Accelerating Alternating Least Squares for Tensor Decomposition by Pairwise Perturbation

Linjian Ma[1] and *Edgar Solomonik*[2]

L·P·N A @CS@Illinois

[1]EECS Department
University of California, Berkeley

[2]Department of Computer Science
University of Illinois at Urbana-Champaign

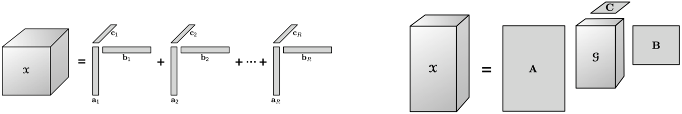
SIAM CSE
Spokane, WA

Outline

1. Overview
2. Alternating Least Squares for CP Decomposition
3. Pairwise Perturbation Algorithm
4. Error Analysis
5. Performance Results
6. Conclusion

Overview

CP and Tucker tensor decompositions[1]

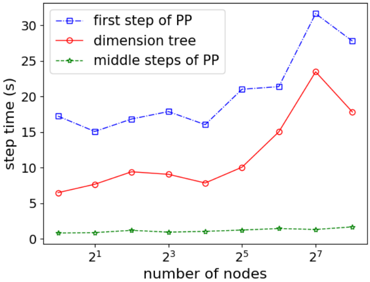


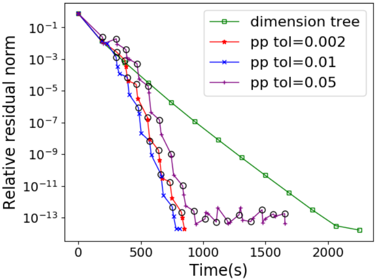
- Alternating least squares (ALS) is most widely used method
- Each ALS sweep optimizes all factor matrices in decomposition
- New algorithm: pairwise perturbation approximates ALS
  - accurate when factor tensors change little at each sweep
  - reduces cost of sweep from $O(s^N R)$ to $O(s^2 R)$ for rank R CP decomposition of order N tensor with dims $s \times \cdots \times s$

[1]Kolda and Bader, SIAM Review 2009

Performance Highlights for CP Decomposition

Pairwise perturbation (PP) outperforms optimized dimension tree ALS

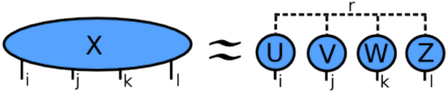




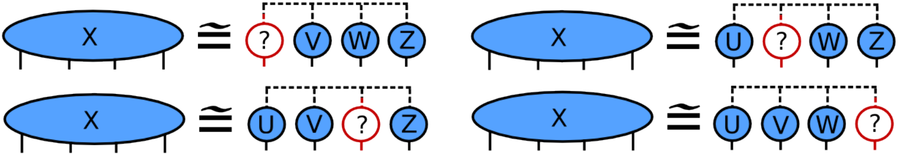
- first step of PP (setup) costs slightly more than ALS sweep
- middle steps (subsequent approximations) up to 10X faster
- overall convergence up to 3X faster for synthetic and real tensors

Consider rank R CP decomposition of an $s \times s \times s \times s$ tensor

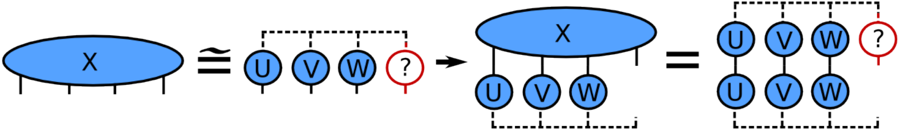
$$x_{ijkl} \approx \sum_{r=1}^{R} u_{ir} v_{jr} w_{kr} z_{lr}$$



ALS updates factor matrices in an alternating manner

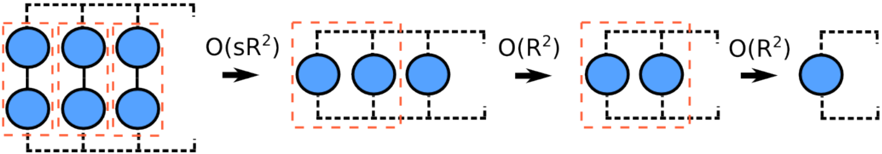


Each quadratic subproblem is typically solved via normal equations

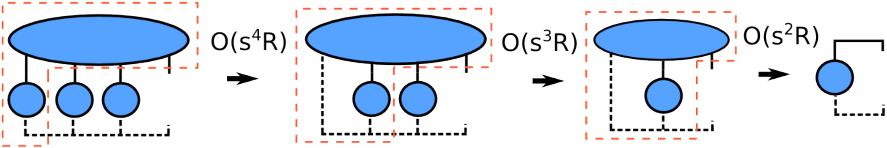


Tensor Contractions in CP ALS

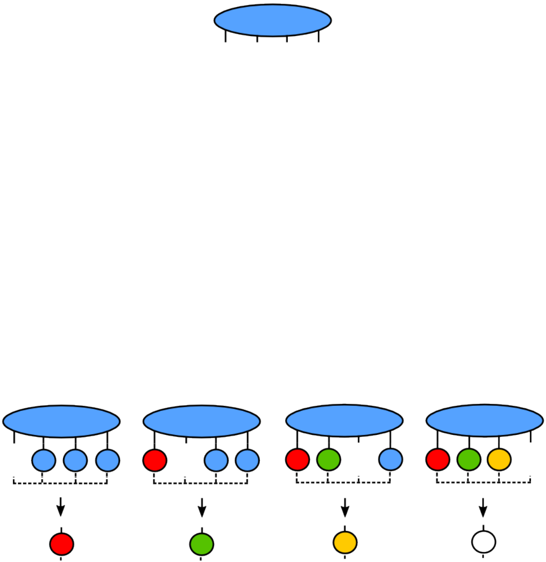
The normal equations are cheap to compute



But forming the right-hand sides requires expensive MTTKRP

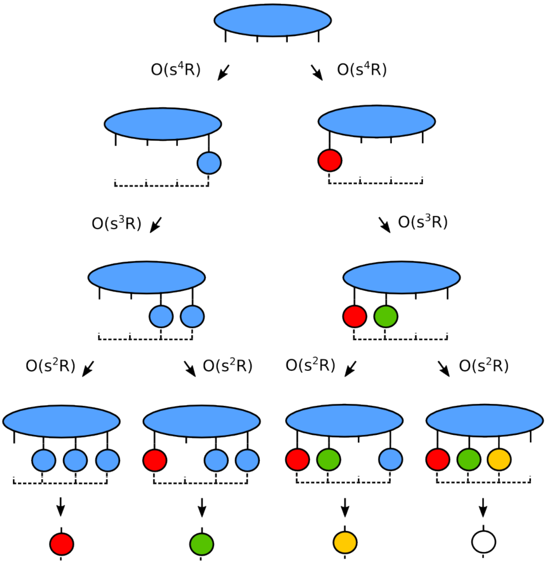


CP ALS Dimension Trees[2]



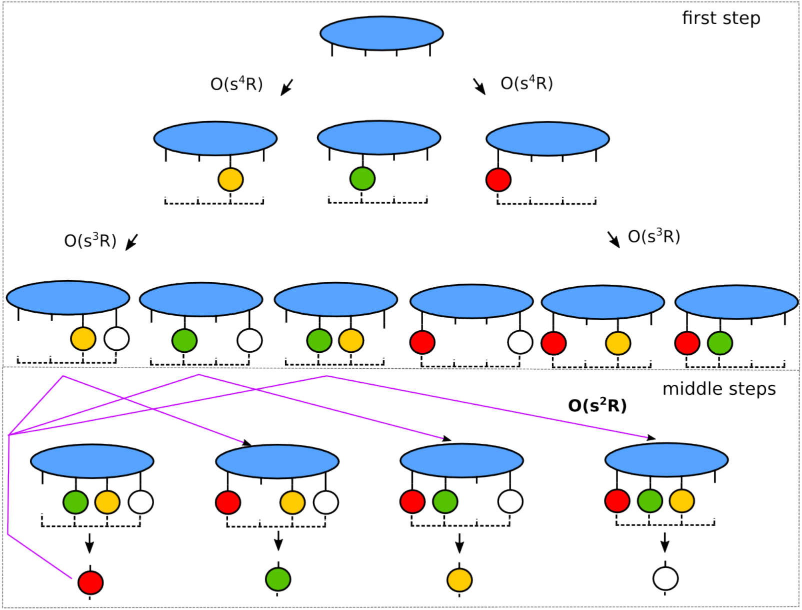
[2]Phan, Tichavskỳ, and Cichocki, IEEE Transactions on Signal Processing 2013

CP ALS Dimension Trees[3]



[3] Phan, Tichavskỳ, and Cichocki, IEEE Transactions on Signal Processing 2013

CP ALS with Pairwise Perturbation



Error Analysis: First Attempt

Consider order $N = 3$ tensor $\boldsymbol{\mathcal{T}}$, let $\boldsymbol{M}^{(3)}$ be the right-hand-sides needed to form the third factor matrix $\boldsymbol{A}^{(3)}$

- Bound columnwise error of $\tilde{\boldsymbol{M}}^{(3)}$ computed by PP middle step
- The ith factor matrix changed by $d\boldsymbol{A}^{(i)}$ since the first step of PP
- Error bound based on conditioning bound of $\boldsymbol{f}_{\boldsymbol{\mathcal{T}}} \in \mathbb{R}^s \times \mathbb{R}^s \to \mathbb{R}^s$,

$$\boldsymbol{z} = \boldsymbol{f}_{\boldsymbol{\mathcal{T}}}(\boldsymbol{x}, \boldsymbol{y}) \Rightarrow z_k = \sum_{j,k} t_{ijk} x_i y_j$$

Theorem (Columnwise Error Bound from Tensor Conditioning)

If $||d\boldsymbol{a}_k^{(l)}||_2 / ||\boldsymbol{a}_k^{(l)}||_2 \leq \epsilon$ *for* $l \in \{1, 2, 3\}$,

$$\frac{||\tilde{\boldsymbol{m}}_k^{(3)} - \boldsymbol{m}_k^{(3)}||_2}{||\boldsymbol{m}_k^{(3)}||_2} \leq \frac{\max_{\boldsymbol{u},\boldsymbol{v} \in \mathbb{S}^{s-1}} ||\boldsymbol{f}_{\boldsymbol{\mathcal{T}}}(\boldsymbol{u}, \boldsymbol{v})||_2}{\min_{\boldsymbol{x},\boldsymbol{y} \in \mathbb{S}^{s-1}} ||\boldsymbol{f}_{\boldsymbol{\mathcal{T}}}(\boldsymbol{x}, \boldsymbol{y})||_2} O(\epsilon^2).$$

MTTKRP is Ill-Posed for Most Tensors

- Error bound relies on worst-case behavior of $\boldsymbol{f_{\mathcal{T}}} \in \mathbb{R}^s \times \mathbb{R}^s \to \mathbb{R}^s$,

$$\boldsymbol{z} = \boldsymbol{f_{\mathcal{T}}}(\boldsymbol{x}, \boldsymbol{y}) \Rightarrow z_k = \sum_{j,k} t_{ijk} x_i y_j$$

- If $\min_{\boldsymbol{x},\boldsymbol{y} \in \mathbb{S}^{s-1}} ||\boldsymbol{f_{\mathcal{T}}}(\boldsymbol{x}, \boldsymbol{y})||_2 = 0$, bound is trivial
- There exist $2 \times 2 \times 2$, $4 \times 4 \times 4$, and $8 \times 8 \times 8$ tensors for which $||\boldsymbol{f_{\mathcal{T}}}(\boldsymbol{x}, \boldsymbol{y})||_2 = 1$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{s-1}$
  - thanks to Fan Huang for finding the $s = 8$ tensor
- However, for any $s \notin \{1, 2, 4, 8\}$, any $s \times s \times s$ tensor $\boldsymbol{\mathcal{T}}$ has $\min_{\boldsymbol{x},\boldsymbol{y} \in \mathbb{S}^{s-1}} ||\boldsymbol{f_{\mathcal{T}}}(\boldsymbol{x}, \boldsymbol{y})||_2 = 0$
- Tensors that are well-conditioned in this sense correspond to solutions to the Hurwitz problem (1898), which exist only for $s \in \{2, 4, 8\}$
  - thanks to Daniel Kressner for pointing out this connection

Error Analysis: Second Attempt

Again, consider order $N = 3$ tensor $\boldsymbol{\mathcal{T}}$, let $\boldsymbol{M}^{(3)}$ be the right-hand-sides needed to form the third factor matrix $\boldsymbol{A}^{(3)}$

- Bound columnwise error of *approximate update* $\tilde{\boldsymbol{H}}^{(1,3)}$ to $\tilde{\boldsymbol{M}}^{(3)}$ computed by PP middle step due to change in $\boldsymbol{A}^{(1)}$

Theorem (Columnwise Error Bound from Matricization Conditioning)

For $\epsilon_k = ||d\boldsymbol{a}_k^{(2)}||_2/||\boldsymbol{a}_k^{(2)}||_2 < 1$ *and matricization* $\hat{\boldsymbol{T}} = {\boldsymbol{T}_{(2)}}^T$*,*

$$\frac{||\tilde{\boldsymbol{h}}_k^{(1,3)} - \boldsymbol{h}_k^{(1,3)}||_2}{||\boldsymbol{h}_k^{(1,3)}||_2} \leq \kappa(\hat{\boldsymbol{T}})\epsilon_k, \ \textit{where } \kappa(\hat{\boldsymbol{T}}) = \frac{\sigma_{\max}(\hat{\boldsymbol{T}})}{\sigma_{\min}(\hat{\boldsymbol{T}})}$$

- For higher-order tensors, can look at conditioning of matricizations of order-3 slices of MTTKRP intermediates
- Higher-order absolute error terms scale as $O(\epsilon_k\epsilon_l)$, but can dominate, so have no relative error bound

Implementation

We used Cyclops Tensor Framework[4] to implement standard dimension tree ALS and pairwise perturbation

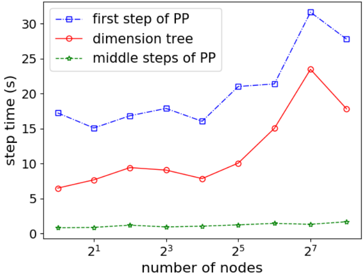
- Cyclops is a C++ library that distributes each tensor over MPI
- Used in chemistry (PySCF, QChem), quantum circuit simulation (IBM/LLNL), and graph analysis (betweenness centrality)
- Summations and contractions specified via Einstein notation

```
E["aixbjy"] += X["aixbjy"] - U["abu"]*V["iju"]*W["xyu"]
```

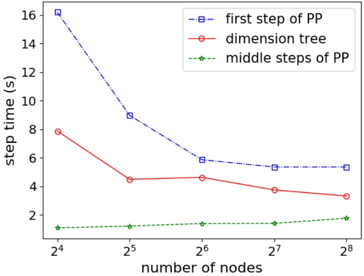
- Best distributed contraction algorithm auto-selected at runtime
- Sparse tensors supported but unused here
- Python interface, OpenMP, and GPU support present but unused
- Used interface to ScaLAPACK SVD to solve linear systems

[4] https://github.com/cyclops-community/ctf

Strong and Weak Scaling Microbenchmarks



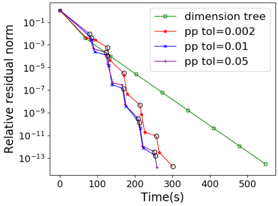
(a) Weak scaling



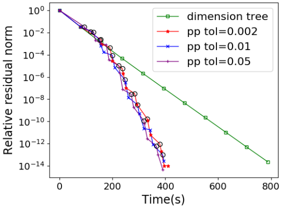
(b) Strong scaling

- Experiments performed on Stamepde2 TACC supercomputer
- Weak scaling: dimension $s = \lfloor 32p^{1/6} \rfloor$ and rank $R = \lfloor 4p^{1/6} \rfloor$
- Strong scaling: dimension $s = 50$ and rank $R = 6$
- First step of PP (setup) costs slightly more than ALS sweep
- Middle steps (subsequent approximations) up to 10X faster

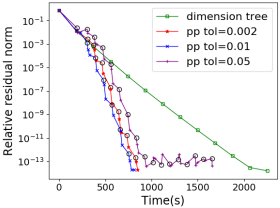
Results for Synthetic Tensors



(a) Random tensor on 1 node



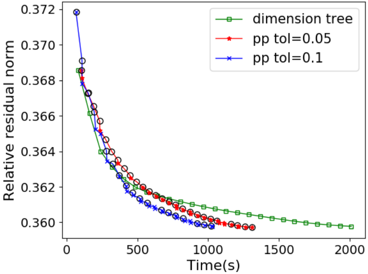
(b) Random tensor on 16 nodes



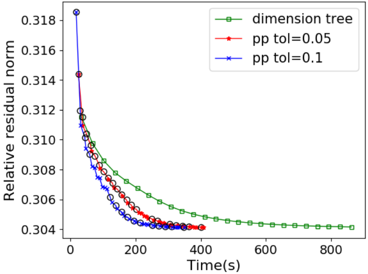
(c) Random tensor on 256 nodes

- Order 6 tensor, dimension $s = \lfloor 32p^{1/6} \rfloor$ and rank $R = \lfloor 4p^{1/6} \rfloor$
- Low-rank with random factor matrices
- Overall convergence up to 3X faster
- Better performance for larger tensors

Results for Real Tensors



(a) Coil Dataset



(b) Time-Lapse Dataset

- Coil Dataset dimension: $128 \times 128 \times 3 \times 7200$
- Time-Lapse Dataset dimension: $1024 \times 1344 \times 33 \times 9$
- Single node (KNL) execution with MPI
- Overall convergence up to 2.5X faster

Summary and Conclusion

- Introduced new pairwise perturbation algorithm to approximate ALS in CP and Tucker decomposition
  - Approximate sweep faster for CP by factor of $O(s^{N-2})$ and for Tucker by factor of $O(s^{N-2}/R^{N-2})$
  - Error scales with change to factor matrices from first PP step
  - For Tucker additional error bounds hold since generally computed result (core tensor) is large in norm
  - Both CP and Tucker ALS with dimension trees and with PP implemented using Cyclops[5]
  - Speed-ups of about 3X for a range of problems on Stampede2 (thanks XSEDE/TACC!)
  - For pseudocodes, analysis, and results, see arXiv:1811.10573

[5] https://github.com/LinjianMa/pairwise-perturbation