots > \sigma_\mu\}$ of M with different decay rates. The corresponding worst cases are computed numerically.

For $\sigma_k = \exp(-k\omega)$, the following values are obtained:[7]

ω	0.01	0.05	0.1	0.2	0.4	0.5	0.8	1	2	5
q_ω	1.463	1.287	1.273	1.267	1.267	1.272	1.231	1.193	1.070	1.003

If $\omega \geq 0.34$, the worst case is given by

$$B = \operatorname{diag}\{\sigma_1, \ldots, \sigma_{r-1}, 0, \lambda_{r+1}, 0, \ldots\} \tag{6.1}$$

with suitable λ_{r+1}.

For $\sigma_k = \exp(-\sqrt{k\omega})$, the numerical results for $\mu = 50$, $r = 25$, and varying ω are[8]:

ω	0.01	0.05	0.1	0.2	0.5	1	2	5	10	20
q	1.447	1.336	1.303	1.279	1.264	1.261	1.259	1.263	1.257	1.267

In this case, the results depend weakly on the choice of r.

For $\sigma_k = 1/k^2$, the bounds are mainly between 1.12 and 1.25.

For $\sigma_k = 1/k$, typical values are between 1.17 and 1.27 (the latter value corresponds to $\mu = 2r$).

In particular, the bound improves for the asymptotic case of $\|\Delta\| \to 0$. Here, the asymptotic starts when $\|\Delta\|$ is clearly smaller than the singular value σ_r:

$$\frac{\|A - B\|}{\|\Delta\|} \leq 1 + \varepsilon + \mathcal{O}(\varepsilon^2) \qquad \text{as } \varepsilon := \frac{\|\Delta\|}{\sigma_r} \to 0.$$

The prototypic example is $r = 1$, $\sigma_1 = 1$, $M = \operatorname{diag}\{1, \varepsilon\}$, $A = \operatorname{diag}\{1, 0\}$, $B = \operatorname{diag}\{0, \vartheta\}$, $\Delta = \operatorname{diag}\{-1, \vartheta - \varepsilon\}$. The worse case is given by $\vartheta = \frac{1}{2}\varepsilon + \frac{1}{2}\sqrt{4 + \varepsilon^2}$ leading to $\frac{1+\vartheta^2}{1+(\vartheta-\varepsilon)^2} = 1 + \varepsilon + \frac{1}{2}\varepsilon^2 + \mathcal{O}(\varepsilon^3)$.

We conclude that for a realistic behaviour of the singular values, the bounds are smaller than $(1+\sqrt{5})/2$. Under the additional condition of Remark 4.6, values about 1.05 to 1.07 are rather probable.

7 Analysis for Unidirectional Partitions

We call a partition *unidirectional* if only partition rule A or only rule B is applied (cf. §3.1). Example 3.3 is unidirectional, whereas the other examples are not. We shall see that in the unidirectional case certain orthogonality properties hold allowing much better estimates. Without loss of generality, we assume in the following that only rule B is applied; i.e., all blocks $b \in T^{(\ell)}$ are of the 'horizontal' form $\tau \times J$ with $\tau \subset I$. Proposition 5.1 shows that in the unidirectional case additional properties are valid. However, we shall not use Proposition 5.1, but estimate in a different way.

Equation (5.8) describes the addition of the remainder E_b caused by the singular value decomposition in the block b at level $\ell = \operatorname{level}(b)$. Now we explicitly add up the remainders of all previous levels.

[7] The particular values are $\mu = 50$ and $r = 25$. For $\omega = 0$, the choice of $2r = \mu$ yields the worse case $q = (1+\sqrt{5})/2$. However, for larger values of ω, the results are rather independent of μ and r.

[8] For $\omega \geq 12$, the worse case is again given by (6.1).

First, we consider blocks $b \in P$ and set formally

$$E_b^{(\ell)} := 0 \qquad \text{for all } \ell > \mathrm{level}(b),\ b \in P.$$

Nonzero remainders $E_b^{(\ell)}$ occur only for blocks $b \in T^{(\ell)}$, since a singular value decomposition is performed at these blocks. We sum the distributions $E_{b'}^{(\ell)}$ of a fixed level ℓ over bigger blocks and define

$$E_b^{(\ell)} := \mathrm{Aggl}\{E_{b'}^{(\ell)} : b' \in T^{(\ell)}, b' \subset b\}.$$

Now the error F_b in $A_b = B|_b - F_b$ (cf. (5.6)) can be written as

$$F_b = \Delta|_b + \sum_{k=\ell}^{L} E_b^{(k)} \qquad \text{for } b \in T^{(\ell)}. \tag{7.1}$$

Lemma 7.1 *Let $b \in T^{(\ell)}$. Then $E_b^{(\ell')}$ and $E_b^{(\ell'')}$ are pairwise orthogonal for all $\ell' \neq \ell''$ with $\ell \le \ell', \ell'' \le L$. Furthermore, there exists a subspace V such that*[9] *$\hat{A}_b \in V \otimes \mathbb{R}^J$, while $E_b^{(k)} \in V^\perp \otimes \mathbb{R}^J$ for $\ell + 1 \le k \le L$.*

Proof. Consider $b = \tau \times J \in T^{(\lambda-1)}$ with $b = b_1 \cup b_2$, $\tau = \tau_1 \cup \tau_2$, where $b_i = \tau_i \times J \in T^{(\lambda)}$ (cf. rule B in §3.1). On the level λ, the singular value decomposition generates $\hat{A}_{b_i} = A_{b_i} + E_{b_i}^{(\lambda)}$, where $A_{b_i} \in V_i \otimes \mathbb{R}^J$ and $E_{b_i}^{(\lambda)} \in V_i^\perp \otimes \mathbb{R}^J$ $(i = 1, 2)$. At the next level $\lambda - 1$, the submatrix $\hat{A}_b := \mathrm{Aggl}\{A_{b_1}, A_{b_2}\}$ belongs to $\hat{V} \otimes \mathbb{R}^J$, where $\hat{V}$ is the agglomeration of subspaces defined by

$$\hat{V} := \mathrm{Aggl}\{V_1, V_2\} := \{v \in \mathbb{R}^\tau : v = \mathrm{Aggl}\{v_1, v_2\}, v_1 \in V_1, v_2 \in V_2\}.$$

The singular value decomposition of $\hat{A}_b$ yields $A_b + E_b^{(\lambda-1)}$. Both terms again belong to $\hat{V} \otimes \mathbb{R}^J$. Note that $\hat{V}$ is orthogonal to $\mathrm{Aggl}\{V_1^\perp, V_2^\perp\}$, so that

$$E_b^{(\lambda-1)} \perp E_b^{(\lambda)} = \mathrm{Aggl}\{E_{b_1}^{(\lambda)}, E_{b_2}^{(\lambda)}\}.$$

More precisely, $A_b \in V \otimes \mathbb{R}^J$ holds for a subspace $V \subset \hat{V}$, while $E_b \in (V^\perp \cap \hat{V}) \otimes \mathbb{R}^J$ belongs to an orthogonal subspace. By induction, one obtains $E_b^{(\ell)} \perp E_b^{(\lambda)}$ for all $b \in T^{(\ell)}$, $\ell < \lambda$. ∎

Note that the statements about the subspaces V_1, V_2, V do not hold for the corresponding minimal subspaces W_1, W_2, W with $A_{b_i} \in V_i \otimes W_i$ and $A_b \in V \otimes W$. The reason is that $\hat{A}_b := \mathrm{Aggl}\{A_{b_1}, A_{b_2}\}$ belongs to $\hat{V} \otimes \hat{W}$ with $\hat{W} = W_1 + W_2$. Since $W_2^\perp$ need not be orthogonal to W_1, one cannot conclude that $\hat{W}$ is orthogonal to $W_1^\perp + W_2^\perp$.

We still have to estimate $E_b^{(\ell)}$ for $b \in T^{(\ell)}$. As in the first approach, we represent the matrix $\hat{A}_b = A_b + E_b^{(\ell)}$ as a perturbation of $B|_b$:

$$\hat{A}_b = B|_b - \Delta|_b - \sum_{k=\ell+1}^{L} E_b^{(k)} \qquad \text{for } b \in T^{(\ell)}.$$

Let $\hat{A}_b \in V \times \mathbb{R}^J$ be the representation from Lemma 7.1 and define $P : \mathbb{R}^I \to V \subset \mathbb{R}^I$ as the orthogonal projection onto V. Since P maps $V^\perp$ into zero, we obtain $PE_b^{(k)} = 0$ $(\ell < k \le L)$ and therefore

$$\hat{A}_b = P\hat{A}_b = P(B|_b) - P(\Delta|_b).$$

Note that again $P(B|_b) \in \mathcal{R}(r, b)$. Therefore, Lemma 4.1a (with M, B, Δ replaced with $\hat{A}_b, P(B|_b), P(\Delta|_b)$) proves the estimate $\|E_b^{(\ell)}\| \le \|P(\Delta|_b)\| \le \|\Delta|_b\|$ for $b \in T^{(\ell)}$. The agglomeration of several $E_{b_i}^{(\ell)}$ yields again $\|E_b^{(\ell)}\| \le \|\Delta|_b\|$ for any $b \in T^{(k)}$, $k \le \ell$.

The estimate of F_b from (7.1) becomes

$$\begin{aligned}\|F_b\| &\le \left\|\Delta|_b + \sum_{k=\ell}^{L} E_b^{(k)}\right\| \le \|\Delta|_b\| + \left\|\sum_{k=\ell}^{L} E_b^{(k)}\right\| \le \|\Delta|_b\| + \sqrt{\sum_{k=\ell}^{L} \|E_b^{(k)}\|^2} \le \|\Delta|_b\| + \sqrt{\sum_{k=\ell}^{L} \|\Delta|_b\|^2} \\ &= \left(1 + \sqrt{L - \ell + 1}\right) \|\Delta|_b\|.\end{aligned}$$

The particular case for $\ell = 0$ proves the next theorem.

[9] For subspaces $V \subset \mathbb{R}^I$ and $W \subset \mathbb{R}^J$, the tensor space $V \otimes W$ is defined by $\mathrm{span}\{vw^\mathsf{T} : v \in V, w \in W\}$ (cf. [5, Remark 1.3]).

Theorem 7.2 *Assume that T describes a unidirectional partition of depth L. Let $M = B - \Delta$ with $\operatorname{rank}(B) \le r$, while A is the result of the recursive truncation. Then the following estimate holds:*

$$\frac{\|A - B\|}{\|\Delta\|} \le 1 + \sqrt{L+1}.$$

The computational cost is described in Proposition 5.4.

The representation $A = B - \Delta - \sum_{\ell=0}^{L} E_{I\times J}^{(\ell)}$ together with $M = B - \Delta$ yields the next result, which states that the recursive truncation is optimal up to a factor $\sqrt{L+1}$.

Corollary 7.3 *Under the conditions of Theorem 7.2 we have*

$$\frac{\|M - A\|}{\|\Delta\|} \le \sqrt{L+1}.$$

8 Mixed Application

The block decomposition of both Examples 3.4 and 3.5 yields a partition into an $n/r \times n/r$ block matrix consisting of $r \times r$ blocks. Such a partition can also be obtained by two unidirectional partitions.

Example 8.1 *Let $M \in \mathbb{R}^{I\times J}$ with $\#I = \#J = 2^p r$, where $r, p \in \mathbb{N}$. Apply partition rule B p times and then partition rule A p times. This yields a binary tree of depth $L := 2p$. The partition obtained in Example 3.3 is the result of the first p steps, while the final partition is the same as in Examples 3.4 and 3.5.*

The first p steps are unidirectional. Therefore, $A_b = B|_b - F_b$ ($b \in T^{(p)}$) holds with $\|A_b - B|_b\| \le \left(1 + \sqrt{L+1}\right)\|\Delta|_b\|$. The second p steps are again unidirectional, so that $\|A - B\| \le \left(1 + \sqrt{L+1}\right)\|F\|$, where $F := Aggl\{F_b : b \in T^{(p)}\}$. Together, we obtain the following result.

Proposition 8.2 *The recursive truncation described in Example 8.1 satisfies*

$$\frac{\|A - B\|}{\|\Delta\|} \le L + 2\sqrt{L+1} + 2.$$

The computational work of Example 8.1 is $\left((\frac{643}{3} + 6L)r + \frac{2}{3r} - L\right) n^2 + (6L + \frac{2}{r} - 40)nr^2$.

We recall that the decompositions of the Examples 3.4, 3.5, and 8.1 yield the same partition P. The third approach is the cheapest one. Furthermore, this approach offers the minimal amplification factor.

In principle, also Example 3.4 is of mixed type: partition step A is unidirectional as well as partition step B, but not their combination. Correspondingly, one can prove that subsequent terms $E_b^{(k)}$ in (7.1) satisfy $E_b^{(k)} \perp E_b^{(k+1)}$. However, it is not obvious whether this property can be exploited to obtain better estimates than (5.10).

9 Numerical Examples

We consider the Helmholtz equation $\Delta u + k^2 u = 0$ with $k = 100$ on the unit sphere $S = \{x \in \mathbb{R}^3 : |x| = 1\}$. The corresponding single-layer operator has the kernel function $k(x,y) = \frac{1}{4\pi}\frac{\exp(\mathrm{i}k|x-y|)}{|x-y|}$. As discretisation we use the Galerkin method with piecewise constant elements $\tau \in T$ defined on small flat triangles with corner points on S. The system matrix K has the entries $K_{\tau\sigma} := \int_\tau \int_\sigma k(x,y)\mathrm{d}x\mathrm{d}y$ for all $\tau, \sigma \in T$ (the index set T is of the size $\#T = 1026$). Let Δ' and Δ'' be two large spherical triangles on S both consisting of 128 elements of T. The algorithm is applied to the submatrix $M = K|_b \in \mathbb{R}^{128\times 128}$ where $b = \{(\tau,\sigma) : \tau \subset \Delta', \sigma \subset \Delta''\}$. The spherical triangles Δ' and Δ'' meet in one corner point. Therefore, they satisfy only a weak admissibility condition (cf. [4, §9.3] and [6]). This fact as well as the large wave number k leads to a slow decay of the singular values of the singular values depicted left in Fig. 9.1.

The recursive truncation starts with blocks of size 16×16. The further recursion is performed according to the Examples 3.3 to 3.5. The results for the target ranks $r = 1$ to 65 are shown in Fig. 9.1. In particular,

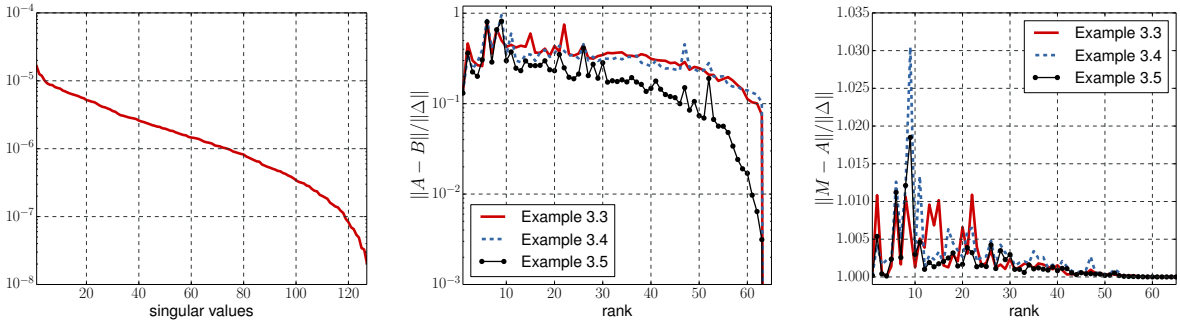


Figure 9.1: Left: singular values; middle and right: errors $\frac{\|A-B\|}{\|\Delta\|}$ and $\frac{\|M-A\|}{\|\Delta\|}$ for various ranks r

the values of $\|M - A\| / \|\Delta\|$ indicate that the result of the recursive truncation is rather close to the best approximation, which is characterised by the value 1.

Other tests yield results where $\|A - B\| / \|\Delta\|$ is close to zero, i.e., the result of the recursive truncation is more or less equal to the SVD_r result. We could not find practical examples for which A deviates strongly from B.

Acknowledgement. We thank Dr. R. Kriemann for providing the numerical tests.

References

[1] M. Bebendorf and W. Hackbusch, *Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients*, Numer. Math., 95 (2003), pp. 1–28.

[2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, 3rd ed., 1996.

[3] L. Grasedyck and W. Hackbusch, *Construction and arithmetics of $\mathcal{H}$-matrices*, Computing, 70 (2003), pp. 295–334.

[4] W. Hackbusch, *Hierarchische Matrizen - Algorithmen und Analysis*, Springer, Berlin, 2009.

[5] ——, *Tensor Spaces and Numerical Tensor Calculus*, vol. 42 of SCM, Springer, Berlin, 2012.

[6] W. Hackbusch, B. Khoromskij, R. Kriemann, *Hierarchical matrices based on a weak admissibility criterion*. Computing, 73 (2004), pp. 207–243.

[7] U. Kandler and C. Schröder, *Spectral error bounds for Hermitian inexact Krylov methods*. Preprint 11-2014, Institute of Mathematics, Technische Universität Berlin, 2014.

[8] Y. Nakatsukasa and N. J. Higham, *Stable and efficient divide and conquer algorithms for the symmetric eigenvalue decomposition and the SVD*, SIAM J. Sci. Comput., 35 (2013), pp. A1325–A1349.

[9] E. Schmidt, *Zur Theorie der linearen und nichtlinearen Integralgleichungen. I. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener*, Math. Ann., 63 (1907), pp. 433–476.