

DISTRIBUTED HIERARCHICAL SVD IN THE HIERARCHICAL TUCKER FORMAT

LARS GRASEDYCK* AND CHRISTIAN LÖBBERT*

Abstract. We consider tensors in the Hierarchical Tucker format and suppose the tensor data to be distributed among several compute nodes. We assume the compute nodes to be in a one-to-one correspondence with the nodes of the Hierarchical Tucker format such that connected nodes can communicate with each other. An appropriate tree structure in the Hierarchical Tucker format then allows for the parallelization of basic arithmetic operations between tensors with a parallel runtime which grows like $\log(d)$, where d is the tensor dimension. We introduce parallel algorithms for several tensor operations, some of which can be applied to solve linear equations $\mathcal{A}X = B$ directly in the Hierarchical Tucker format using iterative methods like conjugate gradients or multigrid. We present weak scaling studies, which provide evidence that the runtime of our algorithms indeed grows like $\log(d)$. Furthermore, we present numerical experiments in which we apply our algorithms to solve a parameter-dependent diffusion equation in the Hierarchical Tucker format by means of a multigrid algorithm.

Keywords: Hierarchical Tucker, HT, Multigrid, Parallel algorithms, SVD, Tensor arithmetic

1. Introduction. High dimensional tensors may e.g. arise in the context of parameter-dependent problems. Consider a linear equation

$$A(p_1, p_2, \dots, p_d) \cdot \vec{x}(p_1, p_2, \dots, p_d) = \vec{b}(p_1, p_2, \dots, p_d),$$

where the matrix A , the right-hand side \vec{b} , and with that also the solution \vec{x} , depend on a (possibly high) number of parameters. Then the solution vector \vec{x} can be regarded as a tensor X (i.e. a multidimensional array) of dimension $d + 1$, where we have one tensor dimension for each of the d parameters plus the vector dimension of \vec{x} :

$$x_i(p_1, p_2, \dots, p_d) = X(p_1, p_2, \dots, p_d, i).$$

In the same way, the right-hand side \vec{b} and the matrix A can be regarded as $(d + 1)$ -dimensional tensors B and \mathcal{A} , where the two vector indices i, j of the matrix A are combined to one index (i, j) :

$$A_{ij}(p_1, p_2, \dots, p_d) = \mathcal{A}(p_1, p_2, \dots, p_d, (i, j)). \quad (1.1)$$

If low-rank representations/approximations of the tensors \mathcal{A} and B are available and if basic arithmetic operations can directly be performed in the underlying low-rank format, then an approximate solution X of the tensor equation $\mathcal{A}X = B$ can be obtained by applying some iterative method directly in the low-rank format. The solution tensor X will then contain all solutions $\vec{x}(p_1, p_2, \dots, p_d)$ for any parameter combination and the solutions for single parameter combinations can easily be extracted from X . Furthermore several postprocessing operations like computing the mean over (all) parameter combinations or computing expected values with respect to a probability distribution of the parameters can directly be carried out for the solution X in the low-rank format.

In this article we present parallel algorithms for basic arithmetic operations on tensors, where we choose the Hierarchical Tucker format as low-rank format for tensors. The Hierarchical Tucker format is briefly introduced in Section 2. For a more detailed description of low-rank tensor approximation techniques we refer to the literature survey [7]. For our

*Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55, 52056 Aachen, Germany. Email: {lgr,loebbert}@igpm.rwth-aachen.de. The authors gratefully acknowledge support by the DFG priority programme 1648 (SPPEXA) under grant GR-3179/4-2.

algorithms the tensors may be stored distributed over several compute nodes. As a consequence the algorithms can also be used for postprocessing operations on tensors which stem from a parallel sampling method on distinct compute nodes, as e.g. described in [8].

In Section 3 we give an overview of our parallel algorithms for the Hierarchical Tucker format. The Hierarchical Tucker format is based on a binary tree (cf. Section 2, Fig. 2.1). For this article we assume the data of each tree node to be stored on its own compute node. Then our algorithms typically require communication between compute nodes which are neighbors with respect to the binary tree. Assuming the number of dimensions d to be a power of two, $d = 2^p$, $p \in \mathbb{N}$, and choosing the tree T_d such that the number of tree levels is minimized, our algorithms can run in parallel on all compute nodes of the same tree level. In Section 3 we refer to this as *level-wise parallelization*. Since we have $p + 1 = \log_2(d) + 1$ tree levels, we expect the parallel runtime of our algorithms to grow like $\log(d)$.

In Section 4 we present runtime tests which provide evidence that the parallel runtime of our algorithms indeed grows like $\log(d)$ for arithmetic operations between tensors of dimension d .

In Section 5 we apply our algorithms to solve a parameter-dependent toy problem in the Hierarchical Tucker format by means of iterative methods. We chose a diffusion equation with piecewise constant diffusion coefficients which are controlled by 9 parameters. Applying a Finite Element discretization we obtain a linear system, the matrix of which depends on the 9 parameters. Due to its affine structure with respect to the parameters, the system matrix can directly be represented as linear operator in the Hierarchical Tucker format, i.e. we can use iterative methods to solve the linear equation, where all arithmetic operations are carried out between tensors in the Hierarchical Tucker format. We illustrate this for a multigrid method containing a Richardson method as smoother. On the coarsest grid level we use a CG method to solve the defect equation. This numerical experiment shows less the benefit of the parallelization (which is demonstrated in Section 4) but rather confirms that our algorithms can in principle be used together with iterative methods.

2. Distributed tensors in the Hierarchical Tucker format. We consider real-valued tensors $A \in \mathbb{R}^{\mathcal{I}}$ for some product index set $\mathcal{I} = \mathcal{I}_1 \times \dots \times \mathcal{I}_d$, where the \mathcal{I}_μ , $\mu = 1, \dots, d$, are index sets of size $n_\mu \in \mathbb{N}$, e.g. $\mathcal{I}_\mu = \{1, \dots, n_\mu\}$. The number $d \in \mathbb{N}$ is referred to as *dimension* of the tensor. A real-valued tensor is thus a mapping $A : \mathcal{I} \rightarrow \mathbb{R}$ from a product index set \mathcal{I} to the set \mathbb{R} of real numbers.

Assuming the index sets \mathcal{I}_μ to be all of the same size $n := n_\mu$, $\mu = 1, \dots, d$, a tensor $A \in \mathbb{R}^{\mathcal{I}}$ consists of n^d entries and may be regarded as a vector in $\mathbb{R}^{n^d} \cong \mathbb{R}^{\mathcal{I}}$. To emphasize the tensor structure, we prefer the notation $A \in \mathbb{R}^{\mathcal{I}}$, $\mathcal{I} = \mathcal{I}_1 \times \dots \times \mathcal{I}_d$, instead of $A \in \mathbb{R}^{n^d}$.

Even for moderate numbers n and d the exponential growth of the number n^d of tensor entries makes it impossible to store the tensor entry-by-entry, not to mention the costs for arithmetic operations. This *curse of dimensionality* led to a variety of different tensor formats which make use of low rank structures in the tensors.

In this article the *Hierarchical Tucker format* [7, 5, 11] is used, which is based on a hierarchy of the tensor dimensions $\{1, \dots, d\}$, as depicted in Fig 2.1 for $d = 8$. The *dimension tree* of Fig. 2.1 is denoted by T_8 , since $d = 8$. Of course one could think of many different ways how to split a node $\{\mu, \dots, \nu\}$ of the tree T_d into its two sons. However we will always assume a *balanced* decomposition where $\{\mu, \dots, \nu\}$ is subdivided into the two sons

$$\left\{ \mu, \dots, \left\lfloor \frac{\mu + \nu - 1}{2} \right\rfloor \right\} \quad \text{and} \quad \left\{ \left\lfloor \frac{\mu + \nu - 1}{2} \right\rfloor + 1, \dots, \nu \right\}. \quad (2.1)$$

This *balanced tree* minimizes the number of tree levels: $\text{depth}(T_d) = \lceil \log_2(d) \rceil$, whereas a

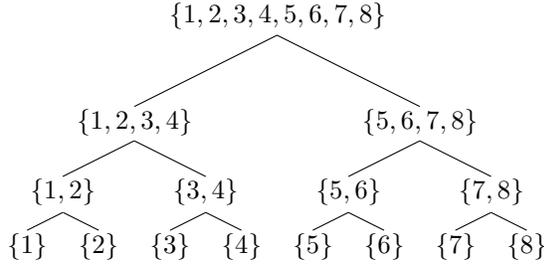


Fig. 2.1: Balanced dimension tree for dimension $d = 8$.

subdivision of $\{\mu, \dots, \nu\}$ into $\{\mu\}$ and $\{\mu + 1, \dots, \nu\}$ would result in $\text{depth}(T_d) = d - 1$. Since our algorithms will allow for level-wise parallelization, the balanced tree appears to be the best choice here.

For simplicity we assume the dimension d to be a power of two: $d = 2^p$ for some $p \in \mathbb{N}$, i.e. the leaves $\{\mu\}$, $\mu = 1, \dots, d$, are all on the highest tree level and we always have 2^ℓ tree nodes on level ℓ , where level 0 is the root and level p is the highest level.

2.1. The Hierarchical Tucker format. Consider a tensor $A \in \mathbb{R}^{\mathcal{I}}$ for a product index set \mathcal{I} of dimension d , as above. Then for a subset $t \subset \{1, \dots, d\}$ let $[t] = \{1, \dots, d\} \setminus t$ denote the complement of t and we define the product index sets \mathcal{I}_t and $\mathcal{I}_{[t]}$ by

$$\mathcal{I}_t = \times_{\mu \in t} \mathcal{I}_\mu \quad \text{and} \quad \mathcal{I}_{[t]} = \times_{\mu \in [t]} \mathcal{I}_\mu.$$

DEFINITION 2.1 (Matricization). *The matricization $\mathcal{M}_t(A) \in \mathbb{R}^{\mathcal{I}_t \times \mathcal{I}_{[t]}}$ of the tensor $A \in \mathbb{R}^{\mathcal{I}}$ with respect to a subset $t \subset \{1, \dots, d\}$ is defined via*

$$\mathcal{M}_t(A)_{(\mathbf{i}_t, \mathbf{i}_{[t]})} = A_{(i_1, \dots, i_d)} \quad \text{for} \quad (i_1, \dots, i_d) \in \mathcal{I} = \mathcal{I}_1 \times \dots \times \mathcal{I}_d,$$

where we used the short notation $\mathbf{i}_s = (i_\mu)_{\mu \in s} \in \mathcal{I}_s$ for subsets $s \subset \{1, \dots, d\}$. For a tensor $A \in \mathbb{R}^{\mathcal{I}}$ the Hierarchical Tucker format is based on matricizations of A with respect to the subsets $t \subset \{1, \dots, d\}$ of some underlying dimension tree T_d .

For each node $t \in T_d$ of the tree let U_t be a matrix, the columns of which span the same vector space as the columns of the matricization $\mathcal{M}_t(A)$:

$$\text{range}(U_t) = \text{range}(\mathcal{M}_t(A)).$$

Such a matrix will be called *frame* from now on.

If $\text{rank}(\mathcal{M}_t(A)) = k_t$, one can find a frame U_t with k_t columns, i.e. $U_t \in \mathbb{R}^{\mathcal{I}_t \times k_t}$. For the root $D := \{1, \dots, d\}$ of T_d the matricization $\mathcal{M}_D(A) \in \mathbb{R}^{\mathcal{I} \times 1}$ is just the rearrangement of the tensor A in one long column and one can choose $U_D = \mathcal{M}_D(A)$, i.e. the frame U_D becomes as hard to handle as the tensor A itself.

In the Hierarchical Tucker format the frames U_t are only stored for the leaves $t = \{\mu\}$ of the tree, for which the $U_{\{\mu\}}$ are of size $n_{\{\mu\}} \times k_{\{\mu\}}$, $\mu = 1, \dots, d$. For a non-leaf node t the row index set \mathcal{I}_t of U_t could already be too large, which is why we do not store U_t explicitly for non-leaf nodes.

Each non-leaf node t has two sons s_1 and s_2 fulfilling $t = s_1 \dot{\cup} s_2$, and one can find coefficients $(B_t)_{i,j,\ell}$, such that

$$(U_t)_{-,i} = \sum_{j=1}^{k_{s_1}} \sum_{\ell=1}^{k_{s_2}} (B_t)_{i,j,\ell} ((U_{s_1})_{-,j} \otimes (U_{s_2})_{-, \ell}), \quad (2.2)$$

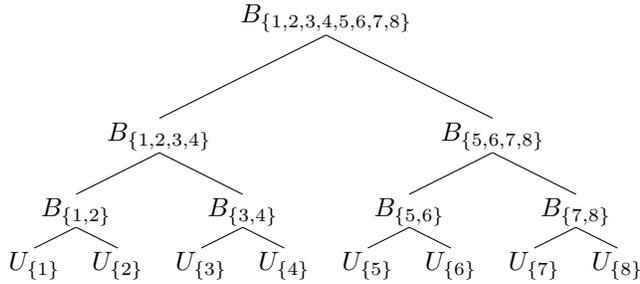


Fig. 2.2: Hierarchical tensor of dimension $d = 8$ with underlying tree T_d of Fig. 2.1

where $(U_t)_{-,i} \in \mathbb{R}^{\mathcal{I}_t}$ denotes the i -th column of U_t .

Instead of $U_t \in \mathbb{R}^{\mathcal{I}_t \times k_t}$ only the coefficients $(B_t)_{i,j,\ell}$ are stored, where $i \in \{1, \dots, k_t\}$, $j \in \{1, \dots, k_{s_1}\}$ and $\ell \in \{1, \dots, k_{s_2}\}$. This results in a tensor B_t of size $k_t \times k_{s_1} \times k_{s_2}$ for each non-leaf node, referred to as *transfer tensor*. Note that we have $k_D = 1$ for the root (U_D is one single column), i.e. B_D is a $k_{t_1} \times k_{t_2}$ -matrix, where $\text{sons}(D) = \{t_1, t_2\}$.

For any index $(i_1, \dots, i_d) \in \mathcal{I}$ we have

$$A_{(i_1, \dots, i_d)} = \mathcal{M}_D(A)_{(i_1, \dots, i_d), 1} = (U_D)_{(i_1, \dots, i_d), 1}$$

and the corresponding tensor entry can be evaluated by recursively using the relation (2.2).

A Hierarchical Tucker tensor (for short, \mathcal{H} -tensor) can be depicted by the underlying tree T_d , cf. Fig 2.2. Writing $n := \max\{n_\mu \mid \mu = 1, \dots, d\}$ and $k := \max\{k_t \mid t \in T_d\}$, we can estimate the number of entries stored in the Hierarchical Tucker format by

$$\#\text{leaves} \cdot n \cdot k + \#\text{(inner nodes)} \cdot k^3 + k^2 = dnk + (d-2)k^3 + k^2 = \mathcal{O}(d). \quad (2.3)$$

The numbers k_t , $t \in T_d$, are called *Hierarchical ranks* of the tensor. To be precise, one can distinguish between the actual *Hierarchical representation ranks* $(k_t)_{t \in T_d}$ and the minimal ranks, since the k_t , $t \in T_d$, need not necessarily to be minimal: In the case

$$k_t > \text{rank}(\mathcal{M}_t(A))$$

we can reduce the complexity of the representation until we have $k_t = \text{rank}(\mathcal{M}_t(A))$. This will, of course, always be the goal, we will even truncate tensors down to lower Hierarchical ranks, which means to find an approximating tensor of lower Hierarchical ranks.

Since we only deal with tensors in the Hierarchical Tucker format, we will sometimes use the short name *tensor ranks* or just *ranks*.

In our runtime tests, we will, for simplicity, choose $k_t = k$ for all $t \in T_d \setminus \{1, \dots, d\}$, and therefore speak of *the (Hierarchical) rank*.

2.2. Distributing \mathcal{H} -tensors. In this article we assume the tensor $A \in \mathbb{R}^{\mathcal{I}}$ to be stored in the Hierarchical Tucker format, where the underlying dimension tree T_d is of balanced type according to (2.1). Furthermore, we assume the data to be distributed among several compute nodes, where the data of each tree node is stored on its own compute node. For tensor dimension d this leads to $2d - 1$ compute nodes. The tensor in Fig. 2.2 would be distributed among 15 compute nodes. Communication is supposed to be possible via the edges of the tree, i.e. each node can communicate with its sons and its father.

The distribution of the tensor data over the $2d - 1$ compute nodes was realized by MPI. In our implementation we store the structure of the tree T_d on each MPI process, which

includes the MPI ranks (i.e. the IDs of the MPI processes) for all the other nodes. In fact it would be sufficient that each MPI process knows the MPI ranks of its sons and its father.

As mentioned in Section 1, our algorithms will typically run in parallel on MPI processes which belong to the same tree level in T_d (level-wise parallelization).

Distributed \mathcal{H} -tensors may arise from parallel tensor sampling [8], where a given output function for solutions of a parameter dependent problem is approximated in the Hierarchical Tucker format.

If $2d - 1$ MPI processes are not available, the distribution of the tree nodes to the available MPI processes would be an interesting issue. This question will, however, not be discussed in this article and may be part of further investigations.

3. Parallel arithmetic for \mathcal{H} -tensors. Given a tensor that has either been assembled in parallel on distributed compute nodes or has been distributed due to its extent of storage, one would like to run tensor computations in parallel to the highest possible extent. In this section we demonstrate how arithmetic operations in the Hierarchical Tucker format can be parallelized level-wise with respect to the underlying tree T_d , where we assume the tree nodes to correspond to the compute nodes (cf. Section 2.2). By using a balanced tree T_d we can thus expect the parallel computing time to grow like $\text{depth}(T_d) = \log_2(d)$, assuming load balancing between the nodes.

In this section we consider the inner product $\langle A, B \rangle \in \mathbb{R}$,

$$\langle A, B \rangle = \sum_{(i_1, \dots, i_d) \in \mathcal{I}} A_{(i_1, \dots, i_d)} \cdot B_{(i_1, \dots, i_d)} \quad , \quad (3.1)$$

of two \mathcal{H} -tensors $A, B \in \mathbb{R}^{\mathcal{I}}$, the addition $A + B$, and the application of an operator \mathcal{L} to an \mathcal{H} -tensor A , both followed by the truncation of the result, either down to prescribed ranks $(k_t)_{t \in T_d}$ or with respect to prescribed accuracy ε .

Perhaps the most basic operation for an \mathcal{H} -tensor $A \in \mathbb{R}^{\mathcal{I}}$ is the evaluation of single tensor entries for given indices $(i_1, \dots, i_d) \in \mathcal{I}$. As we do not store the tensor entry-by-entry, this is as well an issue. The evaluation has briefly been addressed in Section 2.1 and can also be parallelized level-wise.

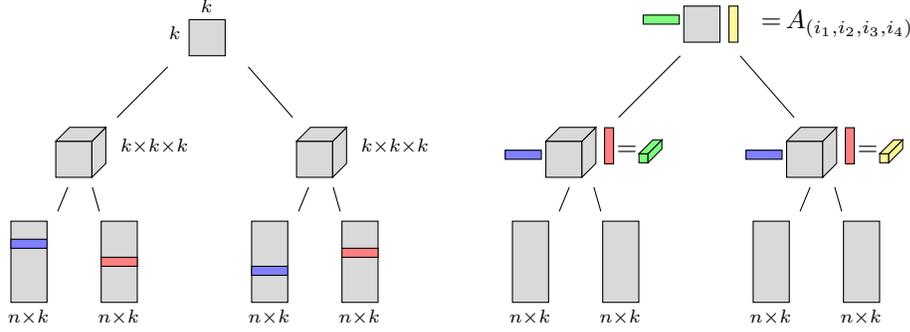
3.1. Evaluation of tensor entries in the Hierarchical Tucker format. Assume the tensor $A \in \mathbb{R}^{\mathcal{I}}$ to be stored in the Hierarchical Tucker format, based on the dimension tree T_d . For a given index $(i_1, \dots, i_d) \in \mathcal{I}$ the tensor entry $A_{(i_1, \dots, i_d)}$ shall be computed. Note that each leaf frame $U_{\{\mu\}} \in \mathbb{R}^{\mathcal{I}_\mu \times k_{\{\mu\}}}$ only depends on the index $i_\mu \in \mathcal{I}_\mu$, $\mu = 1, \dots, d$. Thus in each leaf frame we just choose the corresponding row $(U_{\{\mu\}})_{i_\mu, -}$ and send it to the father node. This can be carried out in parallel for all leaves.

Consider the node $t = \{\mu, \nu\}$ as father of the leaves $\{\mu\}$ and $\{\nu\}$ in T_d . The compute node belonging to t then receives the rows $(U_{\{\mu\}})_{i_\mu, -}$ and $(U_{\{\nu\}})_{i_\nu, -}$ from its sons and can thus compute the row $(U_{\{\mu, \nu\}})_{(i_\mu, i_\nu), -}$ of its own frame by (2.2) and send it to its father (remember that for an inner node t the frame U_t is not stored explicitly - instead the coefficients B_t are stored):

$$(U_t)_{(i_\mu, i_\nu), q} = \sum_{j=1}^{k_{\{\mu\}}} \sum_{\ell=1}^{k_{\{\nu\}}} (B_t)_{q, j, \ell} \cdot (U_{\{\mu\}})_{i_\mu, j} \cdot (U_{\{\nu\}})_{i_\nu, \ell} \quad , \quad q = 1, \dots, k_t.$$

In general each inner node $t \in T_d$ with sons $\text{sons}(t) = \{s_1, s_2\}$ receives the rows $(U_{s_1})_{i_{s_1}, -}$ and $(U_{s_2})_{i_{s_2}, -}$ from its sons and then computes the row $(U_t)_{i_t, -}$ using (2.2):

$$(U_t)_{i_t, q} = \sum_{j=1}^{k_{s_1}} \sum_{\ell=1}^{k_{s_2}} (B_t)_{q, j, \ell} \cdot (U_{s_1})_{i_{s_1}, j} \cdot (U_{s_2})_{i_{s_2}, \ell} \quad , \quad q = 1, \dots, k_t \quad , \quad (3.2)$$



(a) Highest level ($\ell = 2$): In each leaf node $\{\mu\}$, $\mu = 1, 2, 3, 4$ choose the row i_μ and send it to the father node.

(b) Middle level ($\ell = 1$): Multiply the received rows of the sons along the 1st and the 2nd tensor dimension. Send the result to the father (root level, $\ell = 0$), where the tensor entry $A_{(i_1, i_2, i_3, i_4)}$ is finally computed.

Fig. 3.1: Parallel evaluation of one tensor entry for a given index $(i_1, i_2, i_3, i_4) \in \mathcal{I}$ ($d = 4$) in the Hierarchical Tucker format. The operations can be parallelized level-wise ($n_\mu = n$ and $k_t = k$ for $\mu = 1, 2, 3, 4$ and $t \in T_d$).

where $\mathbf{i}_s = (i_\mu)_{\mu \in s}$ denotes the subindex of (i_1, \dots, i_d) corresponding to the subset $s \subset \{1, \dots, d\}$. After computing the row $(U_t)_{\mathbf{i}_t, q}$, it is sent to the father node.

Finally the entry $A_{(i_1, \dots, i_d)}$ is obtained on the root node $D = \{1, \dots, d\}$ after the rows $U_{\mathbf{i}_{t_1}, -}$ and $U_{\mathbf{i}_{t_2}, -}$ of the root sons t_1, t_2 have been received:

$$A_{(i_1, \dots, i_d)} = (U_D)_{((i_1, \dots, i_d), 1)} = \sum_{j=1}^{k_{t_1}} \sum_{\ell=1}^{k_{t_2}} (B_D)_{j, \ell} U_{\mathbf{i}_{t_1}, j} U_{\mathbf{i}_{t_2}, \ell}$$

(remember: the root transfer tensor B_D is only 2-dimensional).

The multiplications in (3.2) need $\mathcal{O}(k_t k_{s_1} k_{s_2})$ operations for each inner node $t \in T_d$. Depending on the maximal rank $k := \max\{k_t \mid t \in T_d\}$, we can therefore estimate the complexity in each node by $\mathcal{O}(k^3)$.

Assuming the index (i_1, \dots, i_d) to be known on each leaf node of the tree, the process of selecting all the rows $(U_{\{\mu\}})_{i_\mu, -}$, $\mu = 1, \dots, d$, in the leaves and sending them to the respective father nodes works independently on each leaf and can therefore be carried out in parallel on all leaves.

Each inner node t on level ℓ of the tree requires only the results of its sons s_1 and s_2 , which are on level $\ell + 1$. Thus no interaction to any other node on level ℓ is required, i.e. all nodes on level ℓ work entirely independent of each other and can act in parallel.

This demonstrates that the evaluation can be parallelized level-wise with respect to the tree T_d as illustrated in Fig. 3.1.

3.2. The inner product of two \mathcal{H} -tensors. The inner product $\langle A, C \rangle$ of two \mathcal{H} -tensors $A, C \in \mathbb{R}^{\mathcal{I}}$, $\mathcal{I} = \mathcal{I}_1 \times \dots \times \mathcal{I}_d$, is defined by (3.1), i.e. the sizes $n_\mu = \#\mathcal{I}_\mu$, $\mu = 1, \dots, d$, have to coincide for A and C . The tensor ranks, however, can be different for A and C , which is why we now write $(k_t)_{t \in T_d}$ for the Hierarchical ranks of A and

$(\bar{k}_t)_{t \in T_d}$ for the Hierarchical ranks of C . Nonetheless we use the same tree T_d for A and C and even suppose them to be distributed on the same compute nodes, i.e. the data of A , which belongs to a certain tree node of T_d is stored on the same compute node as the corresponding data of C belonging to the same tree node. The latter assumption can, however, be left off, which would simply lead to more communication being needed.

As before, we denote the transfer tensors of A by B_t for all non-leaf nodes $t \in T_d$ and write $U_{\{\mu\}}$, $\mu = 1, \dots, d$, for the leaf frames, or more general U_t for the frame of node $t \in T_d$ (these are only stored for the leaves, cf. Section 2.1). For the tensor C we use the notation F_t for the transfer tensors and V_t for the frames.

In order to compute the inner product $\langle A, C \rangle$ in the Hierarchical format we first define recursively the matrices $\Phi_t \in \mathbb{R}^{k_t \times \bar{k}_t}$ for each tree node $t \in T_d$, which will coincide with the inner product matrices $(U_t)^\top V_t$, $t \in T_d$. The inner product will finally be obtained by $\langle A, C \rangle = (U_D)^\top V_D = \Phi_D$, where $D = \{1, \dots, d\}$ is the root of T_d .

DEFINITION 3.1. *Let $A, C \in \mathbb{R}^{\mathcal{I}}$ be \mathcal{H} -tensors, based on the same tree T_d with ranks $(k_t)_{t \in T_d}$ and $(\bar{k}_t)_{t \in T_d}$, i.e. for $t \in T_d$ we have $U_t \in \mathbb{R}^{\mathcal{I}_t \times k_t}$ and $V_t \in \mathbb{R}^{\mathcal{I}_t \times \bar{k}_t}$ (cf. Section 2.1) fulfilling*

$$\text{range}(U_t) = \text{range}(\mathcal{M}_t(A)) \quad \text{and} \quad \text{range}(V_t) = \text{range}(\mathcal{M}_t(C)),$$

as well as transfer tensors B_t and F_t for all non-leaf nodes $t \in T_d$ fulfilling (2.2), which reads

$$(V_t)_{-,i} = \sum_{j=1}^{\bar{k}_{s_1}} \sum_{\ell=1}^{\bar{k}_{s_2}} (F_t)_{i,j,\ell} ((V_{s_1})_{-,j} \otimes (V_{s_2})_{-, \ell}) \quad (3.3)$$

for the tensor C , where $\text{sons}(t) = \{s_1, s_2\}$.

The matrices Φ_t are then defined as $\Phi_{\{\mu\}} = (U_{\{\mu\}})^\top V_{\{\mu\}}$ for the leaves, $\mu = 1, \dots, d$, and recursively by

$$(\Phi_t)_{j_t, \bar{j}_t} = \sum_{j_{s_1}=1}^{k_{s_1}} \sum_{\bar{j}_{s_1}=1}^{\bar{k}_{s_1}} \sum_{j_{s_2}=1}^{k_{s_2}} \sum_{\bar{j}_{s_2}=1}^{\bar{k}_{s_2}} (B_t)_{j_t, j_{s_1}, j_{s_2}} (\Phi_{s_1})_{j_{s_1}, \bar{j}_{s_1}} (\Phi_{s_2})_{j_{s_2}, \bar{j}_{s_2}} (F_t)_{\bar{j}_t, \bar{j}_{s_1}, \bar{j}_{s_2}}. \quad (3.4)$$

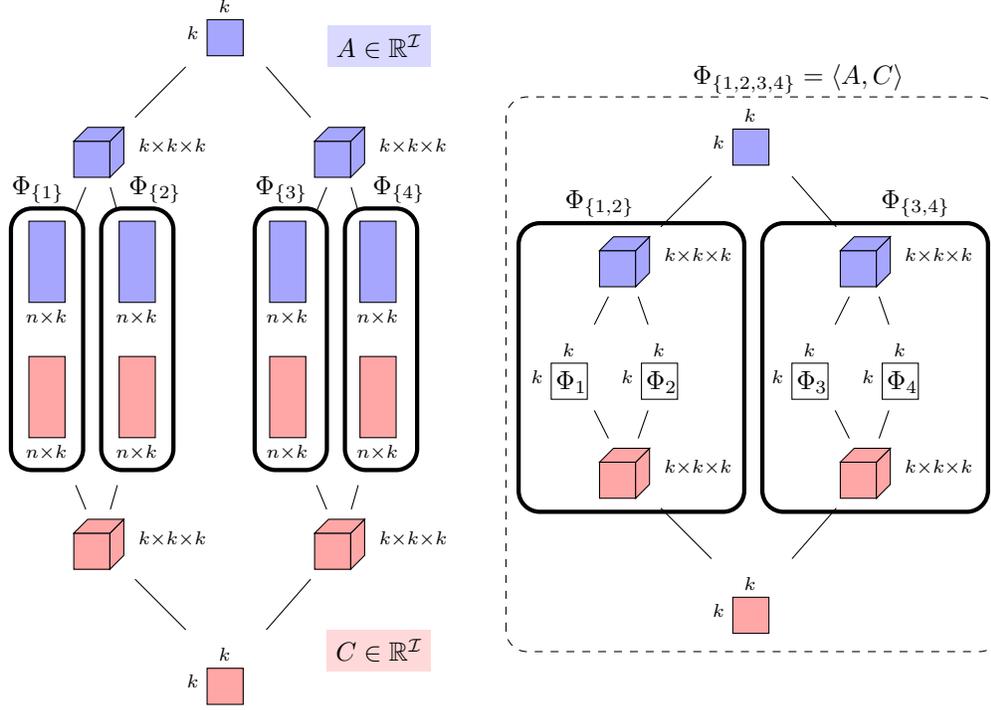
for all non-leaf nodes $t \in T_d$.

LEMMA 3.2. *For each node $t \in T_d$ we have*

$$\Phi_t = (U_t)^\top V_t. \quad (3.5)$$

Proof. On the highest level of the tree (3.5) holds by definition: $\Phi_{\{\mu\}} = (U_{\{\mu\}})^\top V_{\{\mu\}}$, $\mu = 1, \dots, d$. Assume (3.5) being true for level ℓ of the tree. Then for any node $t \in T_d$ of the next lower level $\ell - 1$ we may use that (3.5) holds true for the sons s_1 and s_2 of t , which are on level ℓ . Then (3.4) yields

$$\begin{aligned} & (\Phi_t)_{j_t, \bar{j}_t} \\ &= \sum_{\substack{j_{s_1}, \bar{j}_{s_1} \\ j_{s_2}, \bar{j}_{s_2}}} (B_t)_{j_t, j_{s_1}, j_{s_2}} \left(\sum_{\mathbf{i}_{s_1}} (U_{s_1})_{\mathbf{i}_{s_1}, j_{s_1}} V_{\mathbf{i}_{s_1}, \bar{j}_{s_1}} \right) \left(\sum_{\mathbf{i}_{s_2}} (U_{s_2})_{\mathbf{i}_{s_2}, j_{s_2}} (V_{s_2})_{\mathbf{i}_{s_2}, \bar{j}_{s_2}} \right) (F_t)_{\bar{j}_t, \bar{j}_{s_1}, \bar{j}_{s_2}} \\ &= \sum_{\mathbf{i}_{s_1}, \mathbf{i}_{s_2}} \sum_{j_{s_1}, j_{s_2}} (B_t)_{j_t, j_{s_1}, j_{s_2}} (U_{s_1})_{\mathbf{i}_{s_1}, j_{s_1}} (U_{s_2})_{\mathbf{i}_{s_2}, j_{s_2}} \sum_{\bar{j}_{s_1}, \bar{j}_{s_2}} (F_t)_{\bar{j}_t, \bar{j}_{s_1}, \bar{j}_{s_2}} (V_{s_1})_{\mathbf{i}_{s_1}, \bar{j}_{s_1}} (V_{s_2})_{\mathbf{i}_{s_2}, \bar{j}_{s_2}} \\ &\stackrel{(2.2)}{=} \sum_{\mathbf{i}_{s_1}, \mathbf{i}_{s_2}} (U_t)_{\mathbf{i}_t, j_t} (V_t)_{\mathbf{i}_t, \bar{j}_t} = \sum_{\mathbf{i}_t} (U_t)_{\mathbf{i}_t, j_t} (V_t)_{\mathbf{i}_t, \bar{j}_t} = (U_t)^\top V_t, \end{aligned} \quad (3.3)$$



(a) Highest level ($\ell = 2$): On each leaf node $\{\mu\}$, $\mu = 1, 2, 3, 4$, the matrix $\Phi_{\{\mu\}} = (U_{\{\mu\}})^\top V_{\{\mu\}}$ is computed and sent to the father node(s).

(b) Middle level ($\ell = 1$): The matrices $\Phi_{\{1,2\}}$ and $\Phi_{\{3,4\}}$ are computed and sent to the root node(s) ($\ell = 0$), where the inner product $\langle A, C \rangle = \Phi_{\{1,2,3,4\}}$ is finally computed.

Fig. 3.2: Parallel computation of the inner product $\langle A, C \rangle$ for two \mathcal{H} -tensors $A, C \in \mathbb{R}^{\mathcal{I}}$ ($d = 4$). The operations can be parallelized level-wise.

where we used the notation $\mathbf{i}_s = (i_\mu)_{\mu \in s}$ for subsets $s \subset \{1, \dots, d\}$ again. From this (3.5) follows for all levels ℓ of T_d by induction. \square

Using Definition 3.1 and Lemma 3.2, we can compute the inner product $\langle A, C \rangle$ for two \mathcal{H} -tensors $A, C \in \mathbb{R}^{\mathcal{I}}$ by recursively computing the matrices Φ_t for each node $t \in T_d$. The inner product is then obtained on the root node $D = \{1, \dots, d\}$ as $\langle A, C \rangle = \Phi_D$. The matrices Φ_t , $t \in T_d$, are of size $\mathbb{R}^{k_t \times \bar{k}_t}$ and can be computed in $\mathcal{O}(k_{\{\mu\}} \bar{k}_{\{\mu\}} \#\mathcal{I}_\mu)$ for each leaf node $\{\mu\}$, in $\mathcal{O}(k_t k_{s_1} k_{s_2} \bar{k}_{s_1} + k_t k_{s_2} \bar{k}_{s_1} \bar{k}_{s_2} + k_t \bar{k}_t \bar{k}_{s_1} \bar{k}_{s_2})$ for each inner node t with sons(t) = $\{s_1, s_2\}$ and in $\mathcal{O}(k_{t_1} k_{t_2} \bar{k}_{t_1} + k_{t_2} \bar{k}_{t_1} \bar{k}_{t_2} + k_{t_1} \bar{k}_{t_2})$ for the root $D = \{1, \dots, d\}$, where sons(D) = $\{t_1, t_2\}$. With $k := \max\{k_t, \bar{k}_t \mid t \in T_d\}$ this yields a complexity of $\mathcal{O}(k^4)$, where the ordering of the multiplications in (3.4) might be relevant (in the case of inhomogenous ranks k_t , $t \in T_d$).

For each leaf node, the definition $\Phi_{\{\mu\}} = (U_t)^\top V_t$, $\mu = 1, \dots, d$, requires only data stored on the respective leaf $\{\mu\}$; for each non-leaf node $t \in T_d$ the definition (3.4) of Φ_t requires data stored on t as well as the matrices Φ_{s_1} , Φ_{s_2} of the sons. The computation can therefore be parallelized level-wise, as illustrated in Fig. 3.2.

3.3. The truncation of \mathcal{H} -tensors based on the Hierarchical SVD [5]. The truncation of an \mathcal{H} -tensor down to lower Hierarchical ranks is essential for the addition $A + C$ of two \mathcal{H} -tensors and for the application of some operator \mathcal{L} to A , where the operator $\mathcal{L} \in \mathbb{R}^{\mathcal{J} \times \mathcal{I}}$ is as well stored in the Hierarchical Tucker format. Both operations increase the Hierarchical ranks, as it is well known for the addition of matrices and the matrix rank from linear algebra.

In this section we briefly summarize the truncation techniques from [5] to truncate an \mathcal{H} -tensor $A \in \mathbb{R}^{\mathcal{I}}$ down to lower Hierarchical ranks $(k_t)_{t \in T_d}$, either for prescribed ranks or for prescribed accuracy ε . Moreover, we describe how these techniques can be parallelized level-wise with respect to the underlying tree T_d for the case of distributed tensors.

In general the truncation of a tensor $A \in \mathbb{R}^{\mathcal{I}}$ (not necessarily an \mathcal{H} -tensor) to prescribed ranks $(k_t)_{t \in T_d}$ for some tensor tree T_d can be achieved like this (cf. [5]):

1. For all leaves $\{\mu\}$, $\mu = 1, \dots, d$, compute an SVD $\mathcal{M}_{\{\mu\}}(A) = U\Sigma V^\top$ of the matricization $\mathcal{M}_{\{\mu\}}(A)$ and store the left singular vectors corresponding to the $k_{\{\mu\}}$ largest singular values as frame $U_{\{\mu\}}$:

$$U_{\{\mu\}} = U|_{\mathcal{I}_\mu \times \{1, \dots, k_{\{\mu\}}\}} \quad (\text{singular values in descending order}).$$

2. For all inner nodes $t \in T_d$ with sons $(t) = \{s_1, s_2\}$ compute $U_t = U|_{\mathcal{I}_t \times \{1, \dots, k_t\}}$, where $\mathcal{M}_t(A) = U\Sigma V^\top$ as in (1.). Instead of storing the frame U_t we store the transfer tensor $B_t \in \mathbb{R}^{k_t \times k_{s_1} \times k_{s_2}}$, which can be computed by

$$(B_t)_{i,j,\ell} = \langle (U_t)_{-,i}, (U_{s_1})_{-,j} \otimes (U_{s_2})_{-,,\ell} \rangle,$$

where U_{s_1} and U_{s_2} are the respective singular vectors for the sons of t .

3. For the root $D = \{1, \dots, d\}$ with sons $(D) = \{t_1, t_2\}$ compute the transfer tensor B_D by

$$(B_D)_{j,\ell} = \langle \mathcal{M}_D(A), (U_{t_1})_{-,j} \otimes (U_{t_2})_{-,,\ell} \rangle.$$

In the case of large tensors, which are already given in the Hierarchical Tucker format, computing the full matricizations $\mathcal{M}_t(A)$ can easily get impractical. Even the computation of U_t for a non-leaf node $t \in T_d$ could be no more feasible.

We make use of results from [5], which allow for the truncation of an \mathcal{H} -tensor $A \in \mathbb{R}^{\mathcal{I}}$ down to smaller Hierarchical ranks, without explicitly computing $\mathcal{M}_t(A)$ or U_t for non-leaf nodes $t \in T_d$. For now we assume the columns of the frames $U_t \in \mathbb{R}^{\mathcal{I}_t \times k_t}$ to be orthonormal for all non-root nodes $t \in T_d$ (if this is not the case, the tensor can be orthogonalized, cf. Section 3.4). For all non-root nodes $t \in T_d$ one can find a matrix $V_t \in \mathbb{R}^{\mathcal{I}_t \times k_t}$, $[t] = \{1, \dots, d\} \setminus t$, such that the matricization $\mathcal{M}_t(A)$ can be written as

$$\mathcal{M}_t(A) = U_t(V_t)^\top, \tag{3.6}$$

since $\text{range}(U_t) = \text{range}(\mathcal{M}_t(A))$. From [5] we know how to compute the matrices $\widehat{B}_t := (V_t)^\top V_t \in \mathbb{R}^{k_t \times k_t}$ for all $t \in T_d \setminus \{D\}$ in $\mathcal{O}(dk^4)$, $k := \max\{k_t \mid t \in T_d\}$, for the case that the columns of U_t , $t \in T_d \setminus \{D\}$, are orthonormal. If $V_t^\top = Q_t \Sigma_t W_t^\top$ is a singular value decomposition of V_t^\top , then $\widehat{B}_t Q_t = Q_t \Sigma_t^2$, i.e. we get the left singular vectors and singular values of V_t^\top as the eigenvectors and the square roots of the eigenvalues of \widehat{B}_t . Using (3.6) and the fact that the columns of U_t are orthonormal, we get $U_t Q_t \in \mathbb{R}^{\mathcal{I}_t \times k_t}$ as left singular vectors of $\mathcal{M}_t(A)$ with the same singular values. We can therefore truncate from rank k_t down to rank $r_t < k_t$ with respect to $t \in T_d \setminus \{D\}$ by multiplying U_t only with the first r_t columns of Q_t , i.e.

$$U_t \leftarrow U_t \cdot (Q_t|_{\{1, \dots, k_t\} \times \{1, \dots, r_t\}}). \tag{3.7}$$

From the following Lemma 3.3 (cf. [5]) we will see how the transfer tensors have to be transformed in order to achieve (3.7) for each node $t \in T_d \setminus \{D\}$. Since we will use Lemma 3.3 again in Section 3.4, we formulate it for arbitrary invertible matrices Y, Z :

LEMMA 3.3 (Transfer tensor transformation). *Let $A \in \mathbb{R}^{\mathcal{I}}$ be an \mathcal{H} -tensor with underlying tree T_d and let U_t and B_t be defined as in Section 2.1, i.e.*

$$(U_t)_{-,i} = \sum_{j=1}^{k_{s_1}} \sum_{\ell=1}^{k_{s_2}} (B_t)_{i,j,\ell} (U_{s_1})_{-,j} \otimes (U_{s_2})_{-,,\ell}$$

for all columns of U_t , where $t \in T_d$ is a non-leaf node with $\text{sons}(t) = \{s_1, s_2\}$.

Then for matrices $X \in \mathbb{R}^{k_t \times k_t}$, $Y \in \mathbb{R}^{k_{s_1} \times k_{s_1}}$, $Z \in \mathbb{R}^{k_{s_2} \times k_{s_2}}$, where Y, Z are invertible, the transformations $U_t \leftarrow U_t X$, $U_{s_1} \leftarrow U_{s_1} Y$ and $U_{s_2} \leftarrow U_{s_2} Z$ yield the transformed transfer tensor

$$B_{t,\text{new}} = (X^\top, Y^{-1}, Z^{-1}) \circ B_t, \quad \text{i.e.} \quad (B_{t,\text{new}})_{i,j,k} = \sum_{\ell,m,n} X_{\ell,i} (Y^{-1})_{j,m} (Z^{-1})_{k,n} (B^t)_{\ell,m,n},$$

such that

$$(U_t X)_{-,i} = \sum_{j=1}^{k_{s_1}} \sum_{\ell=1}^{k_{s_2}} (B_{t,\text{new}})_{i,j,\ell} (U_{s_1} Y)_{-,j} \otimes (U_{s_2} Z)_{-,,\ell}.$$

Let us first suppose, we transform $U_t \leftarrow U_t Q_t$ in each non-root node $t \in T_d$, where Q_t is the full orthogonal matrix of eigenvectors for $t \in T_d$ from above, i.e. $Q_t^{-1} = Q_t^\top$. Then the transfer tensors would transform like

$$B_{t,\text{new}} = (Q_t^\top, Q_{s_1}^\top, Q_{s_2}^\top) \circ B_t, \quad (3.8)$$

according to Lemma 3.3, where $\text{sons}(t) = \{s_1, s_2\}$. Truncating the tensor down to ranks $(r_t)_{t \in T_d}$ means leaving only the first r_t columns of Q_t in each non-root node $t \in T_d$ and removing the remaining ones (assuming the eigenvalues/singular values to be in descending order). This can directly be carried out in (3.8) by first truncating the matrices Q_t , Q_{s_1} , Q_{s_2} and multiplying afterwards:

$$B_{t,\text{new}} = ((Q_t |_{\{1, \dots, k_t\} \times \{1, \dots, r_t\}})^\top, (Q_{s_1} |_{\{1, \dots, k_{s_1}\} \times \{1, \dots, r_{s_1}\}})^\top, (Q_{s_2} |_{\{1, \dots, k_{s_2}\} \times \{1, \dots, r_{s_2}\}})^\top) \circ B_t. \quad (3.9)$$

For the two sons t_1 and t_2 of the root $D = \{1, \dots, d\}$, the matrices $\widehat{B}_t = (V_t)^\top V_t \in \mathbb{R}^{k_t \times k_t}$ are just

$$\widehat{B}_{t_1} = B_D (B_D)^\top \quad \text{and} \quad \widehat{B}_{t_2} = (B_D)^\top B_D. \quad (3.10)$$

For the sons s_1 and s_2 of a non-root node t , the computation of \widehat{B}_{s_1} and \widehat{B}_{s_2} involves the matrix \widehat{B}_t as well as the transfer tensor B_t (cf. [5]):

$$(\widehat{B}_{s_1})_{p,q} = \sum_{i_2=1}^{k_t} \sum_{\ell=1}^{k_{s_2}} \sum_{i_1=1}^{k_t} (B_t)_{i_1,p,\ell} (\widehat{B}_t)_{i_1,i_2} (B_t)_{i_2,q,\ell} \quad \text{for } p, q = 1, \dots, k_{s_1} \quad \text{and} \quad (3.11)$$

$$(\widehat{B}_{s_2})_{p,q} = \sum_{i_2=1}^{k_t} \sum_{j=1}^{k_{s_1}} \sum_{i_1=1}^{k_t} (B_t)_{i_1,j,p} (\widehat{B}_t)_{i_1,i_2} (B_t)_{i_2,j,q} \quad \text{for } p, q = 1, \dots, k_{s_2}. \quad (3.12)$$

Notice that the equations from (3.10) can be written as (3.11) and (3.12) if we set $\widehat{B}_D = 1 \in \mathbb{R}^{1 \times 1}$.

The computation of the matrices \widehat{B}_t can thus be parallelized level-wise with respect to the tree T_d , where the computations start at the root node, which computes both matrices from (3.10) and sends them to its sons t_1 and t_2 . Subsequently any non-root node $t \in T_d$ receives the matrix \widehat{B}_t from its father and, if t is not a leaf, computes the matrices \widehat{B}_{s_1} and \widehat{B}_{s_2} for its sons s_1 and s_2 according to (3.11) and (3.12).

Once the matrices \widehat{B}_t are computed for all non-root nodes $t \in T_d$, the truncation can be started on the highest level of T_d . Again, all nodes of one level work entirely independent of each other:

1. For all leaves $\{\mu\}$, $\mu = 1, \dots, d$: Compute the eigenvectors $Q_{\{\mu\}}$ of $\widehat{B}_{\{\mu\}}$ together with the corresponding eigenvalues. Overwrite the frames $U_{\{\mu\}}$ with the truncated singular vectors: $U_{\{\mu\}} \leftarrow U_{\{\mu\}}(Q_{\{\mu\}} |_{\{1, \dots, k_{\{\mu\}}\} \times \{1, \dots, r_{\{\mu\}}\}})$. Send the truncated eigenvector matrix $Q_{\{\mu\}} |_{\{1, \dots, k_{\{\mu\}}\} \times \{1, \dots, r_{\{\mu\}}\}}$ to the father.
2. For all inner nodes $t \in T_d$ with sons $(t) = \{s_1, s_2\}$ receive the truncated matrices $Q_{s_1} |_{\{1, \dots, k_{s_1}\} \times \{1, \dots, r_{s_1}\}}$ and $Q_{s_2} |_{\{1, \dots, k_{s_2}\} \times \{1, \dots, r_{s_2}\}}$ from the sons s_1 and s_2 , compute Q_t , truncate it to $Q_t |_{\{1, \dots, k_t\} \times \{1, \dots, r_t\}}$, send the truncated matrix to the father, and update the transfer tensor B_t according to (3.9).
3. For the root $D = \{1, \dots, d\}$: Receive the truncated matrices $Q_{t_1} |_{\{1, \dots, k_{t_1}\} \times \{1, \dots, r_{t_1}\}}$ and $Q_{t_2} |_{\{1, \dots, k_{t_2}\} \times \{1, \dots, r_{t_2}\}}$ from the sons t_1, t_2 of D and transform B_D like

$$B_D \leftarrow Q_{t_1}^\top B_D Q_{t_2}.$$

This shows that the truncation of \mathcal{H} -tensors down to smaller ranks can be parallelized level-wise.

Note that we assumed the columns of U_t to be orthonormal for all non-root nodes $t \in T_d$ in order to get the matrices \widehat{B}_t for each non-root node t by (3.10), (3.11) and (3.12) (cf. [5]). Furthermore we can only take $U_t Q_t$ as left singular vectors of $\mathcal{M}_t(A)$ if the columns of U_t are orthonormal. If this is not fulfilled, the tensor must be orthogonalized first, cf. Section 3.4.

Further notice that after a truncation the columns of U_t , $t \in T_d \setminus \{D\}$, will probably no longer be orthonormal (one can easily construct counterexamples where $B_{t, \text{new}}$ from (3.9) does not fulfill the respective condition, i.e. the columns of $\mathcal{M}_{2,3}(B_{t, \text{new}})$ are not orthonormal). This, however, does not affect the truncation algorithm which first computes the matrices \widehat{B}_t and the corresponding eigenvectors Q_t for all non-root nodes $t \in T_d$ and afterwards starts truncating. Imagine the truncation to be processed from the root to the leaves, then each non-root node $t \in T_d$ is first truncated, before its orthogonality gets destroyed by the transformations induced by its sons. Since the order of the multiplications in (3.9) does not play a role, we may also start the truncations at the leaf nodes.

3.4. The orthogonalization of \mathcal{H} -tensors. For the truncation of an \mathcal{H} -tensor $A \in \mathbb{R}^{\mathcal{I}}$ down to smaller Hierarchical ranks, as described in Section 3.3, the frames $U_t \in \mathbb{R}^{\mathcal{I}_t \times k_t}$ are required to have orthonormal columns for all non-root nodes t of the underlying dimension tree T_d . An \mathcal{H} -tensor $A \in \mathbb{R}^{\mathcal{I}}$ fulfilling this property, is called *orthogonal*:

DEFINITION 3.4 (Orthogonal frames, orthogonal \mathcal{H} -tensor). *Let $A \in \mathbb{R}^{\mathcal{I}}$ be an \mathcal{H} -tensor with underlying dimension tree T_d .*

A frame U_t is called orthogonal, if its columns are orthonormal.

The \mathcal{H} -tensor A is called orthogonal, if each non-root frame U_t , $t \in T_d \setminus \{\{1, \dots, d\}\}$ is orthogonal.

For the leaves $\{\mu\}$, $\mu = 1, \dots, d$, the frames $U_{\{\mu\}}$ are stored explicitly, i.e. we can directly compute a reduced QR-decomposition $U_{\{\mu\}} = Q_{\{\mu\}}R_{\{\mu\}}$, keep the orthogonal factor $Q_{\{\mu\}}$ and adjust the transfer tensor of the father according to Lemma 3.3. For $t = \{\mu, \nu\}$ we do

$$U_{\{\mu\}} \leftarrow Q_{\{\mu\}} (= U_{\{\mu\}}R_{\{\mu\}}^{-1}) \quad \text{and} \quad U_{\{\nu\}} \leftarrow Q_{\{\nu\}} (= U_{\{\nu\}}R_{\{\nu\}}^{-1})$$

and update

$$B_t \leftarrow (I, R_{\{\mu\}}, R_{\{\nu\}}) \circ B_t,$$

where $I \in \mathbb{R}^{k_t \times k_t}$ is the unit matrix.

Now consider an inner node $t \in T_d$, where the frames U_{s_1} and U_{s_2} of the sons s_1, s_2 have already been orthogonalized. Then U_t being orthogonal is equivalent to

$$\sum_{j=1}^{k_{s_1}} \sum_{\ell=1}^{k_{s_2}} (B_t)_{i,j,\ell} (B_t)_{i',j,\ell} = \delta_{i,i'} \quad \text{with the Kronecker delta } \delta_{i,i'},$$

i.e. the orthogonality of the matrix $\mathcal{M}_{\{2,3\}}(B_t)$ (cf. Definition 2.1). For the inner nodes $t \in T_d$ we can thus compute a reduced QR-decomposition of the matrix $\mathcal{M}_{\{2,3\}}(B_t)$, i.e.

$$(B_t)_{i,j,\ell} = \sum_{q=1}^k (Q_t)_{(j,\ell),q} (R_t)_{q,i}, \quad \text{where } k = \min\{k_t, k_{s_1} \cdot k_{s_2}\}, \quad (3.13)$$

and then update the transfer tensor B_f of the father according to Lemma 3.3.

Altogether we have the following algorithm for the orthogonalization of an \mathcal{H} -tensor $A \in \mathbb{R}^{\mathcal{I}}$:

1. For all leaves $\{\mu\}$, $\mu = 1, \dots, d$: Compute a reduced QR-decomposition $U_{\{\mu\}} = Q_{\{\mu\}}R_{\{\mu\}}$ and overwrite $U_{\{\mu\}} \leftarrow Q_{\{\mu\}}$. Send the right factor $R_{\{\mu\}}$ to the father.
2. For all inner nodes $t \in T_d$ with sons $(t) = \{s_1, s_2\}$ receive the factors R_{s_1} and R_{s_2} from the sons and update the transfer tensor: $B_t \leftarrow (I, R_{s_1}, R_{s_2}) \circ B_t$. Then compute a reduced QR-decomposition of $\mathcal{M}_{\{2,3\}}(B_t)$ like (3.13) and store $(B_t)_{q,j,\ell} \leftarrow (Q_t)_{(j,\ell),q}$, with the indices named as in (3.13). The factor R_t is sent to the father.
3. For the root $D = \{1, \dots, d\}$: Receive the factors R_{t_1}, R_{t_2} of the sons t_1, t_2 and update the transfer tensor $B_D \leftarrow R_{t_1} B_D R_{t_2}^T$.

During the algorithm each non-leaf node $t \in T_d$ requires the right factors R_{s_1} and R_{s_2} of its sons, but works independently of all other nodes on its level. Therefore the orthogonalization of an \mathcal{H} -tensor can be parallelized level-wise with respect to the underlying tree T_d .

Depending on the maximal rank $k := \max\{k_t \mid t \in T_d\}$, the complexity on each node can be estimated by $\mathcal{O}(k^4)$, since the complexity of the QR decomposition of $\mathcal{M}_{\{2,3\}}(B_t)$ can be estimated by $\mathcal{O}(k_t^2 k_{s_1} k_{s_2})$ and the complexity of the multiplication $B_t \leftarrow (I, R_{s_1}, R_{s_2}) \circ B_t$ can be estimated by $\mathcal{O}(k_t k_{s_1}^2 k_{s_2} + k_t k_{s_1} k_{s_2}^2)$. However, in our runtime tests the complexity seems to behave more like $\mathcal{O}(k^3)$, which might be due to optimizations in the LAPACK routines which we are using.

3.5. The addition of two \mathcal{H} -tensors. For the sum $A + C$ of two \mathcal{H} -tensors $A, C \in \mathbb{R}^{\mathcal{I}}$, based on the same dimension tree T_d ,

$$(A + C)_{(i_1, \dots, i_d)} = A_{(i_1, \dots, i_d)} + C_{(i_1, \dots, i_d)} \quad \text{for } (i_1, \dots, i_d) \in \mathcal{I},$$

we assume A and C to be distributed among the same compute nodes, i.e. the data of A , which belongs to a certain node of the underlying tree T_d , is stored on the same compute node as the corresponding data of C . Due to this assumption we do not need any communication for the assembling of the tensor $A + C$ in the Hierarchical format. Otherwise the respective nodes would have to communicate. The assembling of $A + C$ by itself (without subsequent truncation) does not even require any arithmetic operations, as the following Lemma 3.5 illustrates, a proof of which can e.g. be found in [11].

LEMMA 3.5 (The sum of two \mathcal{H} -tensors). *Let $A \in \mathbb{R}^{\mathcal{I}}$ and $C \in \mathbb{R}^{\mathcal{I}}$ be two \mathcal{H} -tensors, based on the same dimension tree T_d . For A we use the previous notation B_t for the transfer tensors and U_t for the frames. The respective objects of C are denoted by F_t and V_t . Furthermore we write $(k_t)_{t \in T_d}$ for the Hierarchical ranks of A and $(\bar{k}_t)_{t \in T_d}$ for those of C . For a non-leaf node t we write $\text{sons}(t) = \{s_1, s_2\}$.*

Then the sum $A + C$ can again be represented in the Hierarchical Tucker format with transfer tensors $H_t \in \mathbb{R}^{(k_t + \bar{k}_t) \times (k_{s_1} + \bar{k}_{s_1}) \times (k_{s_2} + \bar{k}_{s_2})}$ for all inner nodes $t \in T_d$, $H_D \in \mathbb{R}^{(k_{t_1} + \bar{k}_{t_1}) \times (k_{t_2} + \bar{k}_{t_2})}$ for the root $D = \{1, \dots, d\}$, where $\text{sons}(D) = \{t_1, t_2\}$, and leaf frames $W_{\{\mu\}} \in \mathbb{R}^{\mathcal{I}_{\{\mu\}} \times (k_{\{\mu\}} + \bar{k}_{\{\mu\}})}$, $\mu = 1, \dots, d$, defined as follows:

$$(W_{\{\mu\}})_{-j} = \begin{cases} (U_{\{\mu\}})_{-j} & , \quad 1 \leq j \leq k_{\{\mu\}} \\ (V_{\{\mu\}})_{-(j - k_{\{\mu\}})} & , \quad k_{\{\mu\}} + 1 \leq j \leq k_{\{\mu\}} + \bar{k}_{\{\mu\}} \end{cases} \quad \text{for } \mu = 1, \dots, d, \quad (3.14)$$

$$(H_t)_{i,j,\ell} = \begin{cases} (B_t)_{i,j,\ell} & , \quad \begin{cases} 1 \leq i \leq k_t, \\ 1 \leq j \leq k_{s_1}, \\ 1 \leq \ell \leq k_{s_2} \end{cases} \\ (F_t)_{i - k_t, j - k_{s_1}, \ell - k_{s_2}} & , \quad \begin{cases} k_t + 1 \leq i \leq k_t + \bar{k}_t, \\ k_{s_1} + 1 \leq j \leq k_{s_1} + \bar{k}_{s_1}, \\ k_{s_2} + 1 \leq \ell \leq k_{s_2} + \bar{k}_{s_2} \end{cases} \\ 0 & , \quad \text{otherwise} \end{cases} \quad t \in T_d \text{ inner node,}$$

$$(H_D)_{j,\ell} = \begin{cases} (B_D)_{j,\ell} & , \quad 1 \leq j \leq k_{t_1} \text{ and } 1 \leq \ell \leq k_{t_2} \\ (F_D)_{j - k_{t_1}, \ell - k_{t_2}} & , \quad k_{t_1} + 1 \leq j \leq k_{t_1} + \bar{k}_{t_1} \text{ and } k_{t_2} + 1 \leq \ell \leq k_{t_2} + \bar{k}_{t_2} \\ 0 & , \quad \text{otherwise.} \end{cases}$$

According to Lemma 3.5 the sum $A + C$ is built by just putting both \mathcal{H} -tensors A and C together (cf. Fig. 3.3). This process can be run in parallel on each node of the tree and does not need any communication, under the above assumption. In the end, since we do not have to compute anything for the addition, we can of course let the tensors A and C be stored, as they are, and just modify the evaluation routine to evaluate $A + C$ for a given index (i_1, \dots, i_d) instead of A or C .

However, if a large number of additions is needed (eg. for some iterative method), we would like to truncate the sum back to lower ranks (cf. Section 3.3), to save storage. This can be done either with respect to prescribed ranks $(k_t)_{t \in T_d}$ or for prescribed accuracy ε of the truncated tensor.

3.6. The application of an operator to an \mathcal{H} -tensor. Let $\mathcal{I} = \mathcal{I}_1 \times \dots \times \mathcal{I}_d$ and $\mathcal{J} = \mathcal{J}_1 \times \dots \times \mathcal{J}_d$ be two product index sets of dimension $d \in \mathbb{N}$ and consider a tensor $A \in \mathbb{R}^{\mathcal{I}}$ as well as a matrix $\mathcal{L} \in \mathbb{R}^{\mathcal{J} \times \mathcal{I}}$. The matrix \mathcal{L} can as well be seen as a tensor of dimension d by ordering the dimensions like

$$\mathcal{L} \in \mathbb{R}^{(\mathcal{J}_1 \times \mathcal{I}_1) \times \dots \times (\mathcal{J}_d \times \mathcal{I}_d)}, \quad \text{i.e. } \mathcal{L}_{(j_1, i_1), \dots, (j_d, i_d)} \quad \text{instead of } \mathcal{L}_{(j_1, \dots, j_d), (i_1, \dots, i_d)}, \quad (3.15)$$

and regarding each of the indices (j_μ, i_μ) , $\mu = 1, \dots, d$, as one index. We assume both tensors \mathcal{L} and A to be represented in the Hierarchical Tucker format and call them \mathcal{H} -

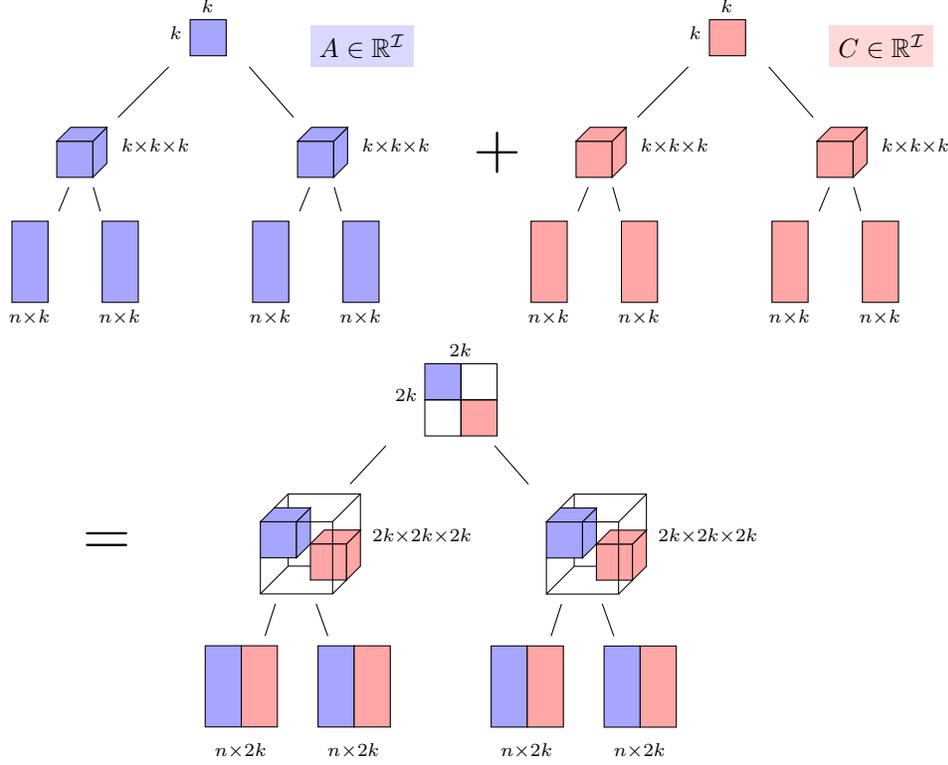


Fig. 3.3: The sum of two \mathcal{H} -tensors $A, C \in \mathbb{R}^{\mathcal{I}}$ is assembled by merging A and C together, cf. Lemma 3.5. The white spaces in the transfer tensors represent zero blocks. The sum can be assembled in parallel on each node of the tree.

operator and \mathcal{H} -tensor. Then the tensor $C := \mathcal{L}A \in \mathbb{R}^{\mathcal{J}}$, i.e.

$$C_{(j_1, \dots, j_d)} = \sum_{(i_1, \dots, i_d)} \mathcal{L}_{(j_1, i_1), \dots, (j_d, i_d)} A_{(i_1, \dots, i_d)}, \quad (3.16)$$

can be computed directly in the Hierarchical Tucker format. The following Lemma 3.6 describes the representation of the \mathcal{H} -tensor $C := \mathcal{L}A$. A proof can be found in [11].

LEMMA 3.6 (The application of an \mathcal{H} -operator to an \mathcal{H} -tensor). *Let $\mathcal{I} = \mathcal{I}_1 \times \dots \times \mathcal{I}_d$ and $\mathcal{J} = \mathcal{J}_1 \times \dots \times \mathcal{J}_d$ be two product index sets of dimension d and let $A \in \mathbb{R}^{\mathcal{I}}$ be an \mathcal{H} -tensor and $\mathcal{L} \in \mathbb{R}^{\mathcal{J} \times \mathcal{I}}$ an \mathcal{H} -operator. Regarding \mathcal{L} as well as tensor of dimension d , cf. (3.15), we suppose A and \mathcal{L} to be both represented in the Hierarchical Tucker format with the same underlying tree T_d . For A we use the previous notation B_t for the transfer tensors and U_t for the frames. The respective objects of \mathcal{L} are denoted by H_t and W_t . Furthermore we write $(k_t)_{t \in T_d}$ for the Hierarchical ranks of A and $(\bar{k}_t)_{t \in T_d}$ for those of \mathcal{L} . For a non-leaf node t we write $\text{sons}(t) = \{s_1, s_2\}$.*

Then the tensor $C := \mathcal{L}A \in \mathbb{R}^{\mathcal{J}}$ can again be represented in the Hierarchical Tucker format with transfer tensors $F_t \in \mathbb{R}^{k_t \bar{k}_t \times k_{s_1} \bar{k}_{s_1} \times k_{s_2} \bar{k}_{s_2}}$ for all non-leaf nodes $t \in T_d$ ($k_D = \bar{k}_D = 1$ for the root D), and leaf frames $V_{\{\mu\}} \in \mathbb{R}^{\mathcal{J}_\mu \times k_{\{\mu\}} \bar{k}_{\{\mu\}}}$, $\mu = 1, \dots, d$, defined as

follows:

$$\begin{aligned}
F_t &= H_t \otimes B_t, \quad \text{i.e.} \quad (F_t)_{(p,i),(q,j),(r,\ell)} = (H_t)_{p,q,r} (B_t)_{i,j,\ell} \\
&\quad (i = 1, \dots, k_t, j = 1, \dots, k_{s_1}, \ell = 1, \dots, k_{s_2}, \\
&\quad p = 1, \dots, \bar{k}_t, q = 1, \dots, \bar{k}_{s_1}, r = 1, \dots, \bar{k}_{s_2}) \quad \text{and} \\
(V_{\{\mu\}})_{j_\mu,(r,\ell)} &= \sum_{i_\mu \in \mathcal{I}_\mu} (W_{\{\mu\}})_{(j_\mu,i_\mu),r} \cdot (U_{\{\mu\}})_{i_\mu,\ell} \\
&\quad (\ell = 1, \dots, k_{\{\mu\}}, r = 1, \dots, \bar{k}_{\{\mu\}}).
\end{aligned} \tag{3.17}$$

From Lemma 3.6 it follows that the product $\mathcal{L}A$ of an \mathcal{H} -operator \mathcal{L} and an \mathcal{H} -tensor A can be computed in parallel on each node $t \in T_d$ without any communication needed, assuming all data of A and \mathcal{L} belonging to one tree node $t \in T_d$ to be stored on the same compute node.

Since the Hierarchical ranks of $C := \mathcal{L}A$ are the products $k_t \bar{k}_t$ of the respective ranks k_t and \bar{k}_t of A and \mathcal{L} , one would most likely want to truncate the tensor C down to given ranks $(r_t)_{t \in T_d}$ or with respect to some prescribed accuracy ε .

Computing the product $F_t = H_t \otimes B_t$ (cf. Lemma 3.6) needs $\mathcal{O}(k_t \bar{k}_t k_{s_1} \bar{k}_{s_1} k_{s_2} \bar{k}_{s_2})$ operations, i.e. the complexity can be estimated by $\mathcal{O}(k^6)$, $k := \max\{k_t, \bar{k}_t \mid t \in T_d\}$, if the ranks of the \mathcal{H} -tensor A and the \mathcal{H} -operator \mathcal{L} are comparable in size.

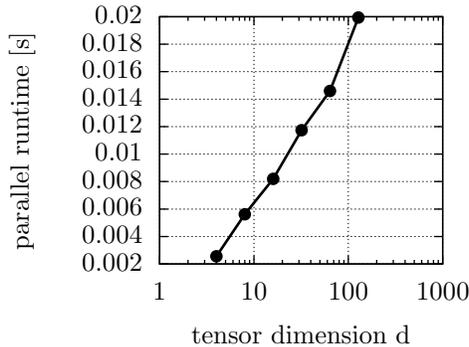
4. Runtime tests. In this section we present some runtime tests. On the one hand we analyze the parallel runtime of our algorithms for varying tensor dimension d . Here we leave the size n of each tensor dimension as well as the rank k ($k_t = k$ for all $t \in T_d \setminus \{1, \dots, d\}$) unchanged. In this case we expect the parallel runtime to grow like $\mathcal{O}(\log(d))$ for the evaluation of tensor entries, the inner product of two \mathcal{H} -tensors, the orthogonalization and the truncation down to lower rank of an \mathcal{H} -tensor.

For the application of an \mathcal{H} -operator to an \mathcal{H} -tensor we would even expect the runtime to be independent of d , since the algorithms can run in parallel on each node (cf. Section 3).

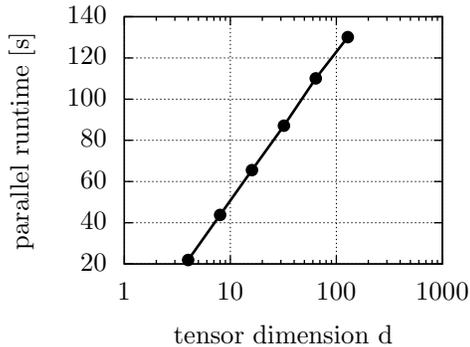
Since the addition of two \mathcal{H} -tensors (without subsequent truncation) does not involve any floating point operation at all (cf. Section 3.5), we do not cover it here. The addition of two \mathcal{H} -tensors together with the truncation behaves like the truncation itself.

Furthermore we analyze the dependence of the parallel runtime on the tensor rank k . For that we leave d and n unchanged while we vary the rank k . Here we expect an $\mathcal{O}(k^4)$ dependence for the inner product (cf. Section 3.2), for the truncation of an orthogonalized \mathcal{H} -tensor (cf. Section 3.3), as well as for the orthogonalization of an \mathcal{H} -tensor (cf. Section 3.4). For the evaluation of tensor entries we expect an $\mathcal{O}(k^3)$ complexity (cf. Section 3.1). For the application of an \mathcal{H} -operator we expect $\mathcal{O}(k^6)$, where also the ranks of the \mathcal{H} -operator are chosen to equal k on each node except for the root node (cf. Sections 3.6).

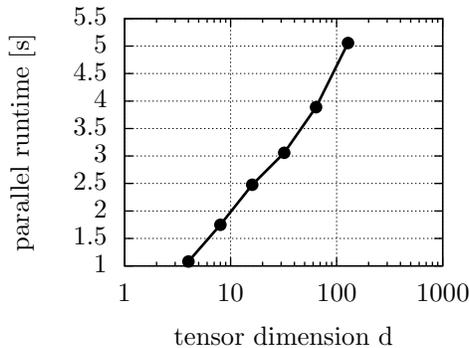
4.1. Parallel runtime for varying tensor dimension d . We choose random \mathcal{H} -tensors where each tensor dimension is of size $n = 10\,000$ and all ranks are of size $k = 100$, i.e. we choose the transfer tensors and the leaf frames as random tensors/matrices. The choice of $n = 10\,000$ and $k = 100$ has initially been made to have the same amount of data on each compute node (except for the root node): $n \cdot k = k^3$, cf. (2.3). However, as long as we choose the rank k large enough, we observe the same behavior of the parallel runtime. The experiments demonstrate that the parallel runtime of our algorithms scales as expected for the evaluation of tensor entries, the inner product of two \mathcal{H} -tensors and for the truncation down to lower rank of an \mathcal{H} -tensor including the orthogonalization. (see Fig 4.1).



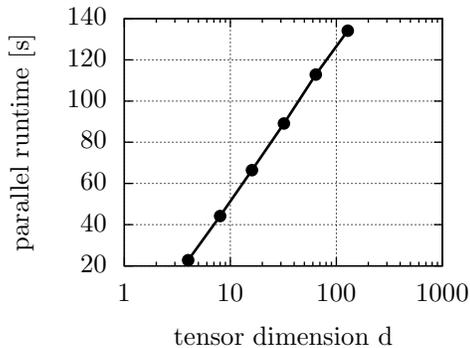
(a) Evaluation of one tensor entry.



(b) Computation of the inner product.



(c) Orthogonalization.



(d) Truncation of the orthogonalized \mathcal{H} -tensor.

Fig. 4.1: Parallel runtimes for random \mathcal{H} -tensors. All tensor dimensions are of the same size $n = 10000$ and the tensor rank on each node (except the root node) is $k = 100$. The parallel runtime of the algorithms for the above operations scales like $\log(d)$.

From Fig. 4.2 we see that the parallel runtime for the application of an \mathcal{H} -operator to an \mathcal{H} -tensor is indeed rather independent of the tensor dimension d .

4.2. Parallel runtime for varying tensor rank k . For varying tensor rank k most of our algorithms reach the predicted complexity estimates (see Fig. 4.3). Only the orthogonalization of an \mathcal{H} -tensor behaves more like $\mathcal{O}(k^3)$ rather than the expected $\mathcal{O}(k^4)$. This might be due to optimizations in the LAPACK routines which we are using.

The ranks for the \mathcal{H} -operator are always chosen equal to those of the \mathcal{H} -tensor, which the operator is applied to.

5. Application to iterative methods. Consider an \mathcal{H} -operator $\mathcal{A} \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$ together with a suitable \mathcal{H} -tensor $B \in \mathbb{R}^{\mathcal{I}}$ as right-hand side, where $\mathcal{I} = \mathcal{I}_1 \times \dots \times \mathcal{I}_d$ is a product index set of dimension $d \in \mathbb{N}$. Then we can use the parallel algorithms for \mathcal{H} -tensor arithmetic (cf. Section 3) to solve $\mathcal{A}X = B$, $X \in \mathbb{R}^{\mathcal{I}}$, directly in the Hierarchical Tucker format.

In this section we apply a multigrid method to a diffusion equation ("*cookie problem*",

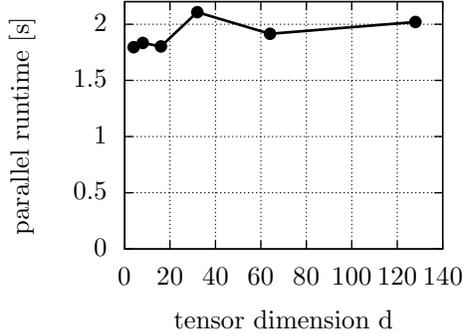


Fig. 4.2: The parallel runtime for the application of an \mathcal{H} -operator to an \mathcal{H} -tensor seems to be independent of the tensor dimension d ($n = 400$, $k = 20$).

see Section 5.1), which depends on 9 parameters. A Finite Element discretizations leads to a parameter-dependent linear system which in our case can directly be transferred to a linear equation in the Hierarchical Tucker format. All arithmetic operations are therefore carried out between tensors in the Hierarchical Tucker format. As a smoother for the multigrid method we choose a Richardson method since it can easily be transferred to the Hierarchical Tucker format. On the coarsest grid level we use a CG algorithm to solve the defect equation. Each parameter will correspond to one tensor dimension. Together with the spacial component of the problem we will thus end up with tensors of dimension $d = 10$.

5.1. The cookie problem. Inspired by [12] we take the *cookie problem*

$$\begin{cases} -\nabla(\sigma(x)\nabla u) = 1 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (5.1)$$

as a test problem, where our "cookies" are squares (instead of circles):

In the domain $\Omega = [0, 7] \times [0, 7]$ we define 9 rectangles ("cookies")

$$\mathcal{D}_\mu = \{x \in \Omega : \|x - c_\mu\|_\infty < 0.5\}, \quad \mu = 1, \dots, 9,$$

with midpoints

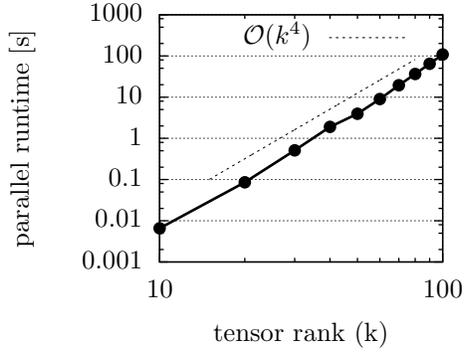
$$\begin{aligned} c_1 &= (1.5, 1.5), & c_2 &= (3.5, 1.5), & c_3 &= (5.5, 1.5), \\ c_4 &= (1.5, 3.5), & c_5 &= (3.5, 3.5), & c_6 &= (5.5, 3.5), \\ c_7 &= (1.5, 5.5), & c_8 &= (3.5, 5.5), & c_9 &= (5.5, 5.5). \end{aligned}$$

The diffusion coefficient $\sigma(x)$, $x \in \Omega$, is defined as

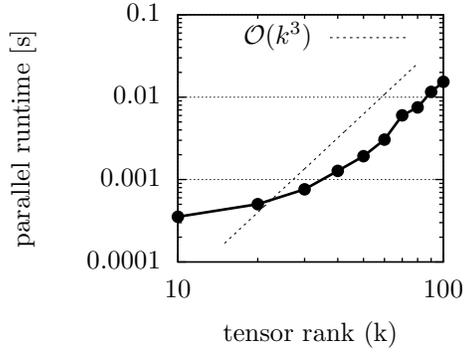
$$\sigma(x) = \begin{cases} 1 + \alpha_\mu & \text{if } x \in \mathcal{D}_\mu, \quad \mu = 1, \dots, 9, \\ 1 & \text{if } x \notin \bigcup_{\mu=1}^9 \mathcal{D}_\mu, \end{cases} \quad (5.2)$$

i.e. depending on 9 parameters α_μ , $\mu = 1, \dots, 9$ (cf. Fig. 5.1). A Finite Element discretization of (5.1) with (5.2) leads to a parameter-dependent linear system

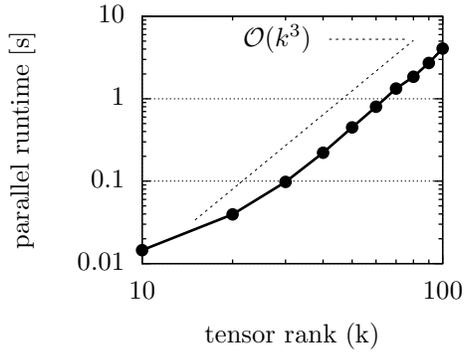
$$A(\alpha_1, \alpha_2, \dots, \alpha_9) \cdot x(\alpha_1, \alpha_2, \dots, \alpha_9) = b, \quad (5.3)$$



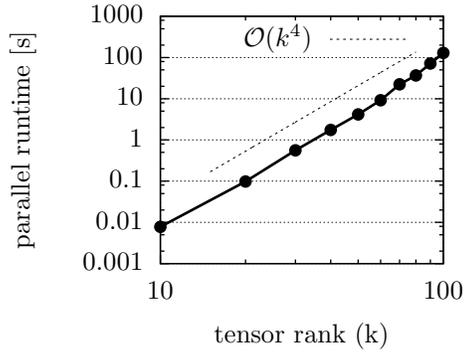
(a) Dot product for $n = 10\,000$.



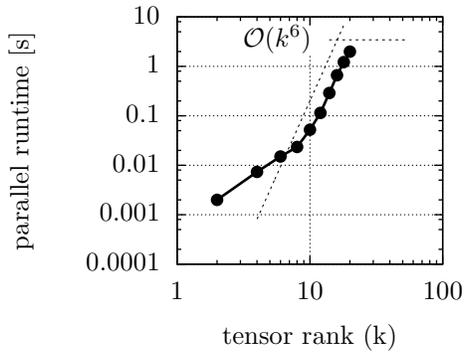
(b) Evaluation for $n = 10\,000$.



(c) Orthogonalization for $n = 10\,000$.



(d) Truncation for $n = 10\,000$.



(e) Operator for $n = 400$.

Fig. 4.3: Parallel runtimes for tensor dimension $d = 64$, fixed sizes n of the tensor dimensions and varying ranks k .

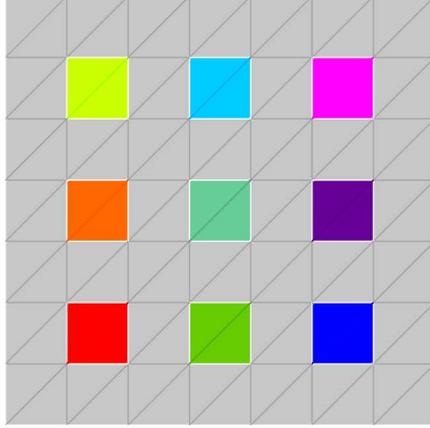


Fig. 5.1: Domain $\Omega = [0, 7] \times [0, 7]$ of the cookie problem. The cookies \mathcal{D}_μ , $\mu = 1, \dots, 9$, are squares of side length 1. Outside the cookies (grey area) the diffusion coefficient equals one: $\sigma(x) \equiv 1$. On each cookie the diffusion coefficient is constant and controlled by a parameter: $\sigma(x) \equiv 1 + \alpha_\mu$, if $x \in \mathcal{D}_\mu$, $\mu = 1, \dots, 9$. Grid and image generated with ProMesh, www.promesh3d.com.

where the right-hand side b is the same for each parameter combination $(\alpha_1, \alpha_2, \dots, \alpha_9)$.

Due to the affine structure of the operator in (5.1) with (5.2), the matrix in (5.3) can directly be represented as an \mathcal{H} -operator: Let $A_0 \in \mathbb{R}^{n \times n}$ be the stiffness matrix of a Finite Element discretization of (5.1) with $\sigma \equiv 1$ and let $A_\mu \in \mathbb{R}^{n \times n}$, $\mu = 1, \dots, 9$, be the stiffness matrices which would result from a Finite Element discretization of (5.1) on the same Finite Element space but with $\sigma \equiv \mathbf{1}_{\mathcal{D}_\mu}$ instead of $\sigma \equiv 1$, where $\mathbf{1}_{\mathcal{D}_\mu} : \Omega \rightarrow \{0, 1\}$ is the indicator function of \mathcal{D}_μ . Since the diffusion coefficient (5.2) can be decomposed as

$$\sigma \equiv 1 + \sum_{\mu=1}^9 \alpha_\mu \mathbf{1}_{\mathcal{D}_\mu},$$

the matrix $A(\alpha_1, \alpha_2, \dots, \alpha_9)$ can be decomposed as

$$A(\alpha_1, \alpha_2, \dots, \alpha_9, i) = A_0(i) + \alpha_1 A_1(i) + \alpha_2 A_2(i) + \dots + \alpha_9 A_9(i), \quad i = 1, \dots, n^2, \quad (5.4)$$

where we combined the two matrix indices to one index i ($\mathbb{R}^{n \times n} \cong \mathbb{R}^{n^2}$). Assuming discretizations

$$a_{\mu, i_\mu}, \quad i_\mu = 1, \dots, n_\mu, \quad \mu = 1, \dots, 9,$$

for the parameters, we can represent (5.4) in the Hierarchical Tucker format, where we take the underlying tree T_{10} as the balanced tree described in Section 2. We choose the first 9 leaf frames $U_{\{\mu\}} \in \mathbb{R}^{n_\mu \times 10}$, $\mu = 1, \dots, 9$, which correspond to the 9 parameter directions, as

$$(U_{\{\mu\}})_{i_\mu, j} = \begin{cases} 1 & \text{if } j \neq \mu + 1 \\ \alpha_{\mu, i_\mu} & \text{if } j = \mu + 1 \end{cases}, \quad i_\mu = 1, \dots, n_\mu, \quad j = 1, \dots, 10, \quad \mu = 1, \dots, 9. \quad (5.5)$$

The last leaf frame $U_{\{10\}} \in \mathbb{R}^{n^2 \times 10}$ just consists of the matrices A_0, \dots, A_9 (written as columns):

$$(U_{10})_{i_{10}, j} = A_{j-1}(i_{10}), \quad i_{10} = 1, \dots, n^2, \quad j = 1, \dots, 10.$$

The transfer tensors are diagonal tensors with ones on the diagonal.

One easily verifies that the above construction yields a representation of the parameter dependent matrix A as \mathcal{H} -operator. This representation is, however, not minimal: Each of the leaf frames in (5.5) contains nine times the column $(1, \dots, 1)^\top$. By only keeping this column once per leaf frame (and adjusting the transfer tensors accordingly), we can always achieve Hierarchical ranks of 2 for these leaf frames.

We thus get an equation $\mathcal{A}X = B$ in the Hierarchical Tucker format, where the \mathcal{H} -operator has Hierarchical ranks $k_t = \text{rank}(\mathcal{M}_t(\mathcal{A}))$, $t \in T_d$, of at most $d = 10$ and the right-hand side B is of rank 1, since it only depends on the spacial variable and not on the parameters. The solution tensor X then contains the solutions of (5.3) for all possible parameter combinations $(\alpha_1, \alpha_2, \dots, \alpha_9)$ and allows for further arithmetic operations in the Hierarchical Tensor format, as e.g. computing averages over spacial domains or over parameters, or computing the expected value or the variance with respect to a probability distribution of the parameters, where the probability distribution has to be represented/approximated in the Hierarchical Tucker format (which is very easy for the case of independent parameters).

5.2. The CG method in parallel \mathcal{H} -tensor arithmetic. We use a CG algorithm (Algorithm 1) to solve the equation $\mathcal{A}X = B$ of the cookie problem (cf. Section 5.1) on the coarsest grid, which is shown in Fig. 5.1. We let each parameter a_μ , $\mu = 1, \dots, 9$, vary between 0.5 and 1.5, using 10 equidistant grid points in $[0.5, 1.5]$. The full tensor X on the coarsest grid would therefore consist of

$$\#(\text{inner grid points}) \cdot 10^9 = 36 \cdot 10^9 \text{ entries,}$$

whereas in the Hierarchical Tucker format, the number of entries can be bounded by

$$\begin{aligned} & (\# \text{inner grid points} + 9 \cdot 10) \cdot k_{\max} + \#(\text{inner tree nodes}) \cdot k_{\max}^3 + k_{\max}^2 \\ & = (36 + 9 \cdot 10) \cdot k_{\max} + 8 \cdot k_{\max}^3 + k_{\max}^2 = 126 \cdot k_{\max} + 8 \cdot k_{\max}^3 + k_{\max}^2 \text{ entries,} \end{aligned}$$

where k_{\max} is an upper bound for the Hierarchical ranks. For $k_{\max} = 50$ these would be about 10^6 entries.

In the CG algorithm (Algorithm 1) we use our parallel algorithms for arithmetic operations between \mathcal{H} -tensors (cf. Section 3). After each addition of two \mathcal{H} -tensors and after each application of the \mathcal{H} -operator \mathcal{A} to an \mathcal{H} -tensor we truncate the result down to lower rank, where we prescribe some accuracy ε such that the truncation error is smaller than ε .

The following Lemma 5.1 considers the simplified case, where after each CG step in Algorithm 1 ($j = 0, 1, 2, \dots$) the iterate X_j is truncated, i.e. truncations in between are neglected. For this simplified case we can reach any accuracy $\delta > 0$ of the approximate solution (i.e. $\|X - X_j\| < \delta$ for some j) if we choose the upper bound $\varepsilon > 0$ for the truncation error small enough. Similar results have already been published in [13] for the general case of convergent iterative processes in Banach spaces.

LEMMA 5.1 (Convergence of the CG algorithm including truncation). *Consider the linear equation $\mathcal{A}X = B$ with symmetric and positive-definite \mathcal{A} . Let \mathcal{C} denote the operation of performing one exact CG step according to the linear system $\mathcal{A}X = B$, i.e. when X_j is the result of j exact CG steps, $X_{j+1} = \mathcal{C}X_j$ would be the next iterate in the exact CG method. Furthermore let \mathcal{T}_ε be the truncation down to lower ranks with error smaller than ε .*

For arbitrary $\delta > 0$ and arbitrary starting tensor X_0 we can achieve

$$\|X - (\mathcal{T}_\varepsilon \mathcal{C})^j X_0\|_{\mathcal{A}} < \delta$$

Data: operator \mathcal{A} , right-hand side B in the Hierarchical Tucker format

Result: Approximate solution X in the Hierarchical Tucker format

$X_0 = B$;

$R_0 = B - \mathcal{A}X_0$;

$R_0 = \mathcal{T}(R_0)$;

$D_0 = R_0$;

$j = 0$;

while $\sqrt{\langle R_j, R_j \rangle} > EPS$ **and** $j < MAX_CG_STEPS$ **do**

$Z = \mathcal{A}D_j$;

$Z = \mathcal{T}(Z)$;

$\alpha_j = \langle R_j, R_j \rangle / \langle D_j, Z \rangle$;

$X_{j+1} = X_j + \alpha_j D_j$;

$X_{j+1} = \mathcal{T}(X_{j+1})$;

$R_{j+1} = R_j - \alpha_j Z$;

$R_{j+1} = \mathcal{T}(R_{j+1})$;

$\beta_j = \langle R_{j+1}, R_{j+1} \rangle / \langle R_j, R_j \rangle$;

$D_{j+1} = \beta_j D_j + R_{j+1}$;

$D_{j+1} = \mathcal{T}(D_{j+1})$;

$j = j + 1$;

end

$X = X_j$;

Algorithm 1: CG algorithm to solve $\mathcal{A}X = B$ in the Hierarchical Tucker format including truncation. The truncation \mathcal{T} can be carried out either down to fixed rank k ($\mathcal{T} = \mathcal{T}_k$) or with prescribed accuracy ε ($\mathcal{T} = \mathcal{T}_\varepsilon$). The stopping criterion depends on a lower bound EPS for the norm of the residual and on an upper bound MAX_CG_STEPS for the number of CG steps.

for some $j \in \mathbb{N}$, if we choose

$$\varepsilon < \delta/C$$

with $C := \sqrt{\|\mathcal{A}\|}(\text{cond}(\mathcal{A}) + 1)$, where $\|\mathcal{A}\| = \lambda_{\max}$ is the spectral norm (largest eigenvalue) of \mathcal{A} , $\text{cond}(\mathcal{A}) = \lambda_{\max}/\lambda_{\min}$ is the condition number (ratio between largest and smallest eigenvalue) of \mathcal{A} and $\|\cdot\|_{\mathcal{A}}$ is the \mathcal{A} -norm, defined as $\|X\|_{\mathcal{A}} := \sqrt{\langle X, \mathcal{A}X \rangle}$.

Proof. For the exact CG algorithm one can show (cf. [14])

$$\|X - \mathcal{C}^j X_0\|_{\mathcal{A}} \leq \gamma \|X - \mathcal{C}^{j-1} X_0\|_{\mathcal{A}} \quad (5.6)$$

with $\gamma = (\text{cond}(\mathcal{A}) - 1)/(\text{cond}(\mathcal{A}) + 1)$.

We now split the error of one perturbed CG step into the error of one exact CG step and one truncation error:

$$\|X - (\mathcal{T}_\varepsilon \mathcal{C})X_0\|_{\mathcal{A}} \leq \|X - \mathcal{C}X_0\|_{\mathcal{A}} + \|\mathcal{C}X_0 - \mathcal{T}_\varepsilon \mathcal{C}X_0\|_{\mathcal{A}}.$$

The first error can be estimated by (5.6). The second error is first transformed to the standard norm $\|\cdot\|$ (Euclidian norm/Frobenius norm), which we use to measure the truncation error, since it can then be bounded by ε :

$$\begin{aligned} \|X - (\mathcal{T}_\varepsilon \mathcal{C})X_0\|_{\mathcal{A}} &\leq \gamma \|X - X_0\|_{\mathcal{A}} + \sqrt{\lambda_{\max}} \|\mathcal{C}X_0 - \mathcal{T}_\varepsilon \mathcal{C}X_0\| \\ &\leq \gamma \|X - X_0\|_{\mathcal{A}} + \varepsilon \sqrt{\lambda_{\max}}. \end{aligned}$$

Similarly (by induction) we can get a bound for the error of j perturbed CG steps:

$$\begin{aligned}
\|X - (\mathcal{T}_\varepsilon \mathcal{C})^j X_0\|_{\mathcal{A}} &\leq \gamma \|X - (\mathcal{T}_\varepsilon \mathcal{C})^{j-1} X_0\|_{\mathcal{A}} + \varepsilon \sqrt{\lambda_{\max}} \\
&\leq \gamma (\gamma \|X - (\mathcal{T}_\varepsilon \mathcal{C})^{j-2} X_0\|_{\mathcal{A}} + \varepsilon \sqrt{\lambda_{\max}}) + \varepsilon \sqrt{\lambda_{\max}} \\
&= \gamma^2 \|X - (\mathcal{T}_\varepsilon \mathcal{C})^{j-2} X_0\|_{\mathcal{A}} + \varepsilon \sqrt{\lambda_{\max}} (\gamma + 1) \\
&\leq \dots \leq \gamma^j \|X - X_0\|_{\mathcal{A}} + \varepsilon \sqrt{\lambda_{\max}} \left(\sum_{\ell=0}^{j-1} \gamma^\ell \right).
\end{aligned}$$

The sum can be bounded by $(1 - \gamma)^{-1}$, thus we have

$$\|X - (\mathcal{T}_\varepsilon \mathcal{C})^j X_0\|_{\mathcal{A}} \leq \gamma^j \|X - X_0\|_{\mathcal{A}} + \varepsilon \sqrt{\lambda_{\max}} (1 - \gamma)^{-1}.$$

Since $\gamma^j \rightarrow 0$ for $j \rightarrow \infty$, we can find $j \in \mathbb{N}$ such that $\gamma^j \|X - X_0\|_{\mathcal{A}} < \delta/2$. To achieve $\|X - (\mathcal{T}_\varepsilon \mathcal{C})^j X_0\|_{\mathcal{A}} < \delta$ we thus have to choose ε in such a way that

$$\varepsilon \sqrt{\lambda_{\max}} (1 - \gamma)^{-1} < \frac{\delta}{2} \quad \Leftrightarrow \quad \varepsilon < \frac{\delta}{C} \quad \text{with} \quad C = \sqrt{\lambda_{\max}} (\text{cond}(\mathcal{A}) + 1),$$

which completes the proof. \square

REMARK 5.2.

1. Lemma 5.1 only states the convergence of the perturbed CG algorithm, i.e. that in principal we can reach arbitrary accuracy (absolute error of the computed solution) if the absolute truncation error is bounded by a sufficiently small ε . Lemma 5.1 does not reveal the convergence rate of such an algorithm. The estimate

$$\|X - \mathcal{C}^j X_0\|_{\mathcal{A}} \leq \left(\frac{\text{cond}(\mathcal{A}) - 1}{\text{cond}(\mathcal{A}) + 1} \right)^j \|X - X_0\|_{\mathcal{A}}, \quad (5.7)$$

which we used in the proof, is quite pessimistic. This bound can already be achieved for the method of steepest descent, where the residual R_{j+1} of one step is typically only orthogonal to the search direction D_j of the last step (cf. Algorithm 1 for notation). In the CG method, however, this residual is orthogonal to the search directions of all previous steps, which yields the (actually much better) estimate

$$\|X - \mathcal{C}^j X_0\|_{\mathcal{A}} \leq 2 \left(\frac{\sqrt{\text{cond}(\mathcal{A})} - 1}{\sqrt{\text{cond}(\mathcal{A})} + 1} \right)^j \|X - X_0\|_{\mathcal{A}}. \quad (5.8)$$

In the proof of Lemma 5.1 we used (5.7) instead of (5.8), since we were interested in the decay of the error for one single CG step (i.e. $j = 1$), where (5.8) can not be used for.

Moreover, in practice, the Euclidian norm of the error instead of the \mathcal{A} -norm will be of interest, which would give us an additional factor $1/\sqrt{\lambda_{\min}} = \sqrt{\|\mathcal{A}^{-1}\|}$ for the constant C .

Much more precise analysis of perturbed CG methods can be found in [14] or [9], where the perturbed CG algorithm is reinterpreted as an exact CG algorithm for a perturbed operator.

2. Lemma 5.1 assumes only one truncation of the iterate X_j in each CG step. In practice we truncate after each addition of two \mathcal{H} -tensors as well as after each application of \mathcal{A} to an \mathcal{H} -tensor (cf. Algorithm 1). This includes also truncations

of the residuals R_j and the search directions D_j , which themselves are afterwards involved in a quotient of scalar products. A precise analysis would therefore have to deal with the respective propagation of the errors inside the algorithm, cf. [9], [14].

3. For our example (\mathcal{H} -operator \mathcal{A} for the coarsest grid of the cookie problem with 10 equidistant points in $[0.5, 1.5]$ per parameter) we can estimate $\lambda_{\max} \approx 11$ and $\lambda_{\min} \approx 0.4$, which would yield $C \approx 95$ in Lemma 5.1, and $C \approx 150$ if we measure the error in the Euclidian norm instead of the \mathcal{A} -norm.

In our numerical experiments we used $\varepsilon = 10^{-4}$ as upper bound for the absolute error of each truncation. Since the Hierarchical ranks can temporarily get rather large during the CG algorithm (cf. [1]), we prescribed additional bounds for the maximal rank, by which we implicitly choose a larger ε . In Fig. 5.2 we plotted the relative residual norm $\|\mathcal{A}X_j - B\|/\|B\|$ within 25 CG steps with different bounds for the maximal rank. One observes the principle of Lemma 5.1: the higher the rank bound (i.e. the smaller we choose ε), the smaller the relative residual norm, which can be reached in the CG algorithm.

The plotted residual norms are not the norms of the "residuals" R_j computed in the perturbed CG algorithm, since these can deviate from the exact residual $\mathcal{A}X_j - B$ (cf. [9]). We therefore computed the residual norm after each CG step, i.e. computed $\sqrt{\langle \mathcal{A}X_j - B, \mathcal{A}X_j - B \rangle}$ in \mathcal{H} -tensor arithmetic.

When applying our CG algorithms with fixed rank bounds on finer grids, we need more iterations until we achieve the maximal attainable accuracy with respect to the rank bound, as one would expect it due to the growing condition number of the operator \mathcal{A} . In order to overcome this, one could use preconditioning. We instead apply a multigrid algorithm with respect to the spacial component of the problem (semi-coarsening) to solve $\mathcal{A}X = B$ on finer grids, which is described in Section 5.3.

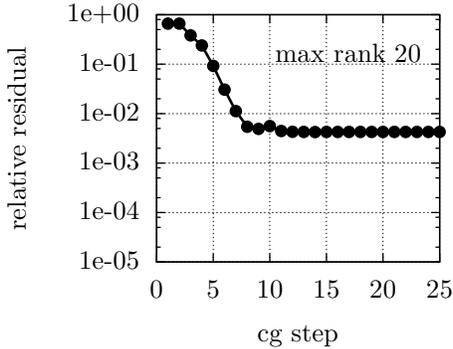
Multigrid methods in the Hierarchical Tucker format have already been studied in [1], where the number of iterations needed to solve the Poisson equation up to some relative residual turned out to be independent of the tensor dimension d (i.e. the number of parameters). This would make them usable for problems depending on a huge number of parameters.

5.3. A multigrid method in parallel \mathcal{H} -tensor arithmetic. Multigrid methods have already been applied for low-rank matrices and \mathcal{H} -matrices in [6] and [4] in order to solve large scale Sylvester and Riccati equations. In [1] multigrid methods have been applied in \mathcal{H} -tensor arithmetic to solve equations on high dimensional domains.

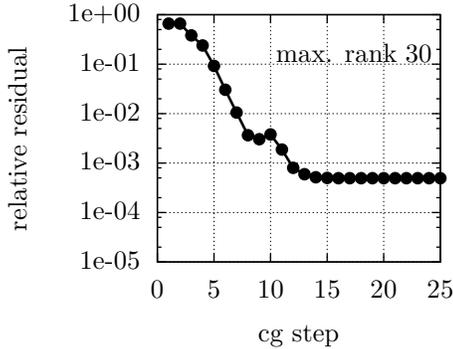
In order to compute a solution of higher resolution, we refine the coarse grid of Fig. 5.1 three times. With every refinement step we add the midpoints of all edges as new grid points. By this we get the grid hierarchy of Fig. 5.3, where the finest grid consist of 3025 inner grid points.

One multigrid step (Algorithm 3) starts on the finest grid and then recursively proceeds to the next coarser grid to solve the system $\mathcal{A}E = D$, where D is the restriction of the residual $B - \mathcal{A}X$ to the next coarser grid. The solution E is then prolonged back to the finer grid and added to the current iterate to get the next iterate. For the multigrid method to work well we need to be able to compute good approximations of the error on the next coarser grid. These can be obtained by several steps of a Richardson method (Algorithm 2), which serves as a smoother for the error. On the coarsest level we use the CG algorithm (cf. Section 5.2) to solve the system. An introduction to multigrid algorithms together with convergence analysis can e.g. be found in [10].

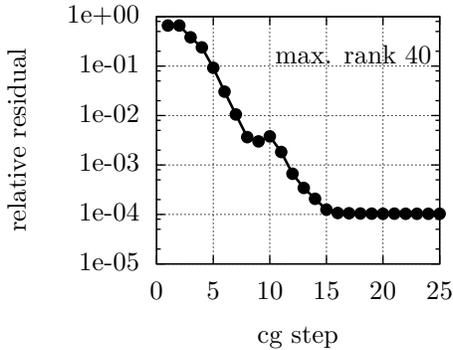
For linear methods (as the Richardson method and the multigrid method), one can obtain similar results to Lemma 5.1: the best attainable accuracy of the solution depends



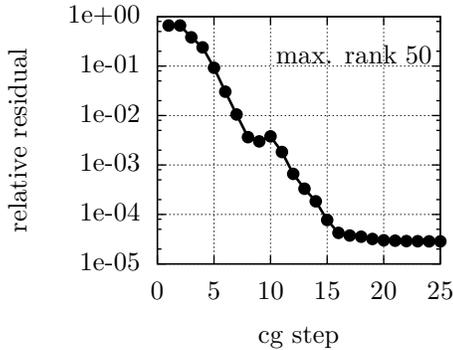
(a) Maximal rank $k = 20$.



(b) Maximal rank $k = 30$.



(c) Maximal rank $k = 40$.



(d) Maximal rank $k = 50$.

Fig. 5.2: CG algorithm using \mathcal{H} -arithmetic for the cookie problem on the coarsest grid, which is depicted in Fig. 5.1. For each of the nine parameters we use 10 equidistant grid points in $[\frac{1}{2}, \frac{3}{2}]$. After each addition of \mathcal{H} -tensors and after each application of the \mathcal{H} -operator \mathcal{A} to an \mathcal{H} -tensor we truncate the resulting tensor down to lower ranks with an absolute error less than 10^{-4} . In addition we prescribe a maximal rank k , since the involved computations have complexity up to $\mathcal{O}(k^4)$.

on the accuracy of the truncations (see also [13]).

We performed 10 multigrid steps for the equation $\mathcal{A}X = B$ on the finest grid (3025 inner grid points, cf. Fig. 5.3) with different upper bounds for the Hierarchical ranks. In Fig. 5.4 we show the relative residual norms for upper rank bounds of 30 and 50. One observes the linear convergence of the multigrid method down to some lower threshold of the relative residual norm. This threshold decreases with increasing rank bounds, which is shown in Fig. 5.5.

6. Conclusions. In this article we presented parallel algorithms which we developed for basic arithmetic operations between distributed tensors in the Hierarchical Tucker format. Throughout the article we assumed the tensor data to be distributed according to the underlying tree of the Hierarchical Tucker format: The data for each tree node is stored on its own compute node. Since our algorithms typically need communication between

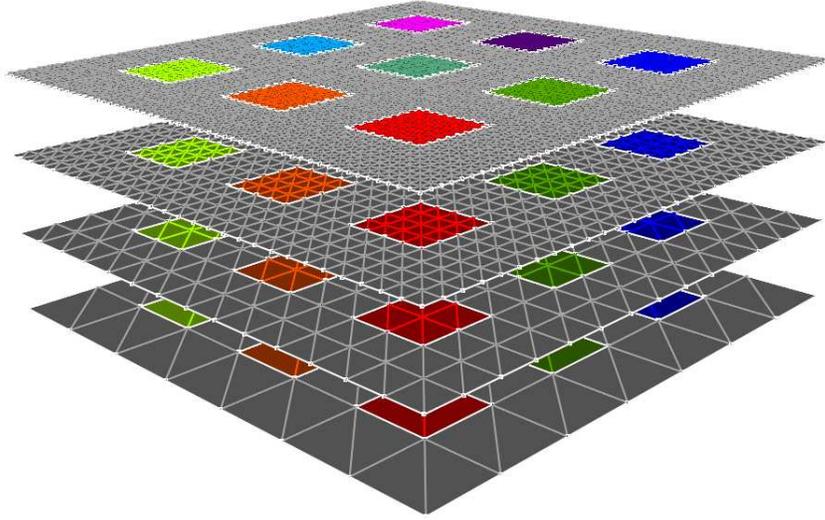
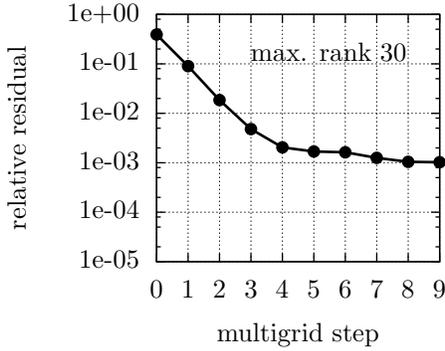
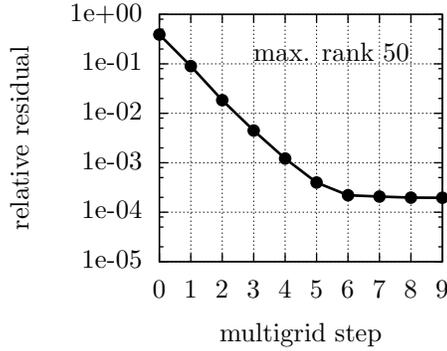


Fig. 5.3: Grid hierarchy for the multigrid method with 36 inner grid points on the coarsest grid and 3025 inner grid points on the finest grid. Grid and image generated with ProMesh, www.promesh3d.com.



(a) Maximal rank $k = 30$.



(b) Maximal rank $k = 50$.

Fig. 5.4: Multigrid method using \mathcal{H} -arithmetic for the cookie problem on the grid hierarchy from Fig. 5.3, once with maximal rank $k = 30$ and once with maximal rank $k = 50$. On the coarsest grid we apply the CG solver from Section 5.2.

different tree levels, we minimized the number of tree levels by choosing a balanced tree. For a tensor of dimension d the number of tree levels grows like $\log(d)$, i.e. a level-wise parallelization of our algorithms yields a parallel runtime, which grows like $\log(d)$. We validated the expected $\log(d)$ -growth of the parallel runtime by several runtime tests in Section 4. The $\log(d)$ -growth of the parallel runtime makes our algorithms applicable for high-dimensional problems which will be part of future work.

Despite the $\log(d)$ -growth of the parallel runtime, we still have a complexity of at least

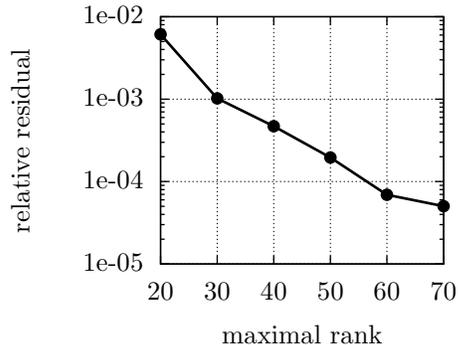


Fig. 5.5: Dependence of the relative residual norm on the upper rank bound in the multigrid method. For each rank bound $20, 30, \dots, 70$ we plotted the best accuracy we obtained after 10 multigrid steps.

$\mathcal{O}(k^3)$ for each compute node (cf. Section 4), where $k \in \mathbb{N}$ is the Hierarchical rank for the respective node. We would therefore expect a further reduction of the runtime by using shared memory parallelization on each node (e.g. `OpenMP`).

In this article we used a parallelized version of the Hierarchical SVD from [5] to truncate tensors in the Hierarchical Tucker format. Since this algorithm involves the computation of $B^\top B$ and BB^\top for matrices B , it may lead to a loss of accuracy. It will be interesting in future research to transfer the techniques from [3, 15] to the Hierarchical SVD.

One interesting application is parameter-dependent problems, which we illustrated by a model problem in Section 5. We applied a multigrid method including our parallel algorithms for \mathcal{H} -arithmetic to a diffusion equation which depends on 9 parameters. Our experiments indicate that the relative residual norm in the iterative method is bounded from below by some barrier which depends on the accuracy of the \mathcal{H} -tensor truncations during the method.

For larger problems preconditioners should be applied, when using the CG method (cf. [12]), which we left out here. Moreover it could be advantageous to apply multilevel sampling techniques, which were introduced in [2].

Our parallel algorithms can as well be used for the postprocessing of (distributed) tensors in the Hierarchical Tucker format, which might have been assembled by some other method (e.g. a sampling method). This could be of interest in the field of uncertainty quantification and could be used for parameter fitting, i.e. adapting model parameters based on some quantity of interest of the solution, which is either known or can be estimated.

REFERENCES

- [1] JONAS BALLANI AND LARS GRASEDYCK, *A projection method to solve linear systems in tensor format*, Numer. Linear Algebra Appl., 20 (2013), pp. 27–43.
- [2] JONAS BALLANI, DANIEL KRESSNER, AND MICHAEL D. PETERS, *Multilevel tensor approximation of PDEs with random data*, Stochastics and Partial Differential Equations: Analysis and Computations, 5 (2017), pp. 400–427.
- [3] SIMON ETTER, *Parallel ALS Algorithm for Solving Linear Systems in the Hierarchical Tucker Representation*, SIAM Journal on Scientific Computing, 38 (2016), pp. A2585–A2609.
- [4] LARS GRASEDYCK, *Nonlinear multigrid for the solution of large scale Riccati equations in low-rank and \mathcal{H} -matrix format*, Numerical linear algebra with applications, 15 (2008), pp. 779–807.

- [5] LARS GRASEDYCK, *Hierarchical Singular Value Decomposition of Tensors*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2029–2054.
- [6] LARS GRASEDYCK AND WOLFGANG HACKBUSCH, *A multigrid method to solve large scale Sylvester equations*, SIAM journal on matrix analysis and applications, 29 (2007), pp. 870–894.
- [7] LARS GRASEDYCK, DANIEL KRESSNER, AND CHRISTINE TOBLER, *A literature survey of low-rank tensor approximation techniques*, GAMM-Mitteilungen, 36 (2013), pp. 53–78.
- [8] LARS GRASEDYCK, RONALD KRIEMANN, CHRISTIAN LÖBBERT, ARNE NÄGEL, GABRIEL WITTUM, AND KONSTANTINOS XYLOURIS, *Parallel tensor sampling in the Hierarchical Tucker format*, Computing and Visualization in Science, 17 (2015), pp. 67–78.
- [9] A. GREENBAUM, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra and its Applications, 113 (1989), pp. 7 – 63.
- [10] WOLFGANG HACKBUSCH, *Multi-Grid Methods and Applications*, vol. 4 of Springer series in computational mathematics, Springer, Heidelberg, 1985.
- [11] ———, *Tensor spaces and numerical tensor calculus*, vol. 42 of Springer series in computational mathematics, Springer, Heidelberg, 2012.
- [12] DANIEL KRESSNER AND CHRISTINE TOBLER, *Low-rank tensor Krylov subspace methods for parametrized linear systems*, SIAM Journal on Matrix Analysis and Applications, 32 (2011), pp. 1288–1316.
- [13] HERMANN G. MATTHIES AND ELMAR ZANDER, *Solving stochastic systems with low-rank tensor compression*, Linear Algebra and its Applications, 436 (2012), pp. 3819 – 3838.
- [14] G. MEURANT, *The Lanczos and conjugate gradient algorithms*, Society for Industrial and Applied Mathematics, 2006.
- [15] E. M. SToudenMIRE AND STEVEN R. WHITE, *Real-space parallel density matrix renormalization group*, Phys. Rev. B, 87 (2013), p. 155137.

Data: iterate X_j

Result: new iterate X_{j+1}

$Z = \mathcal{A}X_j$;

$Z = \mathcal{T}(Z)$;

$Z = Z - B$;

$Z = \mathcal{T}(Z)$;

$X_{j+1} = X_j - \omega Z$;

$X_{j+1} = \mathcal{T}(X_{j+1})$;

Algorithm 2: $X_{j+1} = \text{richardson_step}(\mathcal{A}, X_j, B)$ performs one step of the Richardson method for $\mathcal{A}X = B$ in \mathcal{H} -tensor arithmetic, including truncation. We use this method as a smoother for the multigrid method (Algorithm 3). The truncation \mathcal{T} can be carried out either down to fixed rank k ($\mathcal{T} = \mathcal{T}_k$) or with prescribed accuracy ε ($\mathcal{T} = \mathcal{T}_\varepsilon$). The optimal choice for ω is $\omega_{opt} = 2/(\lambda_{\min} + \lambda_{\max})$, where λ_{\min} and λ_{\max} are the minimal and the maximal eigenvalues of the operator \mathcal{A} .

Data: iterate X_j

Result: new iterate X_{j+1}

if on coarsest grid **then**

 Solve $\mathcal{A}X_{j+1} = B$ with CG (Algorithm 1);

else

 5× presmoothing: $X_j = \text{richardson_step}(\mathcal{A}, X_j, B)$ (Algorithm 2);

$R = B - \mathcal{A}X_j$;

$R = \mathcal{T}(R)$;

$D = \text{restriction}(R)$;

$E = D$;

$E = \text{multigrid_step}(\mathcal{A}, E, D)$;

$P = \text{prolongation}(E)$;

$X_{j+1} = X_j + P$;

$X_{j+1} = \mathcal{T}(X_{j+1})$;

 5× postsmoothing: $X_{j+1} = \text{richardson_step}(\mathcal{A}, X_{j+1}, B)$ (Algorithm 2);

end

Algorithm 3: One multigrid step, $X_{j+1} = \text{multigrid_step}(\mathcal{A}, X_j, B)$, to solve $\mathcal{A}X = B$ in the Hierarchical Tucker format, including truncation. The truncation \mathcal{T} can be carried out either down to fixed rank k ($\mathcal{T} = \mathcal{T}_k$) or with prescribed accuracy ε ($\mathcal{T} = \mathcal{T}_\varepsilon$).