

PLANC: Parallel Low-rank Approximation with Nonnegativity Constraints

SRINIVAS ESWAR and KOBAYASHI, Georgia Institute of Technology

GREY BALLARD, Wake Forest University

RAMAKRISHNAN KANNAN and MICHAEL A. MATHESON, Oak Ridge National Laboratory

HAESUN PARK, Georgia Institute of Technology

We consider the problem of low-rank approximation of massive dense nonnegative tensor data, for example, to discover latent patterns in video and imaging applications. As the size of data sets grows, single workstations are hitting bottlenecks in both computation time and available memory. We propose a distributed-memory parallel computing solution to handle massive data sets, loading the input data across the memories of multiple nodes, and performing efficient and scalable parallel algorithms to compute the low-rank approximation. We present a software package called Parallel Low-rank Approximation with Nonnegativity Constraints, which implements our solution and allows for extension in terms of data (dense or sparse, matrices or tensors of any order), algorithm (e.g., from multiplicative updating techniques to alternating direction method of multipliers), and architecture (we exploit GPUs to accelerate the computation in this work). We describe our parallel distributions and algorithms, which are careful to avoid unnecessary communication and computation, show how to extend the software to include new algorithms and/or constraints, and report efficiency and scalability results for both synthetic and real-world data sets.

Eswar and Hayashi share first authorship.

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy (DOE). This project was partially funded by the Laboratory Director's Research and Development fund. This material is based upon work supported by the National Science Foundation (NSF) under Grants No. OAC-1642385 and No. OAC-1642410. Koby Hayashi acknowledges support from the United States Department of Energy through the Computational Sciences Graduate Fellowship (DOE CSGF) under Grant No. DE-SC0020347.

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the DOE Office of Science (SC). This research used resources of the National Energy Research Scientific Computing Center, a SC User Facility supported by the SC under Contract No. DE-AC02-05CH11231. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF or DOE.

Authors' addresses: S. Eswar, K. Hayashi, and H. Park, Dept. of CSE, Georgia Institute of Technology, Atlanta, GA 30308; emails: {seswar3, khayashi9, hpark}@gatech.edu; G. Ballard, Dept. of CS, Wake Forest University, Winston-Salem, NC 27109; email: ballard@wfu.edu; R. Kannan and M. A. Matheson, Oak Ridge National Laboratory, Oak Ridge, TN 37831; emails: {kannanr, mathesonma}@ornl.gov.

ACM acknowledges that this contribution was authored or co-authored by an employee, contractor, or affiliate of the United States government. As such, the United States government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for government purposes only.

© 2021 Association for Computing Machinery.

0098-3500/2021/06-ART20 \$15.00

<https://doi.org/10.1145/3432185>

CCS Concepts: • **Theory of computation** → **Massively parallel algorithms**; • **Software and its engineering** → **Software libraries and repositories**;

Additional Key Words and Phrases: Tensor factorization, nonnegative least squares, communication-avoiding algorithms

ACM Reference format:

Srinivas Eswar, Koby Hayashi, Grey Ballard, Ramakrishnan Kannan, Michael A. Matheson, and Haesun Park. 2021. PLANC: Parallel Low-rank Approximation with Nonnegativity Constraints. *ACM Trans. Math. Softw.* 47, 3, Article 20 (June 2021), 37 pages.
<https://doi.org/10.1145/3432185>

1 INTRODUCTION

The CP decomposition [11, 23], which is also known as CANDECOMP, PARAFAC, and canonical polyadic decomposition, approximates a tensor, or multidimensional array, by a sum of rank-one tensors. CP is typically used to identify latent factors in data, particularly when the goal is to interpret those hidden patterns, and it is popular within the signal processing, machine learning, and scientific computing communities, among others [1, 21, 38, 56]. Enforcing domain-specific constraints on the computed factors can help to identify interpretable components. We focus in this article on nonnegative dense tensors (when all tensor entries are nonnegative and nearly all of them are positive) and on constraining solutions to have nonnegative entries. Formally, **Nonnegative CP (NCP)**¹ can be defined as

$$\min_{\mathbf{H}^{(i)} \geq 0} \left\| \mathcal{X} - \sum_{r=1}^R \mathbf{H}^{(1)}(:, r) \circ \dots \circ \mathbf{H}^{(N)}(:, r) \right\|^2 \quad (1)$$

for a fixed R , where $\mathbf{H}^{(1)}(:, i) \circ \dots \circ \mathbf{H}^{(N)}(:, i)$ is the outer product of the i^{th} vector from all the N factors that yields a rank-one tensor and $\sum_{r=1}^R \mathbf{H}^{(1)}(:, r) \circ \dots \circ \mathbf{H}^{(N)}(:, r)$ results in a sum of R rank-one tensors that approximate the N th order nonnegative input tensor \mathcal{X} . For example, in imaging and microscopy applications, tensor values often correspond to intensities, and NCP can be used to cluster and analyze the data in a lower-dimensional space [27]. In this work, we consider a brain imaging data set that tracks calcium fluorescence within pixels of a mouse's brain over time during a series of experimental trials [37].

The kernel computations within standard algorithms for computing NCP can be formulated as matrix computations, but the complicated layout of tensors in memory prevents the straightforward use of BLAS and LAPACK libraries. In particular, the matrix formulation of subcomputations involve different views of the tensor data, so no single layout yields a column- or row-major matrix layout for all subcomputations. Likewise, the parallelization approach for tensor methods is not a straightforward application of parallel matrix computation algorithms.

In developing an efficient parallel algorithm for computing a NCP of a dense tensor, the key is to parallelize the bottleneck computation known as Matricized-Tensor Times Khatri-Rao Product (MTTKRP) [2], and a different result is required for each mode of the tensor. The MTTKRP for mode n , for $1 \leq n \leq N$, is defined as

$$\mathbf{M}^{(n)} = \mathbf{X}_{(n)} \left(\mathbf{H}^{(N)} \odot \dots \odot \mathbf{H}^{(n+1)} \odot \mathbf{H}^{(n-1)} \odot \dots \odot \mathbf{H}^{(1)} \right),$$

where $\mathbf{X}_{(n)}$ is a matricization or flattening of the tensor with respect to mode n and \odot is the Khatri-Rao product, or column-wise Kronecker product [2, 38]. The parallelization must load balance the computation, minimize communication across processors, and distribute the results so that the

¹In this article, we use NCP and **Nonnegative Tensor Factorization (NTF)** interchangeably.

rest of the computation can be performed independently. In our algorithm, not only do we load balance the computation, but we also compute and store temporary values that can be used across MTTKRP of different modes using a technique known as dimension trees, significantly reducing the computational cost compared to standard approaches. Our parallelization strategy also avoids communicating tensor entries and minimizes the communication of factor matrix entries, helping the algorithm to remain computation bound and scalable to high processor counts.

We employ a variety of algorithmic strategies to computing NCP, from multiplicative updates to alternating direction method of multipliers. Because the bottleneck computations such as MTTKRP are shared by all update algorithms that compute gradient information, we separate the parallelization strategy for those computations from the (usually local) computations that are unique to each algorithm. In this article, we present an open-source software package called **Parallel Low-rank Approximation with Nonnegativity Constraints (PLANC)** that currently includes six algorithmic options, and we describe how other algorithms can be incorporated into the framework. PLANC can also be used for **nonnegative matrix factorization (NMF)** with dense or sparse matrices. The software is available at <https://github.com/ramkikannan/planc>.

An earlier version of this work appeared as a conference paper [3], which presents the parallelization strategy described in Section 4.1 and the dimension tree optimization detailed in Section 4.2. We summarize the further contributions of this article as follows:

- presentation and description of the open-source PLANC software package,
- utilization of GPUs to alleviate the MTTKRP bottleneck achieving up to $7\times$ speedup over CPU-only execution,
- scaling results for runs from 1 to 16,384 nodes (250,000+ cores) on the Titan supercomputer,
- side by side run-time and convergence results for various update algorithms,
- and new results obtained by applying our code to a mouse brain imaging data set.

2 PRELIMINARIES

2.1 Notation

Tensors will be denoted using Euler script (e.g., \mathcal{T}), matrices will be denoted with uppercase boldface (e.g., \mathbf{M}), vectors will be denoted with lowercase boldface (e.g., \mathbf{v}), and scalars will not be boldface (e.g., s). We use Matlab style notation to index into tensors, matrices, and vectors, and we use 1-indexing. For example, $\mathbf{M}(:, c)$ gives the c th column of the matrix \mathbf{M} .

We use \circ to denote the outer product of two or more vectors. The Hadamard product is the element-wise matrix product and will be denoted using $*$. The Khatri-Rao product, abbreviated KRP, will be denoted with \odot . Given matrices \mathbf{A} and \mathbf{B} that are $I_A \times R$ and $I_B \times R$, the KRP $\mathbf{K} = \mathbf{A} \odot \mathbf{B}$ is $I_A I_B \times R$. It can be thought of as a row-wise Hadamard product, where $\mathbf{K}(i + I_A(j-1), :) = \mathbf{A}(i, :) * \mathbf{B}(j, :)$, or a column-wise Kronecker product, where $\mathbf{K}(:, c) = \mathbf{A}(:, c) \otimes \mathbf{B}(:, c)$.

The CP decomposition of a tensor (also referred to as the CANDECOMP/PARAFAC or canonical polyadic decomposition) is a low-rank approximation of a tensor, where the approximation is a sum of rank-one tensors and each rank-one tensor is the outer product of vectors. We use the notation

$$\mathcal{X} \approx [\![\mathbf{H}^{(1)}, \dots, \mathbf{H}^{(N)}]\!] = \sum_{r=1}^R \mathbf{H}^{(1)}(:, r) \circ \dots \circ \mathbf{H}^{(N)}(:, r)$$

to represent a rank- R CP model, where $\mathbf{H}^{(n)}$ is called a factor matrix and collects the mode- n vectors of the rank-one tensors as columns. The columns of the factor matrices are often normalized, with weights collected into an auxiliary vector $\boldsymbol{\lambda}$ of length R ; in this case, we use the notation $[\![\boldsymbol{\lambda}; \mathbf{H}^{(1)}, \dots, \mathbf{H}^{(N)}]\!]$.

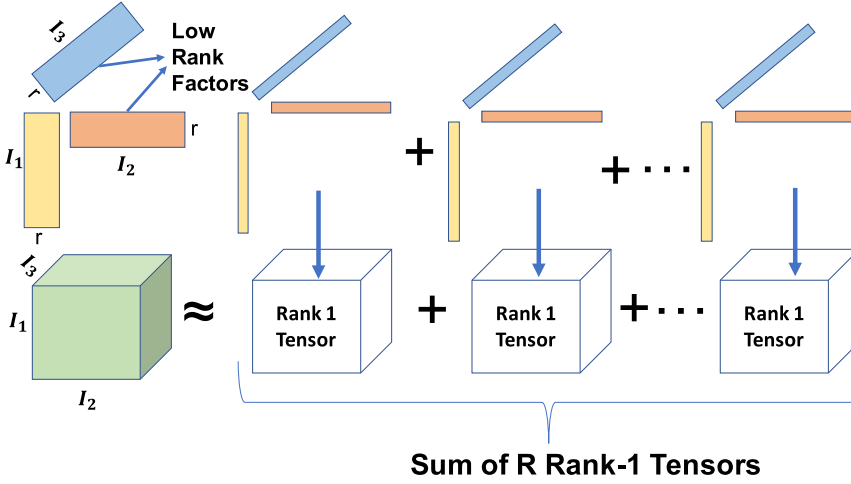


Fig. 1. CP Decomposition.

An NCP constrains the factor matrices to have nonnegative values. In this work, we are interested in NCP models that are good approximations to \mathcal{X} in the least-squares sense. That is, we seek

$$\min_{\mathbf{H}^{(i)} \geq 0} \|\mathcal{X} - \llbracket \boldsymbol{\lambda}; \mathbf{H}^{(1)}, \dots, \mathbf{H}^{(N)} \rrbracket\|,$$

where the tensor norm is a generalization of the matrix Frobenius or vector 2-norm, the square root of the sum of squares of the entries. Here \mathcal{X} is a tensor with dimensions $I_1 \times \dots \times I_N$. We denote the product of its dimensions by I , that is $I = \prod I_n$.

The n th mode matricized tensor denoted by $\mathbf{X}_{(n)}$ is a $I_n \times I/I_n$ matrix formed by organizing the n th mode fibers of \mathcal{X} into the columns of a matrix. The Matricized-Tensor Times Khatri-Rao Product or MTTKRP will be central to this work and takes the form $\mathbf{M}^{(n)} = \mathbf{X}_{(n)} \mathbf{K}^{(n)}$, where $\mathbf{K}^{(n)} = \mathbf{H}^{(N)} \odot \dots \odot \mathbf{H}^{(n+1)} \odot \mathbf{H}^{(n-1)} \odot \dots \odot \mathbf{H}^{(1)}$.

2.2 Nonnegative CP and Alternating-Updating Methods

The CP decomposition is a low-rank approximation of a multi-dimensional array, or tensor, which generalizes matrix approximations like the truncated singular value decomposition. As in Figure 1, CP decomposition approximates the given input matrix as sum of R rank-1 tensors.

Algorithm 1 shows the pseudocode for an alternating-updating algorithm applied to NCP [35]. Lines 11, 12, and 14 compute matrices involved in the gradients of the subproblem objective functions, and Line 13 uses those matrices to update the current factor matrix.

The **nonnegative least-squares (NLS)**-Update in Line 13 can be implemented in different ways. In a faithful **Block Coordinate Descent (BCD)** algorithm, the subproblems are solved exactly; in this case, the subproblem is a nonnegative linear least-squares problem, which is convex. We can use the **Block Principal Pivoting (BPP)** method [35, 36], which is an active-set-like method, to solve the subproblem exactly.

However, as discussed by Kannan et al. [29] for the matrix case, there are other reasonable alternatives to updating the factor matrix without solving the N -block coordinate descent subproblem exactly. For example, we can more efficiently update individual columns of the factor matrix as is done in the **Hierarchical Alternating Least Squares (HALS)** method [14]. In this case, the update rule is

$$\mathbf{H}^{(n)}(:, r) \leftarrow \left[\mathbf{H}^{(n)}(:, r) + \mathbf{M}^{(n)}(:, r) - (\mathbf{H}^{(n)} \mathbf{S}^{(n)})(:, r) \right]_+,$$

ALGORITHM 1: $\llbracket \mathbf{H}^{(1)}, \dots, \mathbf{H}^{(N)} \rrbracket = \text{NCP}(\mathcal{X}, R)$ **Require:** \mathcal{X} is $I_1 \times \dots \times I_N$ tensor, R is approximation rank

```

1: % Initialize data
2: for  $n = 2$  to  $N$  do
3:   Initialize  $\mathbf{H}^{(n)}$ 
4:    $\mathbf{G}^{(n)} = \mathbf{H}^{(n)\top} \mathbf{H}^{(n)}$ 
5: end for
6: % Compute NCP approximation
7: while stopping criteria not satisfied do
8:   % Perform outer iteration
9:   for  $n = 1$  to  $N$  do
10:    % Compute new factor matrix in  $n$ th mode
11:     $\mathbf{M}^{(n)} = \text{MTTKRP}(\mathcal{X}, \{\mathbf{H}^{(1)}, \dots, \mathbf{H}^{(n-1)}, \mathbf{H}^{(n+1)}, \dots, \mathbf{H}^{(N)}\}, n)$ 
12:     $\mathbf{S}^{(n)} = \mathbf{G}^{(1)} * \dots * \mathbf{G}^{(n-1)} * \mathbf{G}^{(n+1)} * \dots * \mathbf{G}^{(N)}$ 
13:     $\mathbf{H}^{(n)} = \text{NLS-Update}(\mathbf{S}^{(n)}, \mathbf{M}^{(n)})$ 
14:     $\mathbf{G}^{(n)} = \mathbf{H}^{(n)\top} \mathbf{H}^{(n)}$ 
15:   end for
16: end while

```

Ensure: $\mathcal{X} \approx \llbracket \mathbf{H}^{(1)}, \dots, \mathbf{H}^{(N)} \rrbracket$

which involves the same matrices $\mathbf{M}^{(n)}$ and $\mathbf{S}^{(n)}$ as BPP. Other possible alternating-updating methods include **Alternating-Optimization Alternating Direction Method of Multipliers (AO-ADMM)** [25, 58] and Nesterov-based algorithms [43]. The details of each of these algorithms are presented in Section 4.3. The parallel algorithm presented in this article is generally agnostic to the approach used to solve the nonnegative least squares subproblems, as all these methods are bottlenecked by the subroutine they have in common, the MTTKRP.

2.3 Parallel Communication Model

To analyze our algorithms, we use the α - β - γ model of distributed-memory parallel computation. In this model, interprocessor communication occurs in the form of messages sent between two processors across a bidirectional link (we assume a fully connected network). We model the cost of a message of size n words as $\alpha + n\beta$, where α is the per-message latency cost and β is the per-word bandwidth cost. Each processor can compute floating point operations (flops) on data that resides in its local memory; γ is the per-flop computation cost. With this communication model, we can predict the performance of an algorithm in terms of the number of flops it performs as well as the number of words and messages it communicates. For simplicity, we will ignore the possibilities of overlapping computation with communication in our analysis. For more details on the α - β - γ model, see the articles by Thakur et al. [63] and Chan et al. [12].

2.4 MPI Collectives

Point-to-point messages can be organized into collective communication operations that involve more than two processors. MPI provides an interface to the most commonly used collectives like broadcast, reduce, and gather, as the algorithms for these collectives can be optimized for particular network topologies and processor characteristics. Chan et al. [12, Figure 1] provide a concise description of the most common collectives. The algorithms we consider use the all-gather, reduce-scatter, and all-reduce collectives, so we review them here, along with their costs. Our analysis

assumes optimal collective algorithms are used [12, 63], though our implementation relies on the underlying MPI implementation.

At the start of an all-gather collective, each of p processors owns data of size n/p . After the all-gather, each processor owns a copy of the entire data of size n . The cost of an all-gather is $\alpha \cdot \log p + \beta \cdot \frac{p-1}{p}n$. At the start of a reduce-scatter collective, each processor owns data of size n . After the reduce-scatter, each processor owns a subset of the sum over all data, which is of size n/p . This single collective is a more efficient way of implementing a reduce followed by a scatter. (Note that the reduction can be computed with other associative operators besides addition.) The cost of an reduce-scatter is $\alpha \cdot \log p + (\beta + \gamma) \cdot \frac{p-1}{p}n$. At the start of an all-reduce collective, each processor owns data of size n . After the all-reduce, each processor owns a copy of the sum over all data, which is also of size n . The cost of an all-reduce is $2\alpha \cdot \log p + (2\beta + \gamma) \cdot \frac{p-1}{p}n$. Note that the costs of each of the collectives are zero when $p = 1$.

3 RELATED WORK

The formulation of NCP with least squares error and algorithms for computing it go back as far as the 1990s [50, 68], developed in part as a generalization of nonnegative matrix factorization algorithms [40] to tensors. Sidiropoulos et al. [56] provide a more detailed and complete survey that includes basic tensor factorization models with and without constraints, broad coverage of algorithms, and recent driving applications. The mathematical tensor operations discussed and the notation used in this article follow Kolda and Bader's survey [38].

Many numerical methods have been developed for the NLS subproblems that arise in NLS. Broadly these methods can be divided into projection-based and active-set-based methods. Projection-based methods are iterative methods that consist of gradient descent and Newton-type algorithms that exploit the fact that the objective function is differentiable and the nonnegative projection operator is easy to compute [15, 22, 34, 44, 47]. Active-set-like methods explicitly partition the variables into zeros and non-zeros. Once the final partition is known the NLS problem can be solved via a simpler unconstrained least-squares problem [10, 36, 39, 65]. We direct the reader to the survey by Kim et al. [35] for a more in-depth discussion on these methods.

Recently, there has been growing interest in scaling tensor operations to bigger data and more processors in both the data mining/machine learning and the high performance computing communities. For sparse tensors, there have been parallelization efforts to compute CP decompositions on shared-memory platforms [41, 60], distributed-memory platforms [32, 33, 59], and GPUs [48, 49, 62], and these approaches can be generalized to constrained problems [58].

Liavas et al. [42] extend a parallel algorithm designed for sparse tensors [59] to the 3D dense case. They use the "medium-grained" dense tensor distribution and row-wise factor matrix distribution, which is exactly the same as our distribution strategy (see Section 4.1.2), and they use a Nesterov-based algorithm to enforce the nonnegativity constraints. A similar data distribution and parallel algorithm for computing a single dense MTTKRP computation is proposed by Ballard, Knight, and Rouse [4]. Another approach to parallelizing NCP decomposition of dense tensors is presented by Phan and Cichocki [52], but they use a dynamic tensor factorization, which performs different, more independent computations across processors. Moon et al. [48] address the data locality optimizations needed during the NLS phase of the algorithm for both shared memory and GPU systems. Ma and Solomonik [45] and Singh et al. [57] compute unconstrained CP decompositions using the Cyclops Tensor Framework [61] as a backend for parallel dense tensor contractions. The former uses a pairwise perturbation technique to approximate MTTKRP computations within ALS, and the latter applies a direct optimization technique based on Gauss-Newton.

The idea of using dimension trees (discussed in Section 4.2) to avoid recomputation within MTTKRPs across modes is introduced in Reference [53] for computing the CP decomposition of dense

tensors. General reuse patterns and mode splitting were present in earlier works on variants of the Tucker Decomposition [6, 20]. It has also been used for sparse CP [33, 41] and sparse Tucker [32].

An alternate approach to speeding up CP computations is by reducing the tensor size either via sampling or compression. A large body of work exists for randomized tensor methods [7, 16, 51, 67], which are recently being extended to the constrained problem [17, 19]. The approach to reduce the tensor size is to first compress the tensor using a different decomposition, like Tucker, and then compute CP on this reduced array. This method has been discussed in further detail by Bro and De Jong [9] and Thomasi and Bro [64], but it becomes more difficult to impose nonnegative constraints on the overall model. A separate approach is to compute the (constrained) CP decomposition of the entire approximation, rather than only the core tensor, exploiting the structure of the Tucker model to perform the optimization algorithm more efficiently [66].

4 ALGORITHMS

4.1 Parallel NCP Algorithm

ALGORITHM 2: $\llbracket \mathbf{H}^{(1)}, \dots, \mathbf{H}^{(N)} \rrbracket = \text{Par-NCP}(\mathcal{X}, R)$

Require: \mathcal{X} is an $I_1 \times \dots \times I_N$ tensor distributed across a $P_1 \times \dots \times P_N$ grid of P processors, so that \mathcal{X}_p is $(I_1/P_1) \times \dots \times (I_N/P_N)$ and is owned by processor $\mathbf{p} = (p_1, \dots, p_N)$, R is rank of approximation

```

1: for  $n = 2$  to  $N$  do
2:   Initialize  $\mathbf{H}_p^{(n)}$  of dimensions  $(I_n/P) \times R$ 
3:    $\bar{\mathbf{G}} = \text{Local-SYRK}(\mathbf{H}_p^{(n)})$ 
4:    $\mathbf{G}^{(n)} = \text{All-Reduce}(\bar{\mathbf{G}}, \text{ALL-PROCS})$ 
5:    $\mathbf{H}_{p_n}^{(n)} = \text{All-Gather}(\mathbf{H}_p^{(n)}, \text{PROC-SLICE}(n, p_n))$ 
6: end for
7: % Compute NCP approximation
8: while not converged do
9:   % Perform outer iteration
10:  for  $n = 1$  to  $N$  do
11:    % Compute new factor matrix in  $n$ th mode
12:     $\bar{\mathbf{M}} = \text{Local-MTTKRP}(\mathcal{X}_p, \{\mathbf{H}_{p_1}^{(1)}, \dots, \mathbf{H}_{p_{n-1}}^{(n-1)}, \mathbf{H}_{p_{n+1}}^{(n+1)}, \dots, \mathbf{H}_{p_N}^{(N)}\}, n)$ 
13:     $\mathbf{M}_p^{(n)} = \text{Reduce-Scatter}(\bar{\mathbf{M}}, \text{PROC-SLICE}(n, p_n))$ 
14:     $\mathbf{S}^{(n)} = \mathbf{G}^{(1)} * \dots * \mathbf{G}^{(n-1)} * \mathbf{G}^{(n+1)} * \dots * \mathbf{G}^{(N)}$ 
15:     $\mathbf{H}_p^{(n)} = \text{NLS-Update}(\mathbf{S}^{(n)}, \mathbf{M}_p^{(n)})$ 
16:    % Organize data for later modes
17:     $\bar{\mathbf{G}} = \mathbf{H}_p^{(n)\top} \mathbf{H}_p^{(n)}$ 
18:     $\mathbf{G}^{(n)} = \text{All-Reduce}(\bar{\mathbf{G}}, \text{ALL-PROCS})$ 
19:     $\mathbf{H}_{p_n}^{(n)} = \text{All-Gather}(\mathbf{H}_p^{(n)}, \text{PROC-SLICE}(n, p_n))$ 
20:  end for
21: end while
```

Ensure: $\mathcal{X} \approx \llbracket \mathbf{H}^{(1)}, \dots, \mathbf{H}^{(N)} \rrbracket$

Ensure: Local matrices: $\mathbf{H}_p^{(n)}$ is $(I_n/P) \times R$ and owned by processor $\mathbf{p} = (p_1, \dots, p_N)$, for $1 \leq n \leq N$, λ stored redundantly on every processor

4.1.1 Algorithm Overview. The basic sequential algorithm is given in Algorithm 1, and the parallel version is given in Algorithm 2. In Algorithm 2, we will refer to both the inner iteration,

in which one factor matrix is updated (Lines 10–20), and the outer iteration, in which all factor matrices are updated (Lines 8–21). In the parallel algorithm, the processors are organized into a logical multidimensional grid (tensor) with as many modes as the data tensor. The communication patterns used in the algorithm are MPI collectives: All-Reduce, Reduce-Scatter, and All-Gather. The processor communicators (across which the collectives are performed) include the set of all processors and the sets of processors within the same processor slice. Processors within a mode- n slice all have the same n th coordinate. Each processor is part of N different PROC-SLICE communicators, which we denote by $\text{PROC-SLICE}(n, p_n)$, where n refers to the mode and p_n refers to the n th processor coordinate (i.e., the p_n th processor slice in mode n).

The method of enforcing the nonnegativity constraints of the linear least-squares solve or update generally affects only local computation, because each row of a factor matrix can be updated independently. In our algorithm, each processor solves the linear problem or computes the update for its subset of rows (see Line 15). The most expensive (and most complicated) part of the parallel algorithm is the computation of the MTTKRP, which corresponds to Lines 12, 13, and 19.

The details that are omitted from this presentation of the algorithm include the normalization of each factor matrix after it is computed and the computation of the residual error at the end of an outer iteration. These two computations do involve both local computation and communication, but their costs are negligible. We discuss normalization and error computation and give more detailed pseudocode in Algorithm 4 (Appendix B).

4.1.2 Data Distribution. Given a logical processor grid of processors $P_1 \times \cdots \times P_N$, we distribute the size $I_1 \times \cdots \times I_N$ tensor \mathcal{X} in a block or Cartesian partition. Each processor owns a local tensor of dimensions $(I_1/P_1) \times \cdots \times (I_N/P_N)$, and only one copy of the tensor is stored. Locally, the tensor is stored linearly, with entries ordered in a natural mode-descending way that generalizes column-major layout of matrices. Given a processor $\mathbf{p} = (p_1, \dots, p_N)$, we denote its local tensor $\mathcal{X}_{\mathbf{p}}$.

Each factor matrix is distributed across processors in a block row partition, so that each processor owns a subset of the rows. We use the notation $\mathbf{H}_{\mathbf{p}}^{(n)}$, which has dimensions $I_n/P \times R$, to denote the local part of the n th factor matrix stored on processor \mathbf{p} . However, we also make use of a redundant distribution of the factor matrices across processors, because all processors in a mode- n processor slice need access to the same entries of $\mathbf{H}^{(n)}$ to perform their computations. The notation $\mathbf{H}_{p_n}^{(n)}$ denotes the $I_n/P_n \times R$ submatrix of $\mathbf{H}^{(n)}$ that is redundantly stored on all processors whose n th coordinate is p_n (there are P/P_n such processors).

Other matrices involved in the algorithm include $\mathbf{M}_{\mathbf{p}}^{(n)}$, which is the result of the MTTKRP computation and has the same distribution scheme as $\mathbf{H}_{\mathbf{p}}^{(n)}$, and $\mathbf{G}^{(n)}$, which is the $R \times R$ Gram matrix of the factor matrix $\mathbf{H}^{(n)}$ and is stored redundantly on all processors.

4.1.3 Inner Iteration. The inner iteration is displayed graphically in Figure 2 for a three-way example and an update of the second factor matrix. The main idea is that at the start of the n th inner iteration (Figure 2(a)), all of the data is in place for each processor to perform a local MTTKRP computation, which can be computed using a dimension tree as described in Section 4.2. This means that all processors in a slice redundantly own the same rows of the corresponding factor matrix (for all modes except n). After the local MTTKRP is computed (Figure 2(b)), each processor has computed a contribution to a subset of the rows of the global MTTKRP $\mathbf{M}^{(n)}$, but its contribution must be summed up with the contributions of all other processors in its mode- n slice (denoted $\text{PROC-SLICE}(n, p_n)$ in Line 13 of Algorithm 2). This summation is performed with a Reduce-Scatter collective across the mode- n processor slice that achieves a row-wise partition of the result (in Figure 2(c), the light gray shading corresponds to the rows of $\mathbf{M}^{(2)}$ to which processor

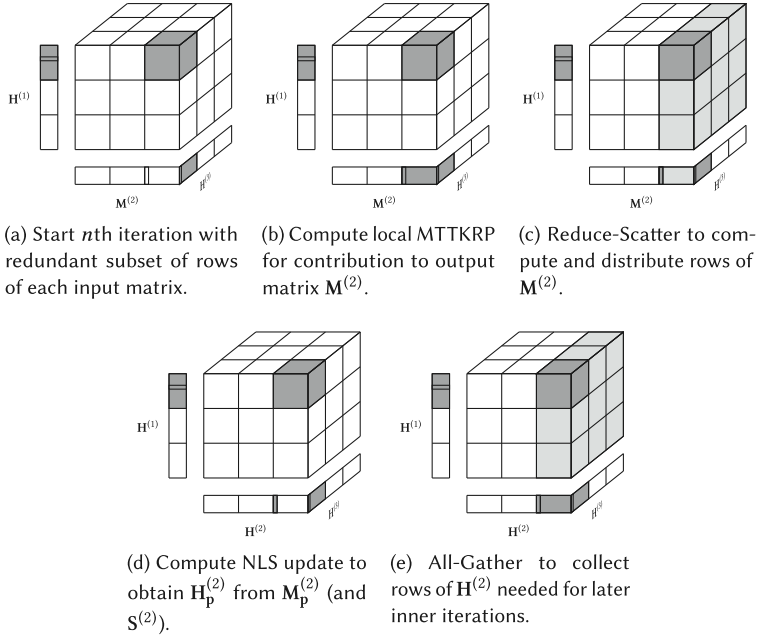


Fig. 2. Illustration of second inner iteration of Par-NCP algorithm for three-way tensor on a $3 \times 3 \times 3$ processor grid, showing data distribution, communication, and computation across steps. Highlighted areas correspond to processor (1, 3, 1) and its processor slice with which it communicates. The column normalization and computation of $G^{(2)}$, which involve communication across all processors, is not shown here.

Table 1. Costs per Outer Iteration (While Loop of Algorithm 2) and per Processor in Terms of Computation (Flops), Communication (Words Moved), and Memory (Words) Required to Compute $S^{(n)}$ and $M^{(n)}$ for Each n , Assuming the Local MTTKRP Uses a Dimension Tree [3]

Computation	Communication	Temporary Memory
$O\left(\frac{R}{P} \prod_n I_n + \frac{R^2}{P} \sum_n I_n\right)$	$O\left(R \sum_n \frac{I_n}{P_n}\right)$	$O\left(R \left(\prod_n \frac{I_n}{P_n}\right)^{1/2} + R \sum_n \frac{I_n}{P_n}\right)$

These costs do not include the computation (and possibly communication) costs of the particular NLS algorithm.

(1, 3, 1) contributes and the dark gray shading corresponds to the rows it receives as output). The output distribution of the Reduce-Scatter is designed so that afterwards, the update of the factor matrix in that mode can be performed row-wise in parallel. $S^{(n)}$ can be computed locally, since the Gram matrices, $G^{(n)}$, are stored redundantly on all processors. Along with $S^{(n)}$ each processor updates its own rows of the factor matrix given its rows of the MTTKRP result (Figure 2(d)). The remainder of the inner iteration is preparing and distributing the new factor matrix data for future inner iterations, which includes an All-Gather of the newly computed factor matrix $H^{(n)}$ across mode- n processor slices (Figure 2(e)) and recomputing $G^{(n)} = H^{(n)\top} H^{(n)}$.

4.1.4 Analysis. We will analyze the cost of a single outer iteration. While the number of outer iterations is sensitive to the NLS method used, the outer iteration time is generally the same across methods. We summarize the analysis in Table 1, which provides the dominant computation,

communication, and memory costs of a single outer-iteration: computing $\mathbf{S}^{(n)}$ and $\mathbf{M}^{(n)}$ for each n that is common to all NLS algorithms.

Computation. The local computation occurs at Lines 12, 14, 15, and 17. The cost of Line 14 is $O(NR^2)$, the cost of Line 15 is $O(R^3 I_n/P)$, which is a loose upper bound for most methods (see Section 4.3), and the cost of Line 17 is $O(R(I_n/P)^2)$. The sum of these three costs across all inner iterations is $O(R^2 N^2 + (R^3/P) \sum I_n + (R/P^2) \sum I_n^2)$, which is dominated by the cost of the MTTKRP. We compute the cost to perform the MTTKRPs using dimension trees amortized over all inner iterations. The dimension tree optimization is explained in detail in Section 4.2. The outer iteration cost is dominated by the two partial MTTKRP computations (from the root of the tree), which together are $O((R/P) \prod I_n) = O(IR/P)$ and dominate the costs of the multi-TTVs. We note that this cost involves the product of all the tensor dimensions, which is why it dominates, and we note that it scales linearly with P .

Communication. The communication within the inner iteration occurs at Lines 13, 18, and 19. Line 18 involves $O(R^2)$ data and a collective across all processors. Lines 13 and 19 involve $O(I_n R/P_n)$ data across a subset of P/P_n processors. Thus, the All-Reduce dominates the latency cost and the Reduce-Scatter/All-Gather dominate the bandwidth cost. Using efficient algorithms for the collectives (Section 2.4), the total outer iteration communication cost is $O(R \sum I_n/P_n)$ words and $O(N \log P)$ messages.

If P is large enough, then the bandwidth cost can achieve a value of $O(NRI^{1/N}/P^{1/N})$ by making the local tensors as cubical (all local tensor dimensions are roughly the same), which is communication optimal [4]. If P is not large enough (or if the tensor dimensions are too skewed) to obtain perfectly cubical tensors, then choosing the processor grid so that local tensors are as cubical as possible is also communication optimal [5] (in this case some of the processor grid dimensions will be 1). We note that the cost in the case of large P scales with $P^{1/N}$, which is far from linear scaling. However, it is proportional to the geometric mean of the tensor dimensions (on the order of one tensor dimension), which is much less than the computation cost dependence on the product of all dimensions. We report the cost for an arbitrary processor grid in Table 1, because the simplified cost expression is not always achievable. In any case of input tensor and number of processors, optimizing the processor grid within our framework is sufficient to obtain the communication lower bounds for MTTKRP [4, 5] to within a constant factor.

Memory. The algorithm requires extra local memory to run. Aside from the memory required to store the local tensor of $O(I/P)$ words and factor matrices of cumulative size $O((R/P) \sum I_n)$, each processor must be able to store a redundant subset of the rows of the factor matrices it needs to perform MTTKRP computations. This corresponds to storing P/P_n redundant copies of every factor matrix, which results in a local memory requirement of $O(R \sum I_n/P_n)$ for a general processor grid. The processor grid that minimizes communication also minimizes local memory, and the extra memory requirement can be as low as $O(NRI^{1/N}/P^{1/N})$, which is typically dominated by $O(I/P)$.

The dimension tree algorithm also requires extra temporary memory space, but the space required tends to be much smaller than what is required to store the local tensor. If the tensor dimensions can be partitioned into two parts with approximately equal geometric means, then the extra memory requirement for running a dimension tree is as small as $O(R\sqrt{I/P})$, which is also typically dominated by $O(I/P)$.

4.2 Dimension Trees

4.2.1 General Approach. An important optimization of the alternating updating algorithm for NCP (and unconstrained CP) is to re-use temporary values across inner iterations [30, 33, 41, 53]. To illustrate the idea, consider a three-way tensor \mathcal{X} approximated by $[[\mathbf{U}, \mathbf{V}, \mathbf{W}]]$ and the two

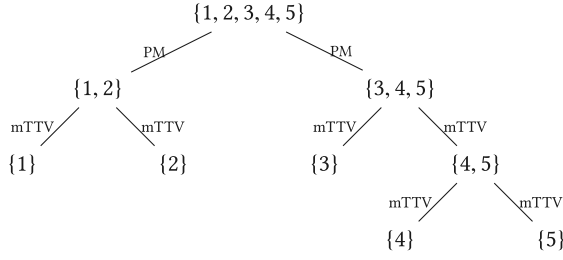


Fig. 3. Dimension tree example for $N = 5$. The data associated with the root node is the original tensor, the data associated with the leaf nodes are MTTKRP results, and the data associated with internal nodes are temporary tensors. Edges labeled with PM correspond to partial MTTKRP computations, and edges labeled with mTTV correspond to multi-TTV computations.

MTTKRP computations $\mathbf{M}^{(1)} = \mathbf{X}_{(1)}(\mathbf{W} \odot \mathbf{V})$ and $\mathbf{M}^{(2)} = \mathbf{X}_{(2)}(\mathbf{W} \odot \mathbf{U})$ used to update factor matrices \mathbf{U} and \mathbf{V} , respectively. The underlined parts of the expressions correspond to the computations that are shared between $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ and depend only on fixed quantities \mathcal{X} and the third factor matrix \mathbf{W} . Indeed, a temporary quantity, which we refer to as a *partial MTTKRP*, can be computed and re-used across the two MTTKRP expressions. We refer to the computation that combines the temporary quantity with the other factor matrix to complete the MTTKRP computation as a multi-tensor-times-vector or *multi-TTV*, as it consists of multiple operations that multiply a tensor times a set of vectors, each corresponding to a different mode.

To understand the steps of the partial MTTKRP and multi-TTV operations in more detail, we consider \mathcal{X} to be $I \times J \times K$ and \mathbf{U} , \mathbf{V} , and \mathbf{W} to have R columns. Then,

$$m_{ir}^{(1)} = \sum_{j,k} x_{ijk} v_{jr} w_{kr} = \sum_j v_{jr} \sum_k x_{ijk} w_{kr} = \sum_j v_{jr} t_{ijr},$$

where \mathcal{T} is an $I \times J \times R$ tensor that is the result of a partial MTTKRP between tensor \mathcal{X} and the single factor matrix \mathbf{W} . Likewise,

$$m_{jr}^{(2)} = \sum_{i,k} x_{ijk} u_{ir} w_{kr} = \sum_i u_{ir} \sum_k x_{ijk} w_{kr} = \sum_i u_{ir} t_{ijr},$$

and we see that the temporary tensor \mathcal{T} can be re-used and is the underlined part of $\mathbf{M}^{(1)} = \mathbf{X}_{(1)}(\mathbf{W} \odot \mathbf{V})$ and $\mathbf{M}^{(2)} = \mathbf{X}_{(2)}(\mathbf{W} \odot \mathbf{U})$. From these expressions, we can also see that computing \mathcal{T} (a partial MTTKRP) corresponds to a matrix-matrix multiplication, and computing each of $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ from \mathcal{T} (a multi-TTV) corresponds to R independent matrix-vector multiplications. In this case, we compute $\mathbf{M}^{(3)}$ using a full MTTKRP.

For a larger number of modes, a more general approach can organize the temporary quantities to be used over a maximal number of MTTKRPs. The general approach can yield significant benefit, decreasing the computation by a factor of approximately $N/2$ for dense N -way tensors. The idea is introduced in Phan et al. [53], but we adopt the terminology and notation of *dimension trees* used by Kaya et al. [30, 31, 33]. In this notation, the root node is labeled $\{1, \dots, N\}$ (we also use the notation $[N]$ for this set) and corresponds to the original tensor, a leaf is labeled $\{n\}$ and corresponds to the n th MTTKRP result $\mathbf{M}^{(n)}$, and an internal node is labeled by a set of modes $\{i, \dots, j\}$ and corresponds to a temporary tensor whose values contribute to the MTTKRP results of modes i, \dots, j .

Figure 3 illustrates a dimension tree for the case $N = 5$. Various shapes of binary trees are possible [30, 53]. For dense tensors, the computational cost is dominated by the root's branches, which correspond to partial MTTKRP computations. We perform the splitting of modes at the

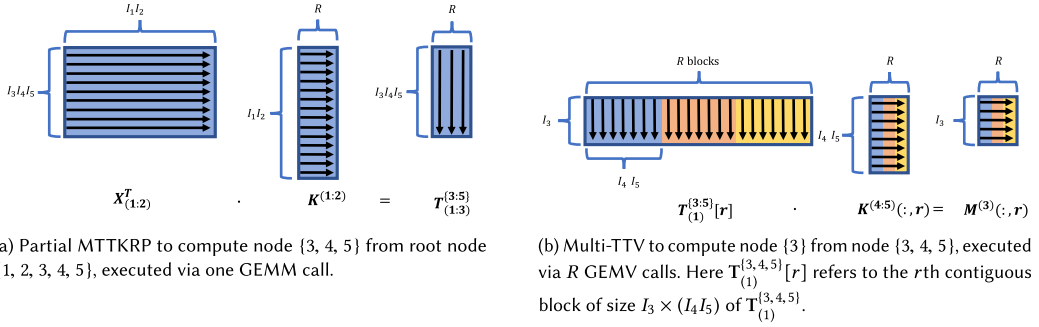


Fig. 4. Data layout and dimensions for two example computations in dimension tree shown in Figure 3. In this notation, $X_{(1:2)}$ is the matricization of input tensor \mathcal{X} with respect to modes 1 through 2, $K^{(1:2)} = H^{(2)} \odot H^{(1)}$, $\mathcal{T}^{(3,4,5)}$ is the temporary $I_3 \times I_4 \times I_5 \times R$ tensor corresponding to node $\{3, 4, 5\}$ in the dimension tree, $K^{(4:5)} = H^{(5)} \odot H^{(4)}$, and $M^{(3)}$ is the MTTKRP result for mode 3. The arrows represent row- vs. column-major ordering in memory.

root so that modes are chosen contiguously with the respect to the layout of the tensor entries in memory. In this way, each partial MTTKRP can be performed via BLAS's GEMM interface without reordering tensor entries in memory. All other edges in a tree correspond to multi-TTVs and are typically much cheaper. By organizing the memory layout of temporary quantities, the multi-TTV operations can be performed via a sequence of calls using BLAS's GEMV interface. By using the BLAS in our implementation, we are able to obtain high performance and on-node parallelism.

Figure 4 shows the data layout and dimensions of a partial MTTKRP and a multi-TTV taken from the example dimension tree in Figure 3. Figure 4(a) shows a partial MTTKRP between the input tensor \mathcal{X} and the Khatri-Rao product of the factor matrices in modes 1 and 2, which produces a temporary tensor \mathcal{T} corresponding to the $\{3, 4, 5\}$ node in the dimension tree. The key to efficiency in this computation is that the matricization of \mathcal{X} that assigns modes 1 through 2 to rows and modes 3 through 5 to columns, which we denote $X_{(1:2)}$, is already column-major in memory. Thus, we can use the GEMM interface and compute the temporary tensor \mathcal{T} without reordering any tensor entries. Note that \mathcal{T} is a four-way tensor in this case, with its last mode of dimension R , and the GEMM interface outputs the matrix $T_{(1:3)}$ (where the first three modes are assigned to rows), which is column-major in memory. Figure 4(b) depicts a multi-TTV that computes the result $M^{(3)}$ from \mathcal{T} and the factor matrices in modes 4 and 5. Here, the tensor \mathcal{T} is matricized with respect to only its first mode (of dimension I_3), but this matricization is also column-major in memory. We choose the ordering of the modes of \mathcal{T} such that each of R contiguous blocks is used to compute one column of the output matrix via a matrix-vector operation with a corresponding column of the Khatri-Rao product of the other factor matrices.

No matter how the dimension tree is designed, the computational cost of each partial MTTKRP is $2IR$, where $I = I_1 \cdots I_N$ is the number of tensor entries and R is the rank of the CP decomposition. This is the same operation count as a full MTTKRP. The computational cost of a multi-TTV is the number of entries in the temporary tensor, which is the product of a *subset* of the original tensor dimensions multiplied by R . Thus, it is computationally cheaper than the partial MTTKRPs, but it is also memory bandwidth bound. The other subroutine necessary for implementing the dimension tree approach is the Khatri-Rao product of contiguous sets of factor matrices. The computational cost of this operation is also typically lower order, but the running time in practice suffers also from being memory bandwidth bound.

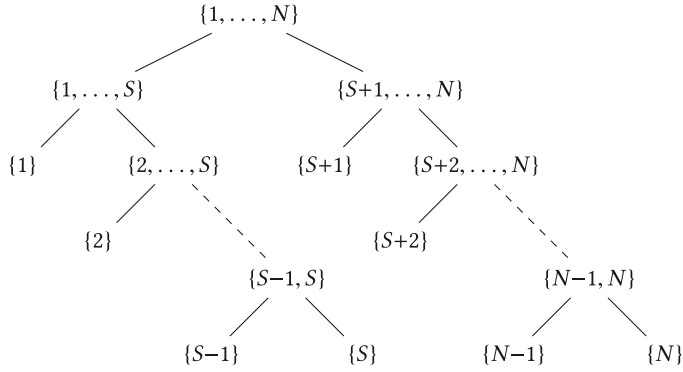


Fig. 5. Dimension tree used in PLANC for general N . Mode S is the “split” mode, chosen so that the product of dimensions in modes $\{1, \dots, S\}$ is approximately equal to that of modes $\{S+1, \dots, N\}$. The splits below the root do not necessarily minimize computation but are chosen for simplicity.

4.2.2 PLANC Implementation. For a given tensor, it is possible to compute the dimension tree that minimizes overall computation and memory. However, for most problems, the computation (and actual running time) will be dominated by the choice of split at the root node, and the other split choices will have negligible effect. The choice of split at the root node has no effect on the computational cost of the two partial MTTKRs, but it does affect the temporary memory requirement as well as the practical running time, as that split will determine the dimensions of the two GEMM calls. The three matrix dimensions in the calls are given by the products of the dimensions of the two subsets of modes and the rank of the decomposition. The amount of additional memory needed is the size of the larger partial MTTKR result and is $O(I)$ if R is less than the smallest tensor dimension.

To minimize temporary memory and optimize GEMM performance, we seek to split the modes such that the products of each subset of modes are nearly equal. To respect the memory layout of the tensor, we consider only contiguous subsets of modes, and thus the split depends on only a single parameter S , which we refer to as the “split” mode, and split the root into nodes $\{1, \dots, S\}$ and $\{S+1, \dots, N\}$. We compute S to be the smallest mode such that the product of the first S modes is greater than the product of the last $N - S$ modes.

Because the splits within the tree have much less effect on the running time and memory, we structure our tree to simplify the software implementation. That is, we compute the factor matrices in order, from 1 to N , and for every internal node of the tree, we split the smallest mode from all other modes. The structure of the tree we use in PLANC is shown in Figure 5, and the pseudocode for its implementation is given by Algorithm 3. Note that the structure of the main left subtree and the main right subtree are identical, and correspondingly the first half of the pseudocode (for modes 1 to S) is nearly identical to the second half (for modes $S+1$ to N), just with different index ranges.

To explain the pseudocode in more detail, we focus on the first half, or modes 1 through S . The first mode ($n = 1$) and the last mode ($n = S$) are special cases, because the first mode involves the partial MTTKR (Line 3) and the last mode does not compute an internal node of the tree. Internal modes ($1 < n < S$) involve computing an internal node of the tree and the MTTKR result for that mode, both of which are computed via multi-TTVs. We use the notation $\mathbf{K}^{(i:j)}$ to represent the reverse Khatri-Rao product of factor matrices $\mathbf{H}^{(i)}$ through $\mathbf{H}^{(j)}$ (i.e., $\mathbf{K}^{(i:j)} = \mathbf{H}^{(j)} \odot \mathbf{H}^{(j-1)} \odot \dots \odot \mathbf{H}^{(i)}$), which are computed in Lines 2, 4, and 8. The partial MTTKR (Line 3) is a matrix multiplication between a matricization of the tensor where the first S modes are mapped to rows

and a partial Khatri-Rao product; the output is the temporary tensor \mathcal{T} , which is computed as a matrix with R columns. Each matrix involved is either column- or row-major ordered in memory as depicted in Figure 4(a), for example, where $N = 5$. We use notation $\mathbf{T}_{(1:S)}^{\{1:S\}}$ for this output, where the subscript defines the matricization and the superscript labels the temporary tensor corresponding to its node in the dimension tree. The multi-TTV operations in Lines 5, 7, 9, and 11 are a set of R matrix-vector multiplications. We use MATLAB-style notation with parentheses to index the r th column of the Khatri-Rao product matrix and the MTTKRP result matrix. We use square-bracket notation to index contiguous column blocks of the temporary tensor. For example, in Line 9, we use $\mathbf{T}_{(1)}^{\{n:S\}}[r]$ to denote the r th column block (which comprises $I_{n+1} \cdots I_S$ columns) of the 1st-mode matricization of temporary tensor $\mathcal{T}^{\{n:S\}}$ (which has dimensions $I_n \times \cdots \times I_S \times R$). This r th column block is the same as the 1st-mode matricization of the r th slice of the tensor. The column blocks are colored distinctly in Figure 4(b), for example, where $R = 3$.

We note that for on-node parallelization, we rely on multi-threaded BLAS for the GEMM and GEMV calls, which can be offloaded to a GPU if available. For the partial Khatri-Rao products, we implement the operation as a row-wise Hadamard product of a set of factor matrix rows, and we use OpenMP parallelization to obtain on-node parallelism.

ALGORITHM 3: MTTKRP via Dimension Tree

Require: \mathcal{X} is original N -way tensor, $\mathcal{T}^{\{i:j\}}$ is temporary tensor of dimension $I_1 \times \cdots \times I_{i-1} \times I_{j+1} \times \cdots \times I_N \times R$

Require: $n \in [N]$ is inner iteration mode (evaluated in order), $S \in [N]$ is fixed split mode

```

1: if  $n = 1$  then
2:    $\mathbf{K}^{(S+1:N)} = \mathbf{H}^{(N)} \odot \cdots \odot \mathbf{H}^{(S+1)}$                                 % partial Khatri-Rao product
3:    $\mathbf{T}_{(1:S)}^{\{1:S\}} = \mathbf{X}_{(1:S)} \cdot \mathbf{K}^{(S+1:N)}$                                 % partial MTTKRP
4:    $\mathbf{K}^{(1:S-1)} = \mathbf{H}^{(S-1)} \odot \cdots \odot \mathbf{H}^{(1)}$                                 % partial Khatri-Rao product
5:    $\mathbf{M}^{(1)}(:, r) = \mathbf{T}_{(1)}^{\{1:S\}}[r] \cdot \mathbf{K}^{(1:S-1)}(:, r)$  for each  $r \in [R]$     % multi-TTV for MTTKRP result
6: else if  $n < S$  then
7:    $\mathbf{T}_{(1:S-n+1)}^{\{n:S\}}(:, r) = \mathbf{T}_{(1)}^{\{n-1:S\}}[r]^T \cdot \mathbf{H}^{(n-1)}(:, r)$  for each  $r \in [R]$  % multi-TTV for internal node tensor
8:    $\mathbf{K}^{(n+1:S)} = \mathbf{H}_{(S)} \odot \cdots \odot \mathbf{H}_{(n+1)}$                                 % partial Khatri-Rao product
9:    $\mathbf{M}^{(n)}(:, r) = \mathbf{T}_{(1)}^{\{n:S\}}[r] \cdot \mathbf{K}^{(n+1:S)}(:, r)$  for each  $r \in [R]$     % multi-TTV for MTTKRP result
10: else if  $n = S$  then
11:    $\mathbf{M}^{(S)}(:, r) = \mathbf{T}_{(1)}^{\{S-1:S\}}[r] \cdot \mathbf{H}^{(S-1)}(:, r)$  for each  $r \in [R]$     % multi-TTV for MTTKRP result
12: else if  $n = S + 1$  then
13:    $\mathbf{K}^{(1:S)} = \mathbf{H}^{(S)} \odot \cdots \odot \mathbf{H}^{(1)}$                                 % partial Khatri-Rao product
14:    $\mathbf{T}_{(1:N-S)}^{\{S+1:N\}} = \mathbf{X}_{(1:S)}^T \cdot \mathbf{K}^{(1:S)}$                                 % partial MTTKRP
15:    $\mathbf{K}^{(S+2:N)} = \mathbf{H}^{(N)} \odot \cdots \odot \mathbf{H}^{(S+2)}$                                 % partial Khatri-Rao product
16:    $\mathbf{M}^{(S+1)}(:, r) = \mathbf{T}_{(1)}^{\{S+1:N\}}[r] \cdot \mathbf{K}^{(S+2:N)}(:, r)$  for each  $r \in [R]$     % multi-TTV for MTTKRP result
17: else if  $n < N$  then
18:    $\mathbf{T}_{(1:N-n+1)}^{\{n:N\}}(:, r) = \mathbf{T}_{(1)}^{\{n-1:N\}}[r]^T \cdot \mathbf{H}^{(n-1)}(:, r)$  for each  $r \in [R]$  % multi-TTV for internal node tensor
19:    $\mathbf{K}^{(n+1:N)} = \mathbf{H}_{(N)} \odot \cdots \odot \mathbf{H}_{(n+1)}$                                 % partial Khatri-Rao product
20:    $\mathbf{M}^{(n)}(:, r) = \mathbf{T}_{(1)}^{\{n:N\}}[r] \cdot \mathbf{K}^{(n+1:N)}(:, r)$  for each  $r \in [R]$     % multi-TTV for MTTKRP result
21: else
22:    $\mathbf{M}^{(N)}(:, r) = \mathbf{T}_{(1)}^{\{N-1:N\}}[r] \cdot \mathbf{H}^{(N-1)}(:, r)$  for each  $r \in [R]$     % multi-TTV for MTTKRP result
23: end if

```

4.3 Update Algorithms

In this subsection, we consider updating algorithms for the NLS updates of the factor at each inner iteration of the algorithm (Line 13 of Algorithm 1). The general problem to be solved in each inner iteration is a constrained least-squares problem of the form

$$\mathbf{X} \leftarrow \arg \min_{\mathbf{X} \geq 0} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F^2. \quad (2)$$

All our updating methods (approximately) solve Equation (2) by first forming $\mathbf{A}^\top \mathbf{A}$ and $\mathbf{A}^\top \mathbf{B}$, matrices that appear in the gradient of the objective function. In the case of updating the factor matrix $\mathbf{H}^{(n)}$, we need to solve Equation (2) with $\mathbf{X} = \mathbf{H}^{(n)\top}$, $\mathbf{A} = \mathbf{K}^{(n)}$, where $\mathbf{K}^{(n)}$ is the KRP of factor matrices leaving out the n th factor matrix and $\mathbf{B} = \mathbf{X}^{(n)}$, where $\mathbf{X}^{(n)}$ is the n th mode matricization of \mathcal{X} . In this case, we have $\mathbf{A}^\top \mathbf{A} = \mathbf{S}^{(n)}$ and $\mathbf{A}^\top \mathbf{B} = \mathbf{M}^{(n)\top}$, which correspond to the inputs to the NLS-Update function in Line 13 of Algorithm 1.

A nice property of the Equation (2) is that it can be decoupled along the columns of \mathbf{X} and thus parallelized as in Algorithm 2. We use the notation \mathbf{X}_p to refer to a subset of the columns of \mathbf{X} owned by processor p , or in the case of Line 15 of Algorithm 2, we use $\mathbf{H}_p^{(n)}$ to refer to a subset of the rows of $\mathbf{H}^{(n)} = \mathbf{X}^\top$. The gradient for this subset of columns depends on the corresponding columns of $\mathbf{A}^\top \mathbf{B} = \mathbf{M}^{(n)\top}$, denoted by $\mathbf{M}_p^{(n)}$, and all of $\mathbf{A}^\top \mathbf{A} = \mathbf{S}^{(n)}$.

Our framework is capable of supporting any alternating-updating NCP algorithm [28]. The updating algorithms that fit this framework and are implemented in PLANC are Multiplicative Update [40], Hierarchical Alternating Least Squares [14, 24], Block Principal Pivoting [36], Alternating Direction Method of Multipliers [26], and Nesterov-type algorithm [43]. We briefly describe the different solvers below. Note that the descriptions are for the general form of the NLS problem as shown in Equation (2).

4.3.1 Multiplicative Update (MU). The MU solve is an elementwise operation [40]. The update rule for element (i, j) of \mathbf{X} is

$$\mathbf{X}(i, j) \leftarrow \mathbf{X}(i, j) \frac{\mathbf{A}^\top \mathbf{B}(i, j)}{(\mathbf{A}^\top \mathbf{A}\mathbf{X})(i, j)}. \quad (3)$$

While this rule does not solve Equation (2) to optimality it ensures a reduction in the objective value from the initial value of \mathbf{X} . Note that Equation (3) breaks down if the denominator becomes zero. In practice a small value is added to the denominator to prevent this situation.

4.3.2 Hierarchical Alternating Least Squares. HALS updates are performed on individual rows of \mathbf{X} [14, 24]. The update rule for row i can be written in closed form as

$$\mathbf{X}(i, :) \leftarrow \left[\mathbf{X}(i, :) + \frac{(\mathbf{A}^\top \mathbf{B})(i, :) - (\mathbf{A}^\top \mathbf{A}(i, :) \mathbf{X})}{(\mathbf{A}^\top \mathbf{A})(i, i)} \right]_+, \quad (4)$$

where $[\cdot]_+$ is the projection operator onto \mathbb{R}_+ . The rows of \mathbf{X} are updated in order so that the latest values are used in every update step. HALS has been observed to produce unbalanced results with either very large or very small values appearing in the factor matrices [24, 35]. Normalizing the rows of \mathbf{X} after every update via Equation (4) has been proposed to alleviate this problem [24, 35]. Within PLANC's parallelization, this step requires explicit communication among processors, because the rows of \mathbf{X} (the columns of $\mathbf{H}^{(n)}$) are distributed across processors.

4.3.3 Block Principal Pivoting. BPP is an active-set like method for solving NLS problems. The main subroutine of BPP is the single right hand side version of Equation (2),

$$\mathbf{x} \leftarrow \arg \min_{\mathbf{x} \geq 0} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2. \quad (5)$$

The **Karush-Kuhn-Tucker (KKT)** optimality conditions for Equation (5) are specified by \mathbf{x} and $\mathbf{y} = \mathbf{A}^\top \mathbf{A}\mathbf{x} - \mathbf{A}^\top \mathbf{b}$: $\mathbf{x} \geq 0$, $\mathbf{y} \geq 0$, and $\mathbf{x} * \mathbf{y} = 0$, where $*$ is the Hadamard product. The complementary slackness criteria from the KKT conditions forces the support sets, i.e., the non-zero elements, of \mathbf{x} and \mathbf{y} to be disjoint. In the optimal solution, the active-set is the set of indices where $x_i = 0$ and the remaining indices are referred to as the passive set. Once the active-set is found, we can find the optimal solution to Equation (5) by solving an unconstrained least-squares problem on the passive set of indices. The BPP algorithm attempts to find the active set by greedily swapping indices between the intermediate active and passive sets until it finds a solution that satisfies the KKT conditions. The unconstrained least squares is solved using the normal equations. Kim and Park [36] discuss the method in greater detail.

4.3.4 Alternating Direction Method of Multipliers. In the ADMM solver [26] the optimization problem Equation (2) is reformulated by introducing an auxiliary variable $\hat{\mathbf{X}}$:

$$\begin{aligned} \min_{\mathbf{X}, \hat{\mathbf{X}}} \quad & \frac{1}{2} \|\mathbf{A}\hat{\mathbf{X}} - \mathbf{B}\|_F^2 + r(\mathbf{X}), \\ \text{subject to } & \mathbf{X} = \hat{\mathbf{X}}, \end{aligned} \quad (6)$$

where $r(\cdot)$ is the penalty function for nonnegativity. It is 0 if $\mathbf{X} \geq 0$ and ∞ otherwise. The updates for the ADMM algorithm are given by

$$\begin{aligned} \hat{\mathbf{X}} &\leftarrow (\mathbf{A}^\top \mathbf{A} + \rho \mathbf{I})^{-1} (\mathbf{A}^\top \mathbf{B} + \rho (\mathbf{X} + \mathbf{U})^\top), \\ \mathbf{X} &\leftarrow \arg \min_{\mathbf{X}} r(\mathbf{X}) + \frac{\rho}{2} \|\mathbf{X} - \hat{\mathbf{X}} + \mathbf{U}\|_F^2, \\ \mathbf{U} &\leftarrow \mathbf{U} + \mathbf{X} - \hat{\mathbf{X}}, \end{aligned} \quad (7)$$

where \mathbf{U} is the scaled version of the dual variables corresponding to the equality constraints $\mathbf{X} = \hat{\mathbf{X}}$ and ρ is a step size specified by the user. \mathbf{U} is initialized as a matrix of all zeros. The advantage of using ADMM is the clever splitting of the nonnegativity constraints into updates of two blocks of variables \mathbf{X} and $\hat{\mathbf{X}}$. This allows for an unconstrained least-squares solve for $\hat{\mathbf{X}}$ and element-wise projections onto \mathbb{R}_+ for \mathbf{X} .

We can accelerate this solve by repeating the updates given by Equation (7) more than once. One important fact to notice is that the same matrix $\mathbf{A}^\top \mathbf{B}$ and matrix inverse $(\mathbf{A}^\top \mathbf{A} + \rho \mathbf{I})^{-1}$ are used for all the updates. We can therefore cache $\mathbf{A}^\top \mathbf{B}$ and the Cholesky decomposition of $(\mathbf{A}^\top \mathbf{A} + \rho \mathbf{I})$ to save some computations during subsequent updates. We stop updating using the stopping criteria described by Huang et al. [26], which is based on $\|\mathbf{X}\|_F$, $\|\hat{\mathbf{X}}\|_F$, and $\|\mathbf{U}\|_F$. Computing these norms requires communication, because each of these matrices are distributed across processors. We also limit the maximum number of acceleration steps to 5. By default, a good choice for ρ is $\|\mathbf{A}\|_F^2/R$, where R is the number of columns of \mathbf{A} (rank of the CP decomposition) [26]. For a comprehensive guide to the ADMM method, convergence properties and selection of optimal ρ please refer to article by Boyd et al. [8].

4.3.5 Nesterov-type Algorithm. The Nesterov-type algorithm in PLANC was introduced by Liavas et al. [43]. Their method solves a modified version of NLS problem Equation (2) with the introduction of a proximal term with an auxiliary matrix \mathbf{X}_* . The proximal term is useful to

handle ill-conditioned instances and guarantee strong convexity. The objective function tackled is

$$f_p(\mathbf{X}) := \frac{1}{2} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F^2 + \frac{\lambda}{2} \|\mathbf{X} - \mathbf{X}_*\|_F^2, \quad (8)$$

where \mathbf{X} is constrained to be nonnegative. The gradient of f_p is given by the expression

$$\nabla f_p(\mathbf{X}) = -(\mathbf{A}^\top \mathbf{A} \mathbf{X} - \mathbf{A}^\top \mathbf{B}) + \lambda(\mathbf{X} - \mathbf{X}_*).$$

Updates to \mathbf{X} are performed using the gradient of f_p ,

$$\begin{aligned} \nabla f_p(\mathbf{Y}_k) &= (\mathbf{A}^\top \mathbf{B} - \lambda \mathbf{X}_*) + (\lambda \mathbf{I} - \mathbf{A}^\top \mathbf{A}) \mathbf{Y}_k, \\ \mathbf{X}_{k+1} &\leftarrow [\mathbf{Y}_k - \alpha \nabla f_p(\mathbf{Y}_k)]_+, \\ \mathbf{Y}_{k+1} &\leftarrow \mathbf{X}_{k+1} + \beta_{k+1} (\mathbf{X}_{k+1} - \mathbf{X}_k), \end{aligned} \quad (9)$$

where $[\cdot]_+$ is the projection operator onto \mathbb{R}_+ . Notice that we can update \mathbf{X} multiple times reusing $\mathbf{A}^\top \mathbf{A}$ and $\mathbf{A}^\top \mathbf{B}$. This is the acceleration performed for every inner iteration in Line 15 of Algorithm 2. They are repeated until a termination criteria is triggered; different criteria are discussed in the original paper [43]. The termination criteria are bounds checks on the minimum and absolute maximum values of \mathbf{X} and require communication, because $\mathbf{X} = \mathbf{H}^{(n)\top}$ is distributed across processors. We also limit the total number of inner iterations to 20.

The selection of hyperparameters λ , α , and β depends on the singular values of \mathbf{A} and is necessary for developing a Nesterov-like method for solving Equation (8). The matrix \mathbf{X}_* is generally \mathbf{X} from the previous outer iteration (Line 8 of Algorithm 2). Details of the selection procedure and different cases can be found in the original paper [43].

In addition to the acceleration performed during each NLS solve, Equation (9), we can also perform an acceleration step for every outer iteration in the while loop (Line 8 of Algorithm 2). In this step all factor matrices are updated using the previous outer iteration values until the objective stops decreasing. The outer acceleration step for iteration i will be

$$\begin{aligned} \mathbf{H}_{new}^{(1)} &\leftarrow \mathbf{H}_i^{(1)} + s_i (\mathbf{H}_i^{(1)} - \mathbf{H}_{i-1}^{(1)}), \\ \mathbf{H}_{new}^{(2)} &\leftarrow \mathbf{H}_i^{(2)} + s_i (\mathbf{H}_i^{(2)} - \mathbf{H}_{i-1}^{(2)}), \\ &\vdots \\ \mathbf{H}_{new}^{(N)} &\leftarrow \mathbf{H}_i^{(N)} + s_i (\mathbf{H}_i^{(N)} - \mathbf{H}_{i-1}^{(N)}). \end{aligned} \quad (10)$$

The results of Equation (10) will be accepted as the next iterate only if the overall objective error with the new factor matrices, $\llbracket \mathbf{H}_{new}^{(1)}, \dots, \mathbf{H}_{new}^{(N)} \rrbracket$, is lower than that of $\llbracket \mathbf{H}_i^{(1)}, \dots, \mathbf{H}_i^{(N)} \rrbracket$. To compute the relative error, we need an extra MTTKRP computation per outer acceleration. Typically $s_i = i^{1/N}$ but its value can change as the overall algorithm progresses [43].

5 SOFTWARE

In this section, we give a brief overview of the PLANC software package structure and ways we expect users to interact with it. PLANC consists of the following modules—shared memory NMF, shared memory NTF, distributed memory NMF, and distributed memory NTF. A detailed description of the NTF module is presented with the NMF modules following a similar hierarchy. We expect users of PLANC to add new NLS solvers and use PLANC to quickly prototype their efficacy on large problems (see References [18, 46], for example). A case study of this interaction is shown. Another possible extension would be to improve the MTTKRP and matrix multiplication operations, as done by Moon et al. [48]. These routines are implemented in the abstract class, so

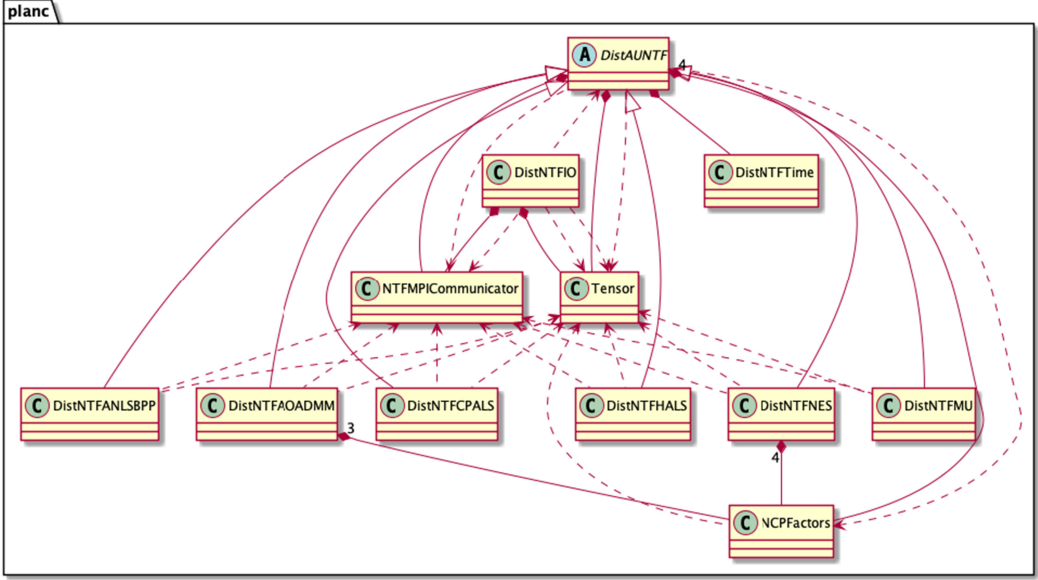


Fig. 6. PLANC UML class diagram. Solid lines with arrow heads represent is-a relationships (inheritance): the six algorithm classes near the bottom of the diagram all derive from the abstract class `DistAUNTF` at the top. Solid lines with diamond heads represent has-a relationships (composition): `NES` and `AOADM` classes have extra `NCPFactors` objects compared to the abstract `DistAUNTF` class, for example. Dotted lines represent dependencies.

any performance improvement immediately benefits all NLS solvers. Another possible interaction with the code would be to utilize the optimized MTTKRP and matrix multiplication operations. While they are not explicitly exposed by PLANC, extending these classes is relatively easy and has already been demonstrated by others [13].

5.1 Class Organization

We briefly describe the overall class hierarchy of the PLANC package as illustrated in Figure 6. PLANC offers both shared and distributed memory implementations of NTF and the classes used in each type are distinguished by the prefix `Dist` in their names (e.g., `DistAUNTF` versus `AUNTF`). We shall cover the distributed implementation of NTF in this section. Most of the descriptions can be directly applied to the shared memory case as well.

There are broadly two types of classes present. Utility classes are primarily for managing data, setting up the processor grid, and interacting with the user. Algorithm classes perform all the computations needed for NTF and implement the different NLS solvers.

5.1.1 Utility Classes.

Data. The `Tensor` and `NCPFactors` classes contain the input tensor \mathcal{X} and the factor matrices $\llbracket \mathbf{H}^{(1)}, \dots, \mathbf{H}^{(N)} \rrbracket$. The `Tensor` class stores the input tensor as a standard data array. The tensor \mathcal{X} is stored as its mode-1 unfolding $\mathbf{X}_{(1)}$ in column major order. Each processor contains its local part of the tensor (see Section 4.1.2). The `NCPFactors` class contains all the factor matrices. Each factor matrix is an Armadillo matrix [55]. The matrices are usually column normalized and the column norms are stored in the vector λ , which is present as a member of this class (see Algorithm 4). The vector λ is replicated in all processors whereas the rows of the factor matrices are distributed

across the processor grid (see Section 4.1.2). There is no global view of the entire input tensor or factor matrices and care must be taken to communicate parts of either among the processor grid.

Communication. The `NTFMPICommunicator` class creates the MPI processor grid for Algorithm 2. In addition to the communicator for the entire grid, it contains a slice communicator for each mode of the processor grid. The slice communicators are used in the Reduce-Scatter of Line 13 and All-Gather of Line 19 in Algorithm 2.

I/O. The `DistNTFIO` utility class is used to read in the input tensor from user-specified files. `DistNTFIO` also contains methods to generate random tensors and to write out the factor matrices to disk. The `ParseCommandLine` class contains all the command line options available in PLANC. As the name suggests it parses the different combinations of user inputs to instantiate the driver class and run the NTF algorithm. Some example user inputs are the target rank of the decomposition, number of outer iterations, NLS solver, and regularization parameters.

5.1.2 Algorithm Classes.

`DistAUNTF`. This is the major workhorse class of the package. It is used to implement Algorithm 2. Some of the important member functions are:

- `computeNTF`: This is the outer iteration (Line 8 in Algorithm 2).
- `distmttkrp`: Computes the distributed MTTKRP in Lines 12 and 13 in Algorithm 2.
- `gram_hadamard, update_global_gram`: These functions are used to compute the Gram matrix used in the NLS solvers.
- `computeError`: This function calculates the relative objective error of the factorization as described in Appendix B.2.

Derived classes. There exist derived classes, such as `DistNTFANLSBPP`, `DistNTFMU`, and so on, for each of the NLS solvers described in Section 4.3. There are two main functions that are present in the derived classes, which are described below. Auxiliary variables needed to implement certain NLS solvers like ADMM and Nesterov-type algorithm are also maintained in this class.

- `update`: This function is the NLS solve function. It returns the updated factor matrix using the current local MTTKRP result and global Gram matrix (see Section 4.3).
- `accelerate`: This implements the outer iteration acceleration (Line 8 in Algorithm 2). Currently only the Nesterov-type algorithm has an outer acceleration step.

5.2 Algorithm Extension

Extending PLANC to include different solvers is a simple task and we list the steps to do so below.

- (1) Create a derived class with the newly implemented update function. This is the new NLS method needed to update the factor matrices.
- (2) The constructor for the new class should contain information on whether the algorithm requires an outer acceleration step. If the method requires an outer acceleration step, then it needs to be implemented in the derived class.
- (3) Update the command line parsing class `ParseCommandLine` to include additional configuration options for the algorithm.
- (4) Include the new algorithm as an option in the utilities and the driver files.

Case Study. We describe the different steps needed to extend PLANC to include the Nesterov-type algorithm.

- (1) We first create the `DistNTFNES` class, which is derived from `DistAUNTF`.
- (2) We implement the update and accelerate functions in the derived class.

- The Nesterov NLS updates require the previous iterate values (for the auxiliary term as described in Section 4.3), which may be thought of as a persistent “state” of the algorithm. We utilize an extra `NCPFactors` object to hold these variables.
 - The Nesterov update function needs synchronization to terminate its local (iterative) NLS solve. This involves accessing the communicators found in `DistNTFMPICommunicator` class for the distributed algorithm.
 - Finally, Nesterov-type algorithms generally involve an outer acceleration step, which is also implemented in the derived class `DistNTFNES`.
- (3) We then update the `ParseCommandLine` class to include Nesterov as an algorithm.
 - (4) We update the driver file `distntf.cpp` to include the Nesterov algorithm.

6 PERFORMANCE RESULTS

6.1 Experimental Setup

The entire experimentation was performed on Titan, a supercomputer at the Oak Ridge Leadership Computing Facility. Titan is a hybrid-architecture Cray XK7 system that contains both advanced 16-core AMD Opteron **central processing units (CPUs)** and NVIDIA Kepler **graphics processing units (GPUs)**. It features 299,008 CPU cores on 18,688 compute nodes, a total system memory of 710 terabytes with 32 GB on each node, and Cray’s high-performance Gemini network.

We use Armadillo [55] for matrix representations and operations. In Armadillo, the elements of the dense matrix are stored in column major order. For dense BLAS and LAPACK operations, we linked Armadillo with the default LAPACK/BLAS wrappers from Cray. We use the GNU C++ Compiler (g++ (GCC) 6.3.0) and Cray’s MPI library. The code can also compile and run on other commodity clusters with entirely open source libraries such as OpenBLAS and OpenMPI.

6.2 Datasets

6.2.1 Mouse Data. The “Mouse” data is a 3D dataset that images a mouse’s brain over time and over a sequence of identical trials [37]. Each entry of the tensor represents a measure of calcium fluorescence in a particular pixel during a time step of a single trial. The calcium imaging is performed using an epi-fluorescence microscope viewing the brain through an artificial crystal skull. Each image has dimension $1,040 \times 1,392$, and the minimum number of time steps across 25 trials is 69. By flattening the pixel dimensions and discarding time steps after 69 for each trial, we obtain a tensor of size $1,446,680 \times 69 \times 25$. Every trial is performed with the same mouse and tracks the same task. The mouse is presented with visual simulation (starting at frame 3), and after a delay is rewarded with water (starting at frame 25). The CP decomposition can be used as an unsupervised learning technique to discover underlying patterns within the data. Because the data is nonnegative, an NCP can be more easily interpreted. We show an example interpretation of a component in Figure 15.

6.2.2 Synthetic. Our synthetic data sets are constructed from a CP model with an exact low rank with no additional noise. In this case, we can confirm that the residual error of our algorithm with a random start converges to zero. For the purposes of benchmarking, we run a fixed number of iterations of the NTF algorithms rather than using a convergence check.

6.3 Performance Breakdown Categories

The list below gives a brief description of all the categories shown in the breakdown plots and their role in the overall algorithm.

Table 2. Characteristics of the Various Update Algorithms that can Potentially Affect Performance

Alg	Communication	Extra MTTKRP	Iterative	Tuning
UCP	×	×	×	×
MU	×	×	×	×
HALS	✓ Column Norms	×	×	×
BPP	×	×	✓	×
ADMM	✓ Stopping Criteria	×	✓	✓ Step Size
NES	✓ Stopping Criteria	✓	✓	✓ See Section 4.3

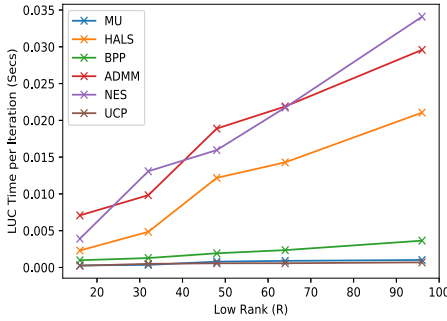
The columns are as follows: (1) if the local update requires communication, (2) if the update requires additional MTTKRP computations, (3) if the local update itself is iterative, (4) if the algorithm's performance are significantly impacted by parameter tuning. A ✓ corresponds to the algorithm having the characteristic, and a × means it does not.

- (1) Gram: the Gram matrix computation includes both the Gram computation of the local factor matrices and the Hadamard product of global Gram matrices for each factor matrix. This computation is performed on each inner iteration but is cheap under the assumption that R is small relative to the tensor dimensions.
- (2) NLS: the cost of a nonnegative least-squares update can vary drastically with the algorithm used. The various characteristics that may affect run time for each NLS algorithm are discussed in Section 6.4.
- (3) MTTKRP: the (partial) MTTKRP is a purely local computation performed on each node, and can be offloaded to the GPU. Using the dimension tree optimization (Section 4.2), we perform 2 partial MTTKRPs for each outer iteration, regardless of the number of modes N . Both operations are cast as GEMM calls, where the dimensions are given by the product of the first S mode dimensions (S is the split mode), the product of the last $N - S$ mode dimensions, and the rank R .
- (4) MultiTTV: the MultiTTVs are purely local computations performed on each node. Each MultiTTV is cast as a set of R GEMV calls, which are typically memory bandwidth bound.
- (5) ReduceScatter: the ReduceScatter collective is used to sum MTTKRP results and distribute portions of the sum appropriately across processors. It is called for each inner iteration.
- (6) AllGather: the AllGather collective is used to collect the updated factor matrices to each processor in the slice corresponding to the mode being updated. It is called after each inner iteration.
- (7) AllReduce: the AllReduce is used to compute the Gram matrices and for computing norms and other quantities required for stopping criteria of some algorithms.

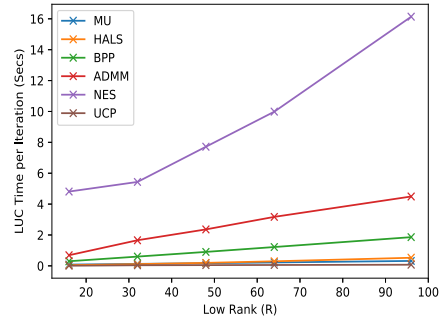
6.4 Updating Algorithm Distinctions

Table 2 highlights the distinct aspects of each updating algorithm that can affect performance. The rows of Table 2 denote the different local update algorithms implemented in PLANC. The algorithms names and acronyms in order from top to bottom in Table 2 are as follows: **Unconstrained CP (UCP)**, MU, HALS, BPP, ADMM, and **Nesterov-type algorithm (NES)**. The aspects of each algorithm that are displayed in Table 2 are as follows:

- Communication: a check mark and description in this column indicates that the local update algorithm requires some amount of communication. For example, the HALS algorithm requires the communication of the updated column norms. Additional communication requirements can affect performance by incurring additional latency and bandwidth costs. These penalties become significant when the number of processors is high.



(a) Synthetic data: $384 \times 384 \times 384 \times 384$ tensor on 81 nodes arranged in $3 \times 3 \times 3 \times 3$ grid



(b) Mouse data: $1447680 \times 69 \times 25$ tensor on 64 nodes arranged in $64 \times 1 \times 1$ grid

Fig. 7. Per Iteration LUC comparison of NLS algorithms on 4D synthetic and 3D Mouse tensors.

- **Extra MTTKRPs:** the NES algorithm has an acceleration step that requires an additional MTTKRP to be performed. This can potentially increase the run time if the acceleration step does not decrease the objective function. Experimentally, on both real and synthetic data sets, we observe that NES run time is significantly increased by the additional MTTKRP computations.
- **Iteration:** this column indicates the iterative nature of the local update algorithm. Note that all of the algorithms we present here are iterative in terms of the outer iteration. UCP, MU, and HALS all have closed-form formulas for the inner iteration, meaning the number of flops can be explicitly computed as a function of the problem size. The rest of the algorithms have flop and communication requirements dependent on the number of iterations it takes the algorithm to converge for a particular local update.
- **Tuning:** many optimization algorithms require some tunable input parameters, which can impact performance. For example, setting a step size is a frequent requirement for gradient-based optimization algorithms when an exact line search is too computationally expensive.

6.5 Microbenchmarks

6.5.1 Per-iteration Timing Comparison Across Algorithms. Figure 7 shows the “local update computation” time taken by the different updating algorithms for various low-rank values on a synthetic data set and the Mouse dataset. The synthetic tensor (Figure 7(a)) involves about 20 **Local Update Computations (LUCs)** per processor per iteration whereas the Mouse data (Figure 7(b)) has about 20,000 updates per processor per iteration, which accounts for the difference in the scales of the time seen in the figures. MU and CP are the cheapest algorithms with NES being the most expensive. HALS, ADMM, and NES algorithms all communicate in their update steps and this significantly affects their runtimes, see Figure 13(b). NES has the most expensive inner iteration involving an Eigen decomposition of the Gram matrix and up to 20 iterations of the NLS updater. ADMM has the second most expensive inner iteration with up to 5 iterations of the acceleration step. HALS, however, does not have a very expensive inner iteration but needs a synchronization to normalize every updated column of the factor matrix before proceeding to the next column, causing a slowdown.

6.5.2 Comparison Across Processor Grids. Figure 8 gives a processor grid comparison for a 3-D cubical tensor of size 512. The distributed MTTKRP time dominates the overall run time, and we observe that an even processor distribution results in the best achieved performance for all update

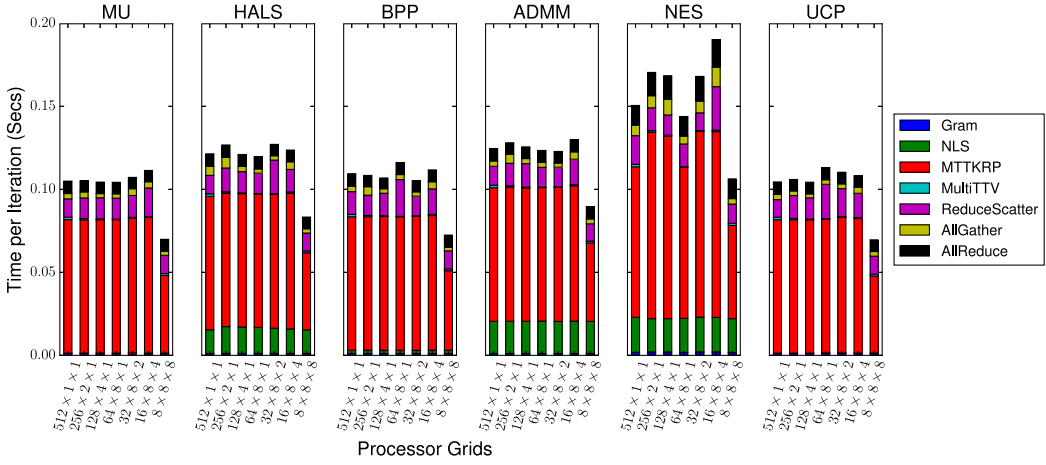


Fig. 8. Processor grid sweep of a $512 \times 512 \times 512$ synthetic 3D low-rank tensor on 512 nodes with low rank 96.

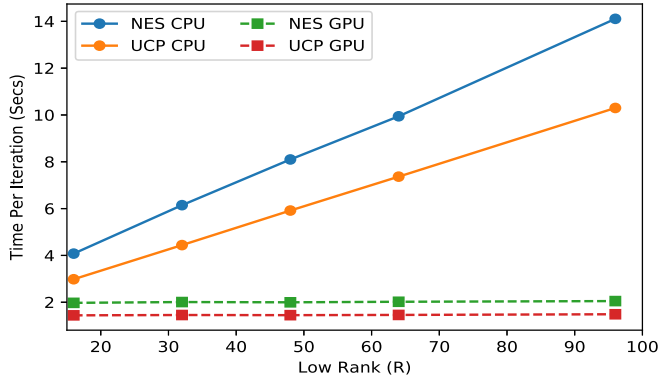


Fig. 9. Timing comparison of CPU and GPU offloading on 4D Synthetic Low-rank Tensor of size $384 \times 384 \times 384 \times 384$ on $3 \times 3 \times 3 \times 3$ processor grid with varying ranks. Only 2 updating algorithms are shown, because all other algorithms had results similar to UCP.

algorithms. This difference in run times is partially accounted for by GEMM performance due to the different shapes of the matrices involved. Besides choosing an even processor grid, we see that configurations 1 through 6 have quite stable run times with the exception of the NES algorithm. The variation in the NES run times can be attributed to the variable number of MTTKRPs needed for the acceleration steps.

6.5.3 Comparison between CPU and GPU Matrix Multiplication Offloading. Figure 9 shows comparison in run times between performing partial MTTKRPs on the CPU versus offloading to the GPU as rank increases. We only plot NES and UCP algorithms as they capture all the behaviours exhibited in offloading the matrix multiplication. All the other NLS algorithms perform similarly to UCP. As expected the CPU run times increase linearly with R as the operation count for UCP is dominated by the MTTKRP, which is linear in R . In the case of the GPU execution, the tested sizes of R are never large enough to saturate the GPU, yielding flat run times even as the rank increases. The NES algorithm takes additional time for both the CPU and GPU executions due

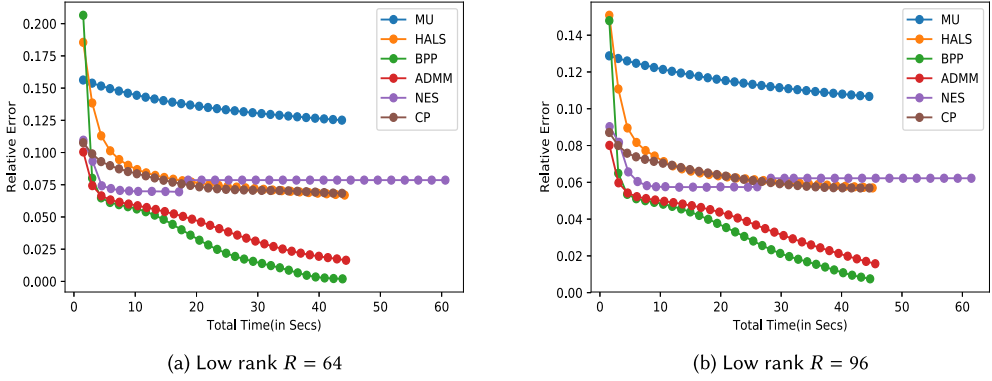


Fig. 10. Relative error over time comparison of updating algorithms on 4D Synthetic Low-rank Tensor of size $384 \times 384 \times 384 \times 384$ on $3 \times 3 \times 3$ processor grid. Each data point corresponds to an outer iteration.

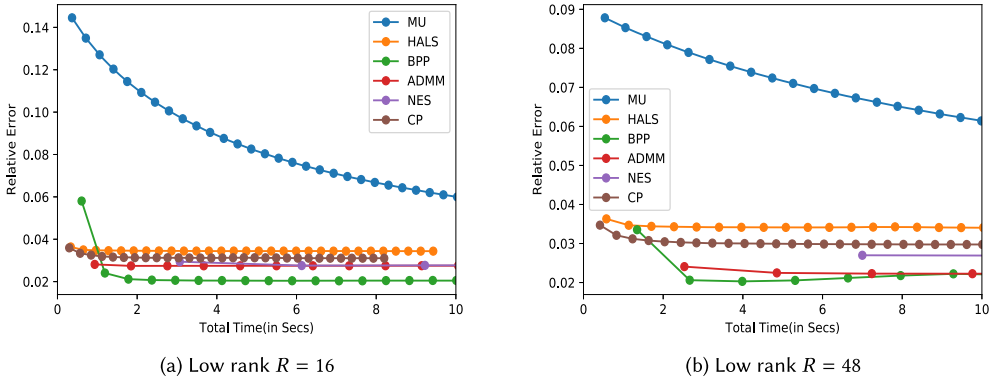
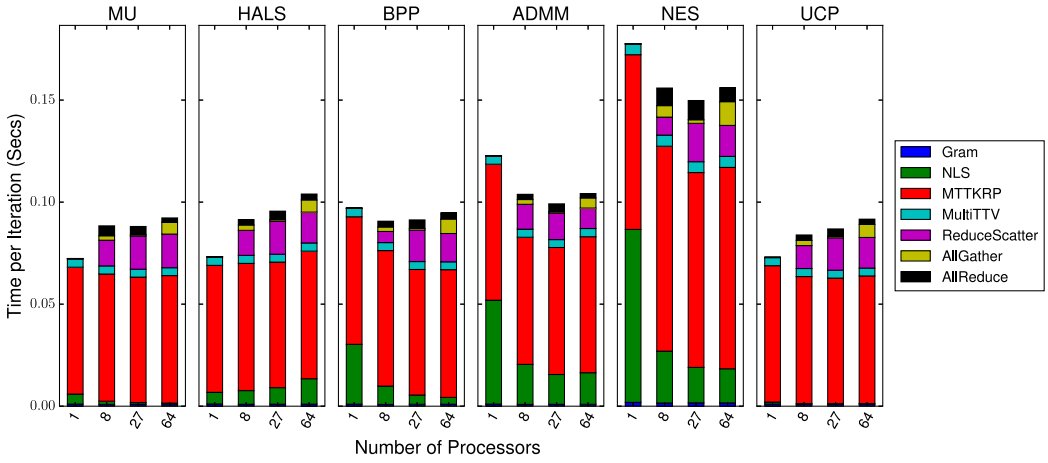


Fig. 11. Relative error comparison of updating algorithms on 3D Real-world Low-rank Tensor of size $1,447,680 \times 69 \times 25$ on a 64 Titan Nodes as $64 \times 1 \times 1$ Processor Grid for 10 s. Each data point corresponds to an outer iteration.

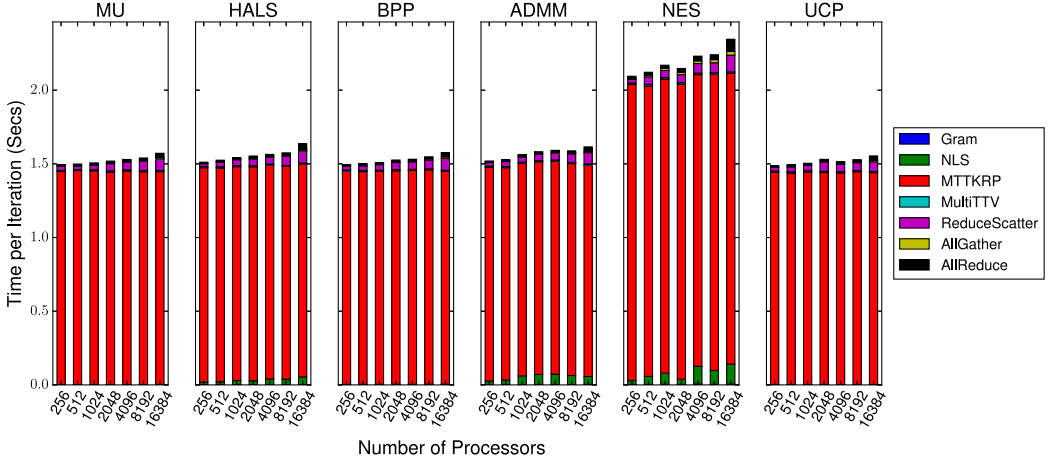
to the additional MTTKRP. In this case, for the chosen tensor size and rank, it is always beneficial to offload the GEMM calls to the GPU, and the maximum achieved speedup with GPU offloading is about 7 \times . However, we have observed in other experiments that NVBLAS can make the incorrect decision to offload the computation to the GPU when it is faster to perform it on the CPU.

6.6 Convergence Comparison Across Algorithms

Figures 10 and 11 show convergence comparisons (error vs. time) for each of the updating algorithms on synthetic low-rank and Mouse data sets, using two different target ranks each. Every algorithm is run for a fixed number (30) of outer iterations for fair comparison. For the Mouse data in Figure 11, we show only the first 10 s, because nearly all algorithms are converging within 30 iterations. The initialized random factors are the same for all algorithms in both tests, and the synthetic tensor is the same for all algorithms. In both the synthetic and real-world cases BPP achieves the lowest approximation error in the shortest amount of time. Overall results are as expected, such as MU achieving the worst error and ADMM achieving the second best in all cases. It is also note-worthy that on the real-world data set the best algorithms, ADMM and BPP, achieve relative errors of ≈ 2 –3%.



(a) Synthetic 3D Low Rank



(b) Synthetic 4D Low Rank

Fig. 12. Weak scaling on synthetic 3D and 4D low-rank tensors. For the 3D case, the input tensors are of size $128 \times 128 \times 128$, $256 \times 256 \times 256$, $378 \times 378 \times 378$, and $512 \times 512 \times 512$ on 1, 8, 27, and 64 Titan Nodes, respectively. The 4D input tensors are $128 \times 128 \times 128 \times 128$, $256 \times 256 \times 256 \times 256$, $512 \times 512 \times 512 \times 512$, $1,024 \times 512 \times 512 \times 512$, $1,024 \times 512 \times 1,024 \times 512$, $1,024 \times 1,024 \times 512 \times 512$, $1,024 \times 1,024 \times 1,024 \times 1,024$, $2,048 \times 1,024 \times 1,024 \times 1,024$, $2,048 \times 1,024 \times 2,048 \times 1,024$ on 1, 16, 256, 512, 1,024, 2,048, 4,096, 8,192, and 16,384 Titan nodes, respectively. For all experiments, the low rank is 96.

6.7 Scaling Studies

6.7.1 Weak Scaling (Synthetic Data). We performed weak scaling analysis on two different cubical tensors with three and four modes (by cubical, we mean all modes have the same dimension). Figure 12 shows the time breakdown for scaling up to 64 nodes of Titan for the 3D case and 16384 nodes for the 4D tensor. In each experiment the size of the local tensor is kept constant at dimension 128 in each mode for all the runs. As expected, the run time is dominated by the cost to compute the MTTKRP, and the domination is more extreme for higher mode tensors. Moreover, we see reasonable weak scaling as the figures remain relatively flat over all processor sizes.

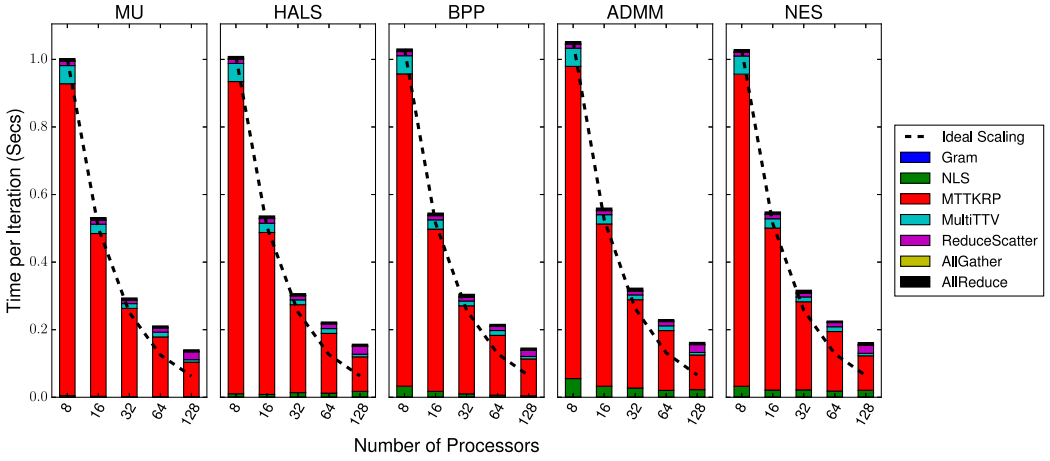
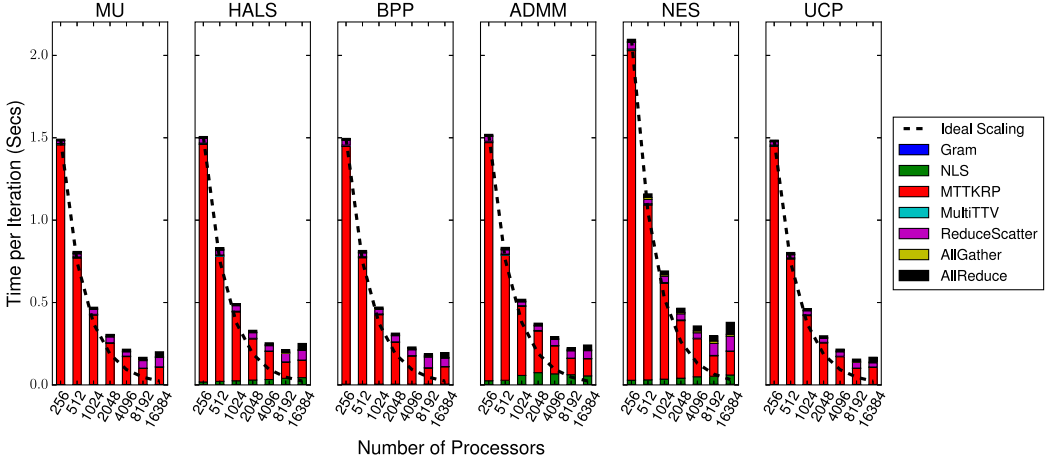
(a) Synthetic 3D Low Rank – $1,024 \times 1,024 \times 1,024$ tensor(b) Synthetic 4D Low Rank – $512 \times 512 \times 512 \times 512$ tensor

Fig. 13. Strong scaling on synthetic 3D and 4D low rank tensors with low rank 96.

The variations occur mainly due to the NLS and communication portions of the algorithm. These do matter in general and especially for the 3D case where the MTTKRP cost is often comparable to NLS times, especially for smaller number of processors. However, NLS times scale well, since they split along processor slices rather than fibers and soon become negligible for large processor grids. The amount of communication per processor remains constant but latency costs increase slowly as we scale up.

6.7.2 Strong Scaling (Synthetic Data). We run strong scaling experiments on two synthetic cubical tensors, one 3D and one 4D. Figure 13 contains these results for each of the local update algorithms ranging from 1 to 16384 processors. Since the tensors are cubical, we try to maintain the processor grids to be as close to cubical as well. For the 3D case the grids used are $2 \times 2 \times 2$, $4 \times 2 \times 2$, $4 \times 4 \times 2$, $4 \times 4 \times 4$ and $8 \times 4 \times 4$. Similarly the grids used for the 4D case are $4 \times 4 \times 4 \times 4$, $8 \times 4 \times 4 \times 4$, $8 \times 4 \times 8 \times 4$, $8 \times 8 \times 8 \times 4$, $8 \times 8 \times 8 \times 8$, $16 \times 8 \times 8 \times 8$, and $16 \times 8 \times 16 \times 8$.

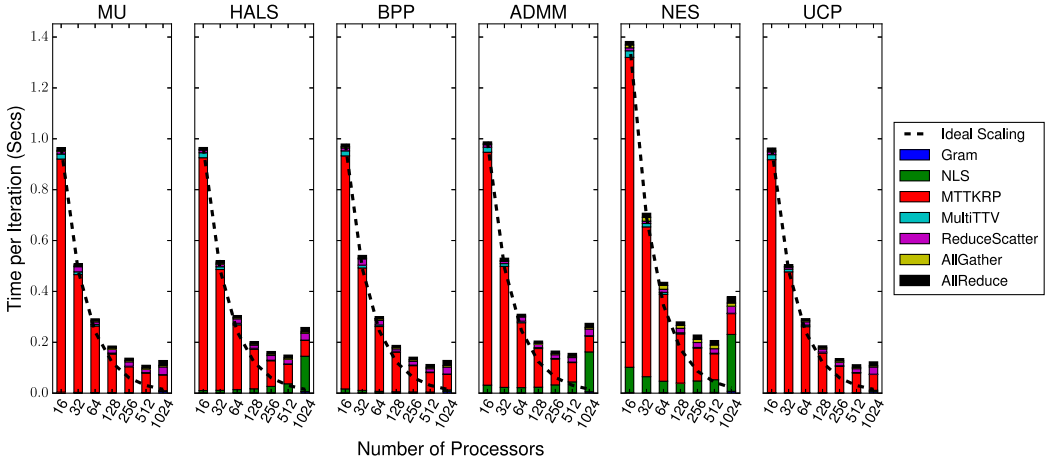


Fig. 14. Strong scaling on Mouse dataset.

We see similar behavior for the 3D and 4D case. For the 3D tensor (Figure 13(a)), we observe good strong scaling up to about 32 nodes and continue to see speed up through 128 nodes. Similarly for the 4D case (Figure 13(b)), the algorithms scale well up to about 1,024 nodes and continue to reduce time until 8,192 nodes; we observe a slowdown when scaling to 16,384 nodes.

One reason for the limit of strong scaling is the communication overheads of AllGather, AllReduce, and ReduceScatter, which become more significant for more processors. The stronger effect is the performance of the local matrix multiplications within MTTKRP's dimension tree. The smallest dimension in the matrix multiplication is typically the low-rank R , which is 96 in these experiments. For a cubical tensor of odd dimension, in our case three, the dimension tree optimization is often forced to cast partial MTTKRP into a very rectangular matrix multiplication, depending on the processor grid. Two of the local dimensions must be grouped together while the other is left alone. This means that the largest dimension would need to be close to the product of the other two in order for there to be an approximately square matrix (multiplying a tall-skinny matrix with k columns, for example). The shape and size of these local multiplications hurts the efficiency of the local computation cost and is the biggest hindrance to strong scalability for these examples.

6.7.3 Strong Scaling (Real World). Figure 14 show strong scaling results on the Mouse dataset. We use a $1D\ P \times 1 \times 1$ processor grid throughout the experiment. The results are in line with the synthetic results. We achieve near-perfect scaling up to ~ 32 nodes and still improve runtimes through 512 nodes. At 1,024 nodes the NLS algorithms, which communicate during the solve steps, perform far worse and show up to $2\times$ slowdown. The non-communicating solvers also degrade in performance but more gracefully.

6.8 Mouse Data Results

The CP decomposition of the Mouse data can be used to interpret brain patterns in response to the light stimulus and water reward given to the mouse. For example, Figure 15 shows a visualization of the factors of the 22nd component of the rank-32 CP decomposition. From the time factor, we see a marked increase in the importance of the component after the reward time frame, which suggests the activity is a response to the reward. Because the same mouse undergoes 25 identical trials, we expect to see no pattern in the time factor of each component. We note that the factors

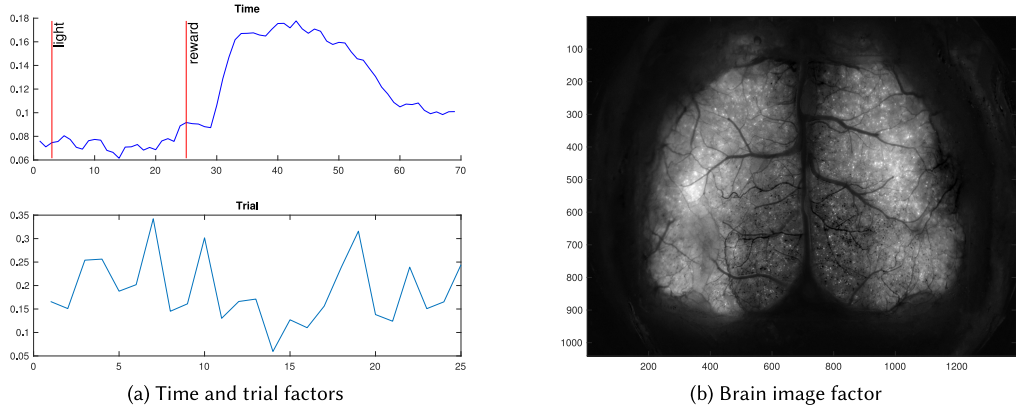


Fig. 15. Visualization of component 22 of rank-32 CP decomposition of Mouse data. The time factor visualization has been annotated with key time points marked, showing when the light stimulus was applied and when the water reward was given. The y-axes of the time and trial factors are unitless loading weights. The brain image factor has been reshaped back into original dimensions to visualize pixels with large weights in the component.

have been normalized, and the absolute magnitude of the y-axis reflects this. The pixel factor has been reshaped to an image of the same dimensions of the original data. We observe higher intensity values in the somatosensory cortex (middle, left), which is associated with bodily sensation. This component, possibly representing a sensory response to the water reward, aligns well with the findings of cell-based analysis [37, Figure 3], which also identified neurons in the somatosensory cortex with intensities that peaked quickly after the reward time frame.

The full set of components for the rank-32 CP decomposition are given in Figures 16 and 17 (Appendix A). We note that the interpretation of these computed components is useful only for exploratory analysis. Their scientific validity would need to be confirmed with tests of robustness with respect to choice of rank, random starting point, and algorithm.

7 CONCLUSION

In this work, we present PLANC, a software library for nonnegative low-rank factorizations that works for tensors of any number of modes and scales to large data sets and high processor counts. The software framework can be adapted to use any NLS algorithm within the context of alternating-updating algorithms. We use a dimension tree optimization to avoid unnecessary recomputation within the bottleneck local MTTKRP computation, and we use an efficient parallelization algorithm that minimizes communication cost. Our performance results show the ability to (weakly) scale well on synthetic data to over 16,000 nodes (35 TB of data), and we show improved performance by strong scaling on a mouse brain imaging data set of size 20 GB on up to 512 nodes.

PLANC is able to offload some of the computation to a GPU, and we show that this can significantly improve the overall runtime. This is possible because in each iteration the bottleneck computation (MTTKRP) can be cast as a pair of matrix multiplications (GEMMs), which benefit from GPU acceleration for sufficiently large dimensions. These dimensions depend on the (local) tensor size and the rank of the CP decomposition. Two of the dimensions can be tuned by the processor grid, which determines the local tensor dimensions, and the choice of dimension tree. Therefore two of the matrix multiplication dimensions are typically large. The third dimension is exactly the rank of the decomposition, so it is typically the smallest dimension and the limitation on GPU efficiency.

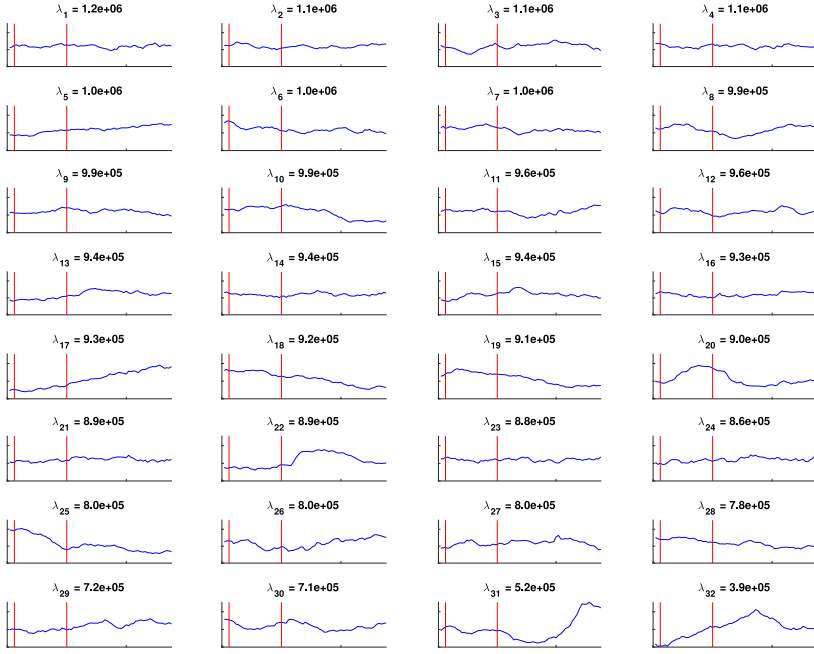
The PLANC software framework is designed to be extensible to NCP algorithms, and we demonstrate how to add an algorithm (the Nesterov-based algorithm) to the library. While previous work argued that overall performance was agnostic to NLS algorithm choice [28], these results show that for NLS algorithms that involve extra communication or significant computation, the per-iteration running time can be noticeably affected. In these cases, the time to solution depends both on the per-iteration time and the convergence rate (number of iterations).

PLANC is available at <https://github.com/ramkikannan/planc>. It provides both shared and distributed memory parallel algorithms for computing dense NMF, sparse NMF, and dense NTF. Sparse NTF is not currently supported by PLANC but there are plans to provide functionality for sparse NTF in the future.

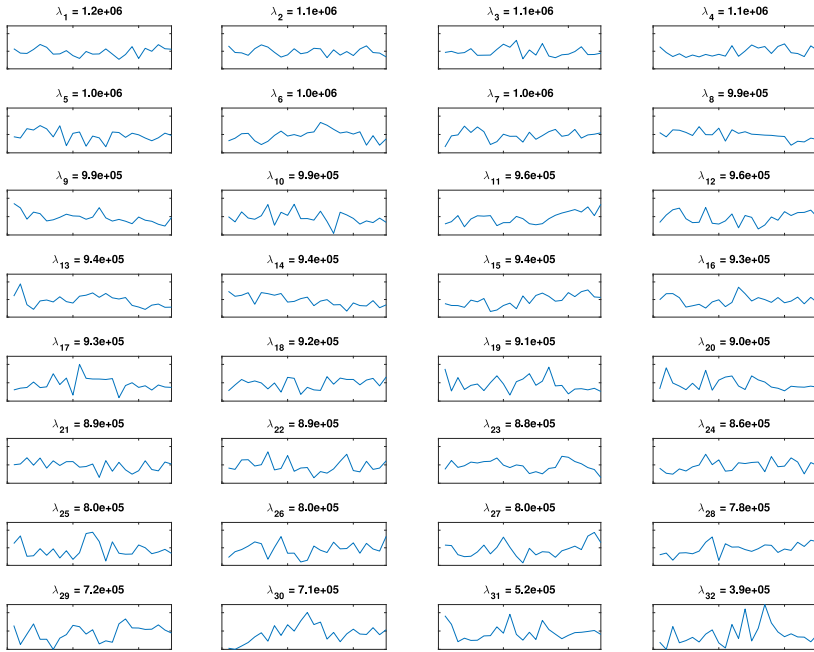
APPENDICES

A FULL RESULTS FOR MOUSE DATA

Figures 16 and 17 show all 32 components of a CP decomposition of the Mouse data. This decomposition includes the component highlighted in Figure 15. The components are ordered by their weight in the λ vector. The vertical bars in Figure 16(a) correspond to the frames of the light stimulus and water reward, respectively.



(a) Time



(b) Trial

Fig. 16. Time and trial factors of rank-32 CP decomposition of Mouse data.

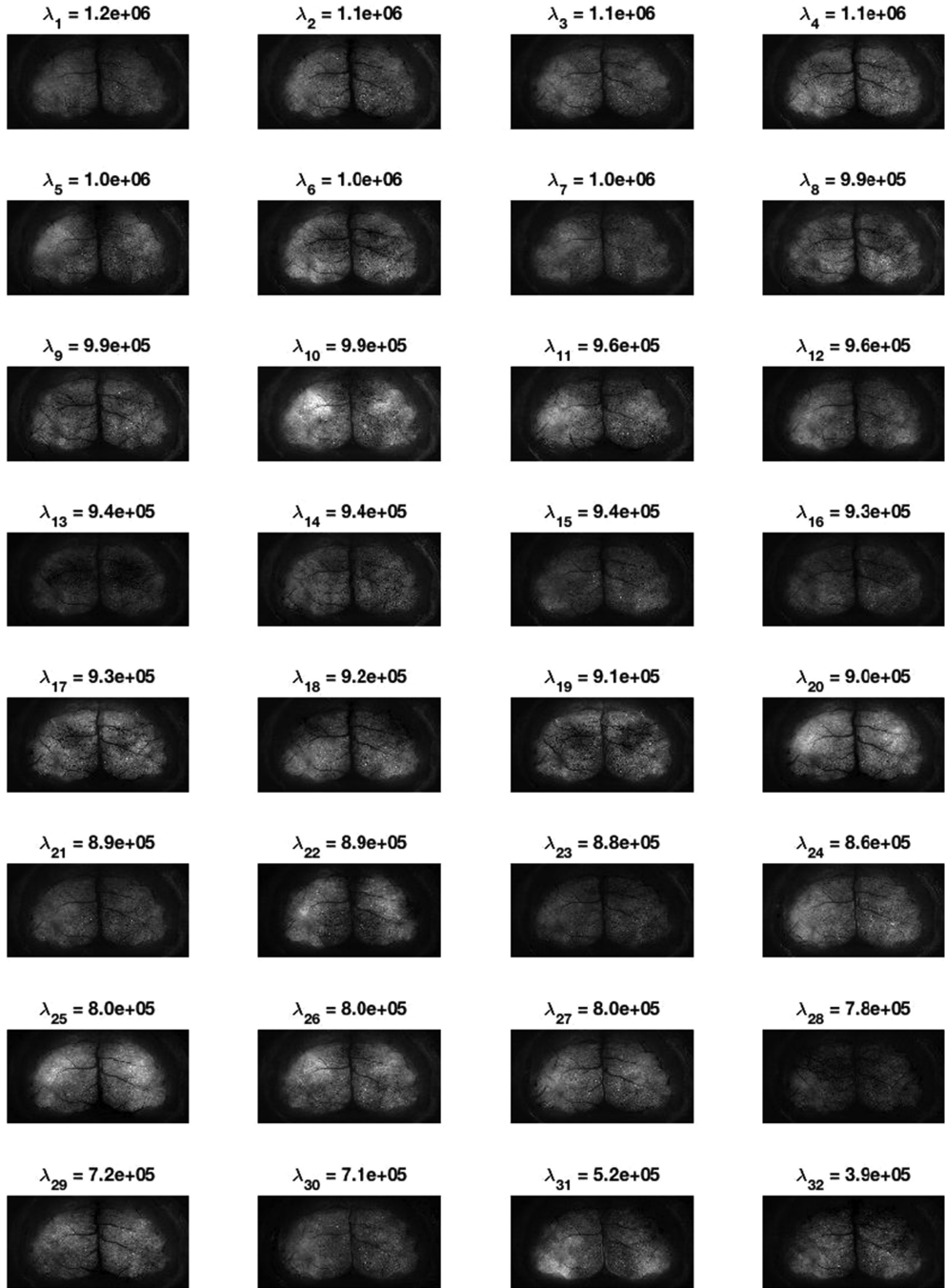


Fig. 17. Brain factors of rank-32 CP decomposition of Mouse data.

B DETAILED PARALLEL ALGORITHM

Algorithm 4 presents a more detailed version of the Parallel NCP Algorithm. The simpler version appears in Section 4 as Algorithm 2. In particular, Algorithm 4 includes the pseudocode for factor matrix normalization and relative error computation as described below.

B.1 Factor Matrix Normalization

The CP decomposition has a scale indeterminacy. To prevent possible growth in factor matrix entries, each time a factor matrix is updated, each of the R columns is normalized using the 2-norm and the weights are stored in an auxiliary vector λ . In the distributed algorithm these steps can be seen on Lines 21–23 in Algorithm 4. Note that communication is required as the global factor matrix column norms are computed.

On an algorithmic level one can observe why this step is necessary from the objective function for updating a single factor matrix in the inner iteration $\min_{\mathbf{H}^{(n)}} \|\mathbf{X}_{(n)} - \mathbf{H}^{(n)} \Lambda (\mathbf{H}^{(N)} \odot \dots \odot \mathbf{H}^{(n+1)} \odot \mathbf{H}^{(n-1)} \dots \odot \mathbf{H}^{(1)})\|_F^2$, where Λ is the diagonal matrix with the λ vector as its diagonal values. To solve, we simply collapse $\mathbf{H}^{(n)} \Lambda$ together. Thus, when the solve occurs, we are actually computing $\mathbf{H}^{(n)} \Lambda$, which is then normalized to obtain both $\mathbf{H}^{(n)}$ and the new λ .

B.2 Relative Error Computation

Given a model $\mathcal{M} = \llbracket \mathbf{H}^{(1)}, \dots, \mathbf{H}^{(N)} \rrbracket$, we compute the relative error $\|\mathcal{X} - \mathcal{M}\|/\|\mathcal{X}\|$ efficiently by using the identity $\|\mathcal{X} - \mathcal{M}\|^2 = \|\mathcal{X}\|^2 - 2\langle \mathcal{X}, \mathcal{M} \rangle + \|\mathcal{M}\|^2$. The quantity $\|\mathcal{X}\|$ is fixed, and the other two terms can be computed cheaply given the temporary matrices computed during the course of the algorithm. The second term can be computed using the identity $\langle \mathcal{X}, \mathcal{M} \rangle = \langle \mathbf{M}^{(N)}, \mathbf{H}^{(N)} \rangle$, where $\mathbf{M}^{(N)} = \mathbf{X}_{(N)} (\mathbf{H}^{(N-1)} \odot \dots \odot \mathbf{H}^{(1)})$ is the MTTKRP result in the N th mode. The third term can be computed using the identity $\|\mathcal{M}\|^2 = \mathbf{1}^T (\mathbf{S}^{(N)} * \mathbf{H}^{(N)T} \mathbf{H}^{(N)}) \mathbf{1}$ where $\mathbf{S}^{(N)} = \mathbf{H}^{(1)T} \mathbf{H}^{(1)} * \dots * \mathbf{H}^{(N-1)T} \mathbf{H}^{(N-1)}$. Both matrices $\mathbf{M}^{(N)}$ and $\mathbf{S}^{(N)}$ are computed during the course of the algorithm for updating the factor matrix $\mathbf{H}^{(N)}$, so the extra computation involved in computing the relative error is negligible. See Lines 29 to 33 in Algorithm 4. These identities have been used previously [38, 43, 54, 59].

ALGORITHM 4: $(\llbracket \lambda; \mathbf{H}^{(1)}, \dots, \mathbf{H}^{(N)} \rrbracket, \epsilon) = \text{Par-NCP}(\mathcal{X}, R)$

Require: \mathcal{X} is an $I_1 \times \dots \times I_N$ tensor distributed across a $P_1 \times \dots \times P_N$ grid of P processors, so that \mathcal{X}_p is $(I_1/P_1) \times \dots \times (I_N/P_N)$ and is owned by processor $\mathbf{p} = (p_1, \dots, p_N)$, R is rank of approximation

```

1: % Initialize data
2:  $a = \text{Norm-Squared}(\mathcal{X}_p)$ 
3:  $\alpha = \text{All-Reduce}(a, \text{ALL-PROCS})$ 
4:  $\epsilon = \text{Inf}$ 
5: for  $n = 2$  to  $N$  do
6:   Initialize  $\mathbf{H}_p^{(n)}$  of dimensions  $(I_n/P) \times R$ 
7:    $\bar{\mathbf{G}} = \text{Local-SYRK}(\mathbf{H}_p^{(n)})$ 
8:    $\mathbf{G}^{(n)} = \text{All-Reduce}(\bar{\mathbf{G}}, \text{ALL-PROCS})$ 
9:    $\mathbf{H}_{p_n}^{(n)} = \text{All-Gather}(\mathbf{H}_p^{(n)}, \text{PROC-SLICE}(n, p_n))$ 
10: end for
11: % Compute NCP approximation
12: while  $\epsilon > \text{tol}$  do
13:   % Perform outer iteration
14:   for  $n = 1$  to  $N$  do
15:     % Compute new factor matrix in  $n$ th mode
16:      $\bar{\mathbf{M}} = \text{Local-MTTKRP}(\mathcal{X}_{p_1 \dots p_N}, \{\mathbf{H}_{p_i}^{(i)}\}, n)$ 
17:      $\mathbf{M}_p^{(n)} = \text{Reduce-Scatter}(\bar{\mathbf{M}}, \text{PROC-SLICE}(n, p_n))$ 
18:      $\mathbf{S}^{(n)} = \mathbf{G}^{(1)} * \dots * \mathbf{G}^{(n-1)} * \mathbf{G}^{(n+1)} * \dots * \mathbf{G}^{(N)}$ 
19:      $\hat{\mathbf{H}}_p^{(n)} = \text{NLS-Update}(\mathbf{S}^{(n)}, \mathbf{M}_p^{(n)})$ 
20:     % Normalize columns
21:      $\bar{\lambda} = \text{Local-Col-Norms}(\hat{\mathbf{H}}_p^{(n)})$ 
22:      $\lambda = \text{All-Reduce}(\bar{\lambda}, \text{ALL-PROCS})$ 
23:      $\mathbf{H}_p^{(n)} = \text{Local-Col-Scale}(\hat{\mathbf{H}}_p^{(n)}, \lambda)$ 
24:     % Organize data for later modes
25:      $\bar{\mathbf{G}} = \mathbf{H}_p^{(n)\top} \mathbf{H}_p^{(n)}$ 
26:      $\mathbf{G}^{(n)} = \text{All-Reduce}(\bar{\mathbf{G}}, \text{ALL-PROCS})$ 
27:      $\mathbf{H}_{p_n}^{(n)} = \text{All-Gather}(\mathbf{H}_p^{(n)}, \text{PROC-SLICE}(n, p_n))$ 
28:   end for
29:   % Compute relative error  $\epsilon$  from mode- $N$  matrices
30:    $\bar{\beta} = \text{Inner-Product}(\mathbf{M}_p^{(N)}, \hat{\mathbf{H}}_p^{(N)})$ 
31:    $\beta = \text{All-Reduce}(\bar{\beta}, \text{ALL-PROCS})$ 
32:    $\gamma = \lambda^\top (\mathbf{S}^{(N)} * \mathbf{G}^{(N)}) \lambda$ 
33:    $\epsilon = \sqrt{(\alpha - 2\beta + \gamma)/\alpha}$ 
34: end while

```

Ensure: $\llbracket \mathcal{X} - \llbracket \lambda; \mathbf{H}^{(1)}, \dots, \mathbf{H}^{(N)} \rrbracket \rrbracket / \|\mathcal{X}\| = \epsilon$

Ensure: Local matrices: $\mathbf{H}_p^{(n)}$ is $(I_n/P) \times R$ and owned by processor $\mathbf{p} = (p_1, \dots, p_N)$, for $1 \leq n \leq N$, λ stored redundantly on every processor

ACKNOWLEDGMENTS

We thank Tony Hyun Kim and Mark Schnitzer for providing the Mouse dataset and help with interpretation of the CP components.

REFERENCES

- [1] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. 2014. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* 15 (2014), 2773–2832. <http://jmlr.org/papers/v15/anandkumar14b.html>
- [2] Brett W. Bader and Tamara G. Kolda. 2007. Efficient MATLAB computations with sparse and factored tensors. *SIAM J. Sci. Comput.* 30, 1 (Dec. 2007), 205–231. DOI : <https://doi.org/10.1137/060676489>
- [3] Grey Ballard, Koby Hayashi, and Ramakrishnan Kannan. 2018. Parallel nonnegative CP decomposition of dense tensors. In *Proceedings of the 25th IEEE International Conference on High Performance Computing (HiPC'18)*. 22–31. DOI : <https://doi.org/10.1109/HiPC.2018.00012>
- [4] Grey Ballard, Nicholas Knight, and Kathryn Rouse. 2018. Communication lower bounds for matricized tensor times Khatri-Rao product. In *Proceedings of the 32nd IEEE International Parallel and Distributed Processing Symposium*. 557–567. DOI : <https://doi.org/10.1109/IPDPS.2018.00065>
- [5] Grey Ballard and Kathryn Rouse. 2020. General memory-independent lower bound for MTTKRP. In *Proceedings of the SIAM Conference on Parallel Processing for Scientific Computing*. 1–11. DOI : <https://doi.org/10.1137/1.9781611976137.1>
- [6] Muthu Baskaran, Benoît Meister, Nicolas Vasilache, and Richard Lethin. 2012. Efficient and scalable computations with sparse tensors. In *Proceedings of the IEEE Conference on High Performance Extreme Computing*. IEEE, 1–6. Retrieved from <https://ieeexplore.ieee.org/abstract/document/6408676>.
- [7] Casey Battaglini, Grey Ballard, and Tamara G. Kolda. 2018. A practical randomized CP tensor decomposition. *SIAM J. Matrix Anal. Appl.* 39, 2 (2018), 876–901. <https://doi.org/10.1137/17M112303>
- [8] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein et al. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* 3, 1 (2011), 1–122. <https://doi.org/10.1561/22000000016>
- [9] Rasmus Bro and Claus A. Andersson. 1998. Improving the speed of multiway algorithms: Part II: Compression. *Chemometr. Intell. Lab. Syst.* 42, 1-2 (1998), 105–113. [https://doi.org/10.1016/S0169-7439\(98\)00011-2](https://doi.org/10.1016/S0169-7439(98)00011-2)
- [10] Rasmus Bro and Sijmen De Jong. 1997. A fast non-negativity-constrained least-squares algorithm. *J. Chemometr.: J. Chemometr. Soc.* 11, 5 (1997), 393–401. [https://doi.org/10.1002/\(SICI\)1099-128X\(199709/10\)11:5<393::AID-CEM483>3.0.CO;2-L](https://doi.org/10.1002/(SICI)1099-128X(199709/10)11:5<393::AID-CEM483>3.0.CO;2-L)
- [11] J. Douglas Carroll and Jih-Jie Chang. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of “Eckart-Young” decomposition. *Psychometrika* 35, 3 (01 Sep 1970), 283–319. DOI : <https://doi.org/10.1007/BF02310791>
- [12] E. Chan, M. Heimlich, A. Purkayastha, and R. van de Geijn. 2007. Collective communication: Theory, practice, and experience. *Concurr. Comput.: Pract. Exper.* 19, 13 (2007), 1749–1783. <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.1206>
- [13] Gopinath Chennupati, Raviteja Vangara, Erik Skau, Hristo Djidjev, and Boian Alexandrov. 2020. Distributed non-negative matrix factorization with determination of the number of latent features. *J. Supercomput.* 76, (2020), 7458–7488. <https://doi.org/10.1007/s11227-020-03181-6>
- [14] Andrzej Cichocki and Anh-Huy Phan. 2009. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE Trans. Fund. Electron. Commun. Comput. Sci.* E92-A (2009), 708–721. Issue 3. <http://dx.doi.org/10.1587/transfun.E92.A.708>
- [15] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. 2008. Nonnegative matrix and tensor factorization [lecture notes]. *IEEE Signal Process. Mag.* 25, 1 (2008), 142–145. <https://doi.org/10.1109/MSP.2008.4408452>
- [16] Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. 2011. Faster least-squares approximation. *Numerische Mathematik* 117, 2 (2011), 219–249. <https://doi.org/10.1007/s00211-010-0331-6>
- [17] N. Benjamin Erichson, Ariana Mendible, Sophie Wihlbörn, and J. Nathan Kutz. 2018. Randomized nonnegative matrix factorization. *Pattern Recogn. Lett.* 104 (2018), 1–7. <https://doi.org/10.1016/j.patrec.2018.01.007>
- [18] S. Eswar, K. Hayashi, G. Ballard, R. Kannan, R. Vuduc, and H. Park. 2020. Distributed-memory parallel symmetric nonnegative matrix factorization. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'20)*. IEEE Computer Society, 1041–1054. Retrieved from <https://dl.acm.org/doi/10.5555/3433701.3433799>.
- [19] Xiao Fu, Cheng Gao, Hoi-To Wai, and Kejun Huang. 2019. Block-randomized stochastic proximal gradient for constrained low-rank tensor factorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19)*. IEEE, 7485–7489. <http://dx.doi.org/10.1109/ICASSP.2019.8682465>

- [20] Lars Grasedyck. 2010. Hierarchical singular value decomposition of tensors. *SIAM J. Matrix Anal. Appl.* 31, 4 (2010), 2029–2054. <https://doi.org/10.1137/090764189>
- [21] Wolfgang Hackbusch. 2014. Numerical tensor calculus. *Acta Numerica* 23 (2014), 651–742. DOI: <https://doi.org/10.1017/S0962492914000087>
- [22] Lixing Han, Michael Neumann, and Upendra Prasad. 2009. Alternating projected Barzilai-Borwein methods for non-negative matrix factorization. *Electron. Trans. Numer. Anal.* 36, 6 (2009), 54–82. Retrieved from <http://etna.mcs.kent.edu/volumes/2001-2010/vol36/abstract.php?vol=36&pages=54-82>.
- [23] Richard A. Harshman. 1970. Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis. *Working Papers Phonet.* 16, 10,085 (1970), 1–84. <http://www.psychology.uwo.ca/faculty/harshman/wpppafac0.pdf>
- [24] Ngoc-Diep Ho. 2008. *Nonnegative Matrix Factorization Algorithms and Applications*. Ph.D. Dissertation. Université Catholique De Louvain. Retrieved from <https://perso.uclouvain.be/paul.vandoreen/ThesisHo.pdf>.
- [25] Kejun Huang, Nicholas D. Sidiropoulos, and Athanasios P. Liavas. 2015. Efficient algorithms for universally constrained matrix and tensor factorization. In *Proceedings of the 23rd European Signal Processing Conference (EU-SIPCO'15)*. IEEE, 2521–2525. <http://dx.doi.org/10.1109/EUSIPCO.2015.7362839>
- [26] Kejun Huang, Nicholas D. Sidiropoulos, and Athanasios P. Liavas. 2016. A flexible and efficient algorithmic framework for constrained matrix and tensor factorization. *IEEE Trans. Signal Process.* 64, 19 (2016), 5052–5065. <http://dx.doi.org/10.1109/TSP.2016.2576427>
- [27] Stephen Jesse, Miaofang Chi, Albina Borisevich, Alexei Belianinov, Sergei Kalinin, Eirik Endeve, Richard K. Archibald, Christopher T. Symons, and Andrew R. Lupini. 2016. Using multivariate analysis of scanning-Ronchigram data to reveal material functionality. *Microsc. Microanal.* 22 (July 2016), 292–293. <http://dx.doi.org/10.1017/S1431927616002312>
- [28] Ramakrishnan Kannan, Grey Ballard, and Haesun Park. 2016. A high-performance parallel algorithm for nonnegative matrix factorization. In *Proceedings of the 21st ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP'16)*. ACM, Article 9, 11 pages. DOI: <https://doi.org/10.1145/2851141.2851152>
- [29] Ramakrishnan Kannan, Grey Ballard, and Haesun Park. 2018. MPI-FAUN: An MPI-based framework for alternating-updating nonnegative matrix factorization. *IEEE Trans. Knowl. Data Eng.* 30, 3 (2018), 544–558. <https://www.computer.org/csdl/trans/tk/2018/03/08089433-abs.html>
- [30] Oguz Kaya. 2017. *High Performance Parallel Algorithms for Tensor Decompositions*. Ph.D. Dissertation. University of Lyon. Retrieved from <https://tel.archives-ouvertes.fr/tel-01623523>.
- [31] Oguz Kaya and Yves Robert. 2019. Computing dense tensor decompositions with optimal dimension trees. *Algorithmica* 81 (2019), 2092–2121. DOI: <https://doi.org/10.1007/s00453-018-0525-3>
- [32] Oguz Kaya and Bora Uçar. 2016. High performance parallel algorithms for the Tucker decomposition of sparse tensors. In *Proceedings of the 45th International Conference on Parallel Processing (ICPP'16)*. 103–112. DOI: <https://doi.org/10.1109/ICPP.2016.19>
- [33] Oguz Kaya and Bora Uçar. 2018. Parallel CANDECOMP/PARAFAC decomposition of sparse tensors using dimension trees. *SIAM J. Sci. Comput.* 40, 1 (2018). DOI: <https://doi.org/10.1137/16M1102744>
- [34] Dongmin Kim, Suvrit Sra, and Inderjit S. Dhillon. 2007. Fast Newton-type methods for the least-squares nonnegative matrix approximation problem. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 343–354. <https://doi.org/10.1137/1.9781611972771.31>
- [35] Jingu Kim, Yunlong He, and Haesun Park. 2014. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *J. Global Optimiz.* 58, 2 (Feb. 2014), 285–319. DOI: <https://doi.org/10.1007/s10898-013-0035-4>
- [36] Jingu Kim and Haesun Park. 2011. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM J. Sci. Comput.* 33, 6 (2011), 3261–3281. <https://doi.org/10.1137/110821172>
- [37] Tony Hyun Kim, Yanping Zhang, Jérôme Lecoq, Juergen C. Jung, Jane Li, Hongkui Zeng, Cristopher M. Niell, and Mark J. Schnitzer. 2016. Long-term optical access to an estimated one million neurons in the live mouse cortex. *Cell Rep.* 17, 12 (2016), 3385–3394. DOI: <https://doi.org/10.1016/j.celrep.2016.12.004>
- [38] T. G. Kolda and B. W. Bader. 2009. Tensor decompositions and applications. *SIAM Rev.* 51, 3 (Sep. 2009), 455–500. DOI: <https://doi.org/10.1137/07070111X>
- [39] Charles L. Lawson and Richard J. Hanson. 1995. *Solving Least Squares Problems*. Vol. 15. SIAM. <https://doi.org/10.1137/1.9781611971217>
- [40] Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788. <https://doi.org/10.1038/44565>
- [41] J. Li, J. Choi, I. Perros, J. Sun, and R. Vuduc. 2017. Model-driven sparse CP decomposition for higher-order tensors. In *Proceedings of the IEEE International Parallel and Distributed Processing Symposium (IPDPS'17)*. 1048–1057. DOI: <https://doi.org/10.1109/IPDPS.2017.80>

- [42] A. P. Liavas, G. Kostoulas, G. Lourakis, K. Huang, and N. D. Sidiropoulos. 2018. Nesterov-based alternating optimization for nonnegative tensor factorization: Algorithm and parallel implementation. *IEEE Trans. Signal Process.* 66, 4 (2018), 944–953.
- [43] Athanasios P. Liavas, Georgios Kostoulas, Georgios Lourakis, Kejun Huang, and Nicholas D. Sidiropoulos. 2017. Nesterov-based parallel algorithm for large-scale nonnegative tensor factorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*. IEEE, 5895–5899. <https://doi.org/10.1109/ICASSP.2017.7953287>
- [44] Chih-Jen Lin. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural Comput.* 19, 10 (2007), 2756–2779. <https://doi.org/10.1162/neco.2007.19.10.2756>
- [45] Linjian Ma and Edgar Solomonik. 2018. Accelerating alternating least squares for tensor decomposition by pairwise perturbation. Retrieved from <https://arxiv.org/abs/1811.10573>.
- [46] Lawton Manning, Grey Ballard, Ramakrishnan Kannan, and Haesun Park. 2020. Parallel hierarchical clustering using rank-two nonnegative matrix factorization. In *Proceedings of the 27th IEEE International Conference on High Performance Computing*.
- [47] Michael Merritt and Yin Zhang. 2005. Interior-point gradient method for large-scale totally nonnegative least-squares problems. *J. Optimiz. Theory Appl.* 126, 1 (2005), 191–202. <https://doi.org/10.1007/s10957-005-2668-z>
- [48] Gordon E. Moon, Aravind Sukumaran-Rajam, Srinivasan Parthasarathy, and P. Sadayappan. 2019. PL-NMF: Parallel locality-optimized non-negative matrix factorization. Retrieved from <https://arxiv.org/abs/1904.07935>.
- [49] Israt Nisa, Jiajia Li, Aravind Sukumaran-Rajam, Richard Vuduc, and P. Sadayappan. 2019. Load-balanced sparse MT-TKRP on GPUs. In *Proceedings of the IEEE International Parallel and Distributed Processing Symposium (IPDPS'19)*. IEEE, 123–133. <https://doi.org/10.1109/IPDPS.2019.00023>
- [50] Pentti Paatero. 1997. A weighted non-negative least-squares algorithm for three-way PARAFAC factor analysis. *Chemometr. Intell. Lab. Syst.* 38, 2 (1997), 223–242. DOI: [https://doi.org/10.1016/S0169-7439\(97\)00031-2](https://doi.org/10.1016/S0169-7439(97)00031-2)
- [51] Evangelos E. Papalexakis, Christos Faloutsos, and Nicholas D. Sidiropoulos. 2012. Parcube: Sparse parallelizable tensor decompositions. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 521–536. https://doi.org/10.1007/978-3-642-33460-3_39
- [52] Anh Huy Phan and Andrzej Cichocki. 2011. PARAFAC algorithms for large-scale problems. *Neurocomputing* 74, 11 (2011), 1970–1984. DOI: <https://doi.org/10.1016/j.neucom.2010.06.030>
- [53] Anh-Huy Phan, Petr Tichavsky, and Andrzej Cichocki. 2013. Fast alternating LS algorithms for high order CAN-DECOMP/PARAFAC tensor factorizations. *IEEE Trans. Signal Process.* 61, 19 (Oct. 2013), 4834–4846. DOI: <https://doi.org/10.1109/TSP.2013.2269903>
- [54] Anh-Huy Phan, Petr Tichavsky, and Andrzej Cichocki. 2013. TENSORBOX: A MATLAB package for tensor decomposition. Retrieved from <https://github.com/phananhhuu/TensorBox>.
- [55] Conrad Sanderson. 2010. *Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments*. Technical Report. NICTA. Retrieved from http://arma.sourceforge.net/armadillo_nicta_2010.pdf.
- [56] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. 2017. Tensor decomposition for signal processing and machine learning. *IEEE Trans. Signal Process.* 65, 13 (July 2017), 3551–3582. DOI: <https://doi.org/10.1109/TSP.2017.2690524>
- [57] Navjot Singh, Linjian Ma, Hongru Yang, and Edgar Solomonik. 2019. *Comparison of Accuracy and Scalability of Gauss-Newton and Alternating Least Squares for CP Decomposition*. Technical Report. Retrieved from <https://arxiv.org/abs/1910.12331>.
- [58] S. Smith, A. Beri, and G. Karypis. 2017. Constrained tensor factorization with accelerated AO-ADMM. In *Proceedings of the 46th International Conference on Parallel Processing (ICPP'17)*. 111–120. DOI: <https://doi.org/10.1109/ICPP.2017.20>
- [59] Shaden Smith and George Karypis. 2016. A medium-grained algorithm for distributed sparse tensor factorization. In *Proceedings of the IEEE 30th International Parallel and Distributed Processing Symposium*. 902–911. DOI: <https://doi.org/10.1109/IPDPS.2016.113>
- [60] S. Smith, N. Ravindran, N. D. Sidiropoulos, and G. Karypis. 2015. SPLATT: Efficient and parallel sparse tensor-matrix multiplication. In *Proceedings of the IEEE International Parallel and Distributed Processing Symposium*. 61–70. DOI: <https://doi.org/10.1109/IPDPS.2015.27>
- [61] Edgar Solomonik, Devin Matthews, Jeff R. Hammond, John F. Stanton, and James Demmel. 2014. A massively parallel tensor contraction framework for coupled-cluster computations. *J. Parallel Distrib. Comput.* 74, 12 (2014), 3176–3190. DOI: <https://doi.org/10.1016/j.jpdc.2014.06.002>
- [62] Bing Tang, Linyao Kang, Yanmin Xia, and Li Zhang. 2018. GPU-accelerated large-scale non-negative matrix factorization using spark. In *Proceedings of the International Conference on Collaborative Computing: Networking, Applications and Worksharing*. Springer, 189–201. https://doi.org/10.1007/978-3-030-12981-1_13
- [63] Rajeev Thakur, Rolf Rabenseifner, and William Gropp. 2005. Optimization of collective communication operations in MPICH. *Int. J. High Perform. Comput. Appl.* 19, 1 (2005), 49–66. DOI: <https://doi.org/10.1177/1094342005051521>

- [64] Giorgio Tomasi and Rasmus Bro. 2006. A comparison of algorithms for fitting the PARAFAC model. *Comput. Stat. Data Anal.* 50, 7 (2006), 1700–1734. <https://doi.org/10.1016/j.csda.2004.11.013>
- [65] Mark H. Van Benthem and Michael R. Keenan. 2004. Fast algorithm for the solution of large-scale non-negativity-constrained least-squares problems. *J. Chemometr.: J. Chemometr. Soc.* 18, 10 (2004), 441–450. <https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.889>
- [66] Nico Vervliet, Otto Debaes, and Lieven De Lathauwer. 2019. Exploiting efficient representations in large-scale tensor decompositions. *SIAM J. Sci. Comput.* 41, 2 (2019), A789–A815. DOI: <https://doi.org/10.1137/17M1152371>
- [67] Yining Wang, Hsiao-Yu Tung, Alexander J. Smola, and Anima Anandkumar. 2015. Fast and guaranteed tensor decomposition via sketching. In *Advances in Neural Information Processing Systems*. 991–999. Retrieved from <https://dl.acm.org/doi/10.5555/2969239.2969350>.
- [68] Max Welling and Markus Weber. 2001. Positive tensor factorization. *Pattern Recogn. Lett.* 22, 12 (2001), 1255–1261. DOI: [https://doi.org/10.1016/S0167-8655\(01\)00070-8](https://doi.org/10.1016/S0167-8655(01)00070-8)

Received August 2019; revised June 2020; accepted October 2020