Accelerating Alternating Least Squares for Tensor Decomposition by Pairwise Perturbation

Linjian Ma[1] and Edgar Solomonik[2]

LPNA @CS@Illinois

[1]EECS Department
University of California, Berkeley

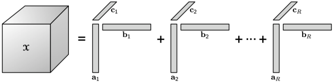
[2]Department of Computer Science
University of Illinois at Urbana-Champaign

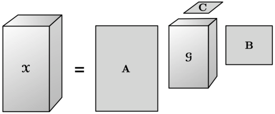
Scientific Computing Seminar
Berkeley, CA

Outline

Overview

CP and Tucker tensor decompositions[1]



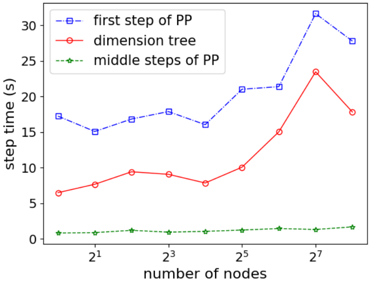


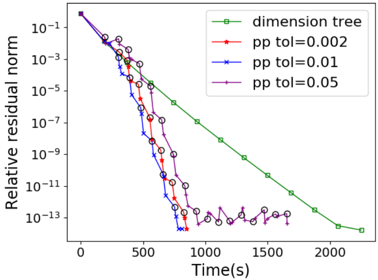
- Alternating least squares (ALS) is most widely used method
- Each ALS sweep optimizes all factor matrices in decomposition
- New algorithm: pairwise perturbation approximates ALS
  - accurate when factor tensors change little at each sweep
  - rank R CP decomposition: it reduces cost of sweep from $O(s^N R)$ to $O(s^2 R)$ for input tensor with dims $s \times \cdots \times s$
  - rank R Tucker decomposition: it reduces cost of sweep from $O(s^N R)$ to $O(s^2 R^{N-1})$

[1] Kolda and Bader, SIAM Review 2009

Performance Highlights for Pairwise Perturbation

Pairwise perturbation (PP) outperforms optimized dimension tree ALS



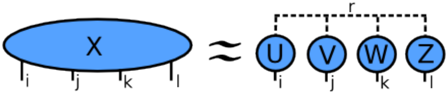


- First step of PP (setup) costs slightly more than ALS sweep
- Middle steps (subsequent approximations) up to 10X faster
- Overall convergence up to 3X faster for synthetic and real tensors

Alternating Least Squares for CP Decomposition

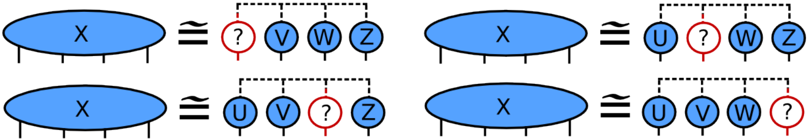
Consider rank R CP decomposition of an $s \times s \times s \times s$ tensor

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

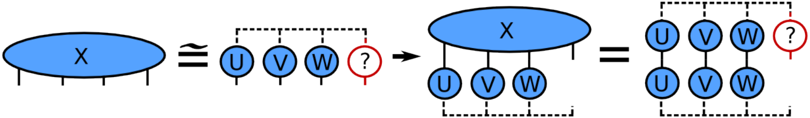


ALS updates factor matrices in an alternating manner

$$\min_{\color{red}{\boldsymbol{A}^{(n)}}} f(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(N)}) = \frac{1}{2} || \boldsymbol{\mathcal{X}} - [\![\boldsymbol{A}^{(1)}, \cdots, \color{red}{\boldsymbol{A}^{(n)}}, \cdots, \boldsymbol{A}^{(N)}]\!] ||_F^2,$$

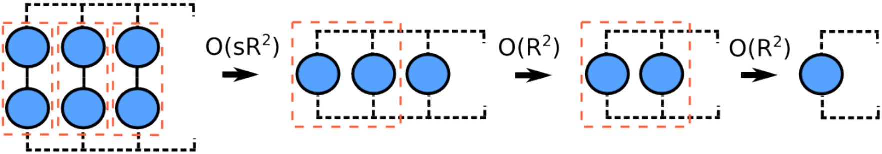


Each quadratic subproblem is typically solved via normal equations

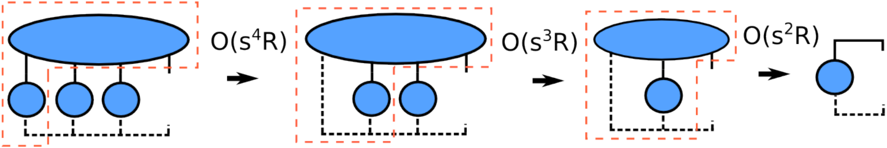


Tensor Contractions in CP ALS

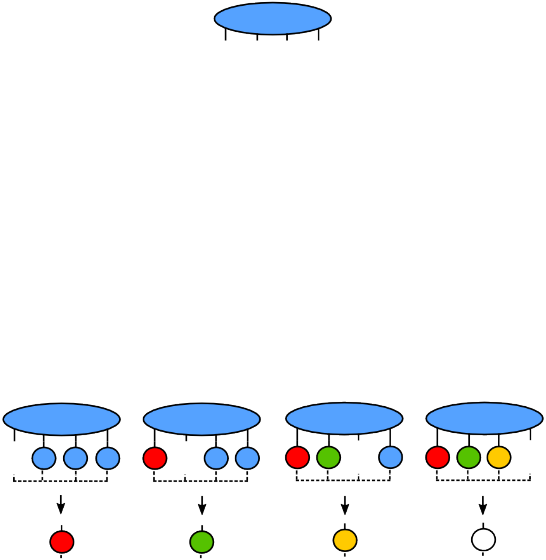
The normal equations are cheap to compute



But forming the right-hand sides ($\boldsymbol{M}^{(n)}$) requires expensive MTTKRP (matricized tensor-times Khatri-Rao product)

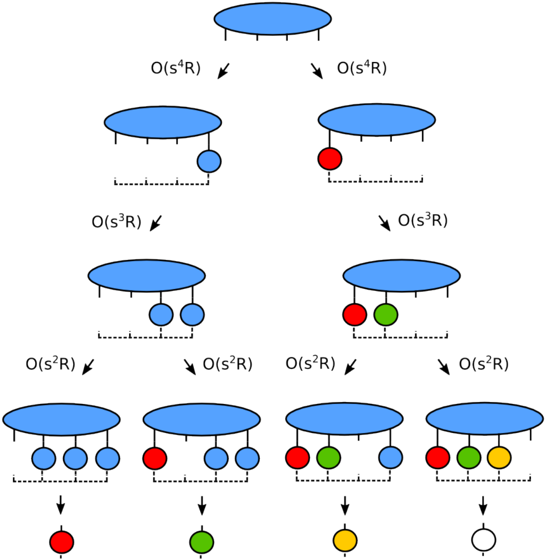


CP ALS Dimension Trees[2]



[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

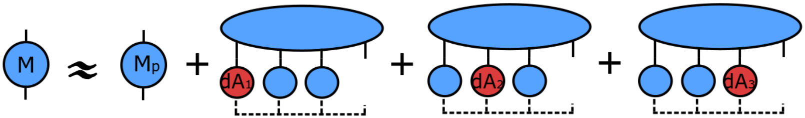
CP ALS with Pairwise Perturbation

Pairwise perturbation (PP) approximates $\boldsymbol{M}^{(n)} \approx \tilde{\boldsymbol{M}}^{(n)}$ using pairwise perturbation operators $\boldsymbol{\mathcal{M}}_p^{(i,n)}$

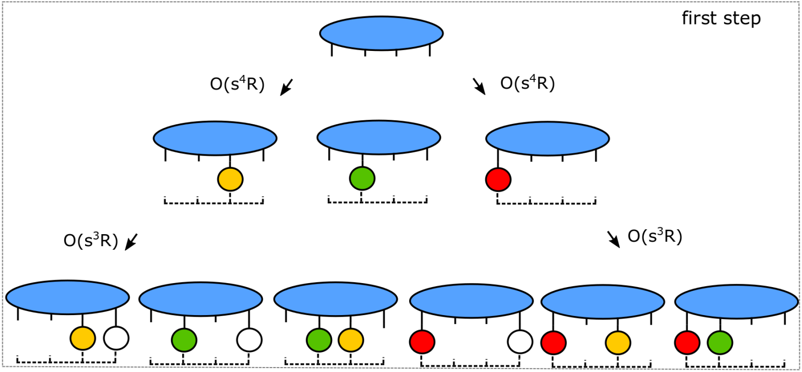
- Write $\boldsymbol{A}^{(n)} = \boldsymbol{A}_p^{(n)} + d\boldsymbol{A}^{(n)} \rightarrow \boldsymbol{M}^{(n)} = \boldsymbol{X}_{(n)} \bigodot_{i=1,i\neq n}^{N} (\boldsymbol{A}_p^{(i)} + d\boldsymbol{A}^{(i)})$
- Elementwise,

$$\boldsymbol{M}^{(n)}(y,k) = \boldsymbol{M}_p^{(n)}(y,k) + \sum_{i=1,i\neq n}^{N} \sum_{x=1}^{s_i} \boldsymbol{\mathcal{M}}_p^{(i,n)}(x,y,k) d\boldsymbol{A}^{(i)}(x,k) +$$

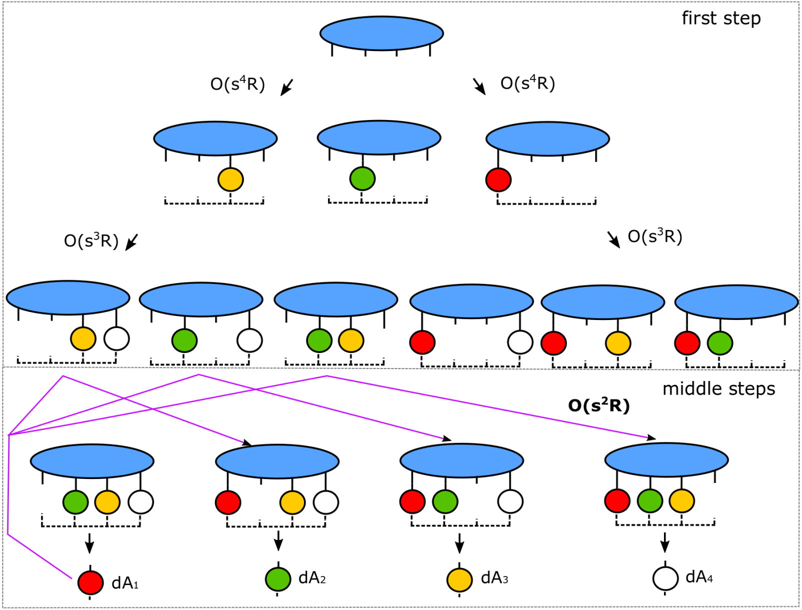
$$\sum_{i=1,i\neq n}^{N} \sum_{j=i+1,j\neq n}^{N} \sum_{x=1}^{s_i} \sum_{z=1}^{s_j} \boldsymbol{\mathcal{M}}_p^{(i,j,n)}(x,z,y,k) d\boldsymbol{A}^{(i)}(x,k) d\boldsymbol{A}^{(j)}(z,k) + \cdots$$



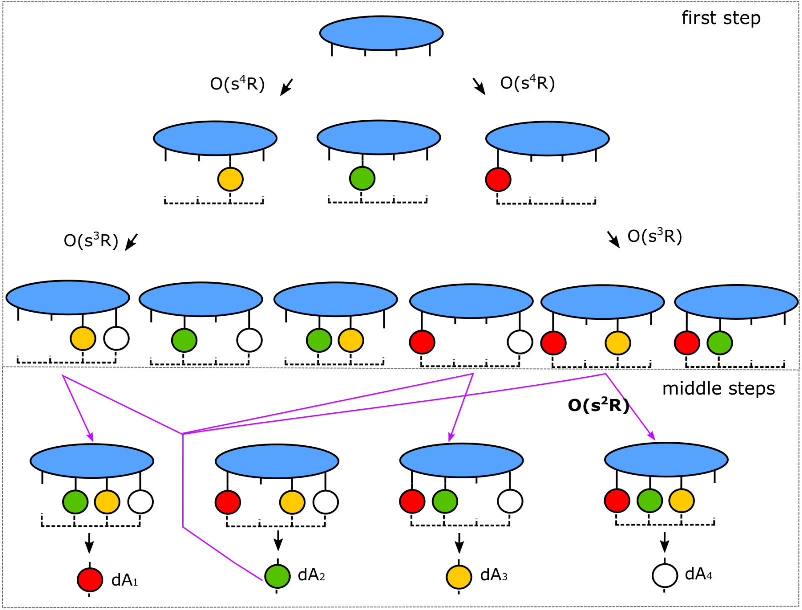
CP ALS with Pairwise Perturbation



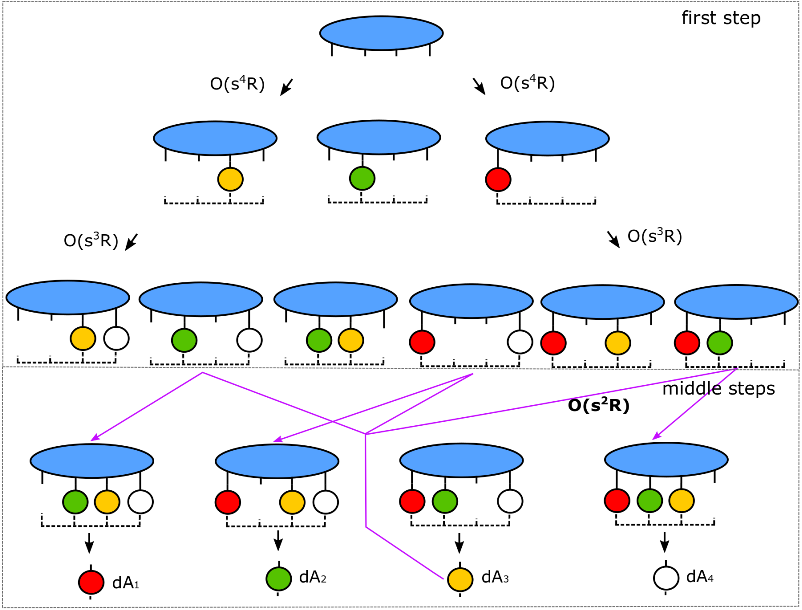
CP ALS with Pairwise Perturbation



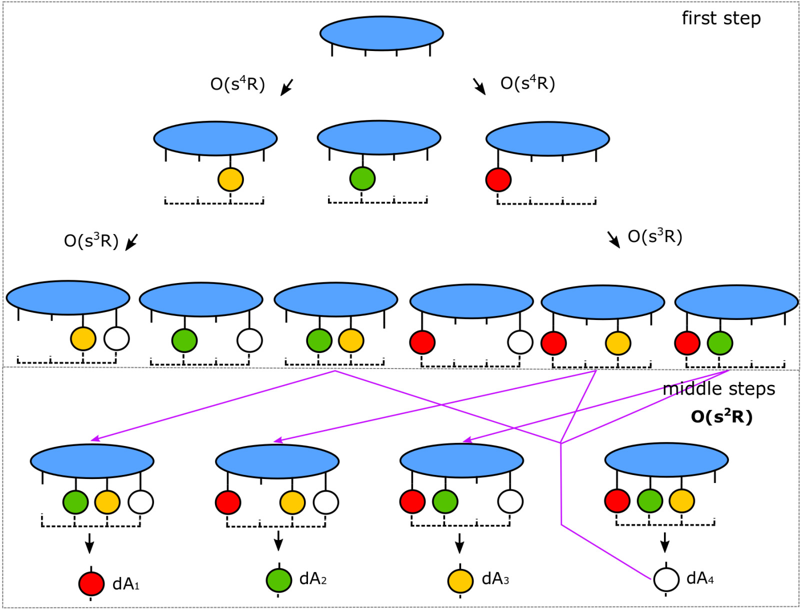
CP ALS with Pairwise Perturbation



CP ALS with Pairwise Perturbation



CP ALS with Pairwise Perturbation



Error Analysis: First Attempt

Consider order $N = 3$ tensor $\boldsymbol{\mathcal{X}}$, 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 $\boldsymbol{f}_{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^s \times \mathbb{R}^s \to \mathbb{R}^s$,

$$\boldsymbol{z} = \boldsymbol{f}_{\boldsymbol{\mathcal{X}}}(\boldsymbol{u}, \boldsymbol{v}) \Rightarrow z_k = \sum_{i,j} x_{ijk} u_i v_j$$

Theorem (Columnwise Error Bound from Tensor Conditioning)

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

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

MTTKRP is Ill-Posed for Most Tensors

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

$$\boldsymbol{z} = \boldsymbol{f_{\mathcal{X}}}(\boldsymbol{u}, \boldsymbol{v}) \Rightarrow z_k = \sum_{i,j} x_{ijk} u_i v_j$$

- If $\min_{\boldsymbol{u},\boldsymbol{v} \in \mathbb{S}^{s-1}} ||\boldsymbol{f_{\mathcal{X}}}(\boldsymbol{u}, \boldsymbol{v})||_2 = 0$, bound is trivial
- There exist $2 \times 2 \times 2$, $4 \times 4 \times 4$, and $8 \times 8 \times 8$ tensors for which $||\boldsymbol{f_{\mathcal{X}}}(\boldsymbol{u}, \boldsymbol{v})||_2 = 1$ for all $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{S}^{s-1}$
  - 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 $\boldsymbol{\mathcal{X}}$ has $\min_{\boldsymbol{u},\boldsymbol{v} \in \mathbb{S}^{s-1}} ||\boldsymbol{f_{\mathcal{X}}}(\boldsymbol{u}, \boldsymbol{v})||_2 = 0$
- 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 $\boldsymbol{\mathcal{X}}$, let $\boldsymbol{M}^{(3)}$ be the right-hand-sides needed to form the third factor matrix $\boldsymbol{A}^{(3)}$

- Define $\boldsymbol{M}^{(3)}_{new} - \boldsymbol{M}^{(3)} = \boldsymbol{H}^{(1,3)} + \boldsymbol{H}^{(2,3)}$
- Define $\boldsymbol{A}^{(i)}_{new} - \boldsymbol{A}^{(i)} = \delta \boldsymbol{A}^{(i)}$
- 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* $\hat{\boldsymbol{T}} = \boldsymbol{\mathcal{X}} \times_1 \delta \boldsymbol{a}_k^{(1)}$,

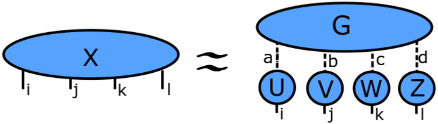
$$\frac{||\tilde{\boldsymbol{h}}_k^{(1,3)} - \boldsymbol{h}_k^{(1,3)}||_2}{||\boldsymbol{h}_k^{(1,3)}||_2} \leq \kappa(\hat{\boldsymbol{T}})\epsilon_k, \text{ where } \kappa(\hat{\boldsymbol{T}}) = \frac{\sigma_{\max}(\hat{\boldsymbol{T}})}{\sigma_{\min}(\hat{\boldsymbol{T}})}$$

- For $N > 3$: higher-order absolute error terms scale as $O(\epsilon_k \epsilon_l)$, but can dominate, so have no relative error bound

Alternating Least Squares for Tucker Decomposition

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

$$x_{ijkl} \approx \sum_{a,b,c,d} g_{abcd} u_{ia} v_{jb} w_{kc} z_{ld}$$



- $\boldsymbol{\mathcal{G}}$ is the core tensor with dimension $R \times R \times R \times R$
- Factor matrices have orthonormal columns
- Tucker Decomposition is usually initialized by HOSVD (Higher Order Singular Value Decomposition)
- Interlaced HOSVD:

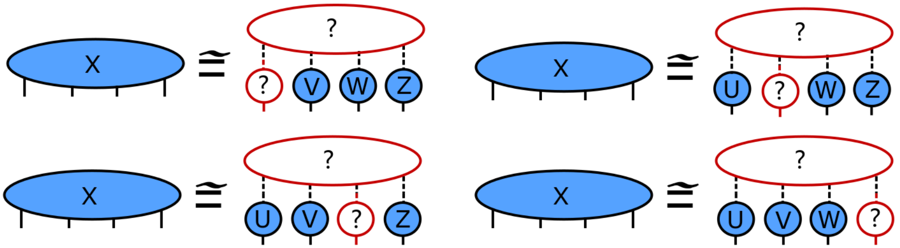
$$\boldsymbol{A}^{(1)} \leftarrow \text{R leading singular vectors of } \boldsymbol{X}^{(1)}$$

$$\boldsymbol{A}^{(2)} \leftarrow \text{R leading singular vectors of } [\boldsymbol{\mathcal{X}} \times_1 \boldsymbol{A}^{(1)T}]^{(2)}$$

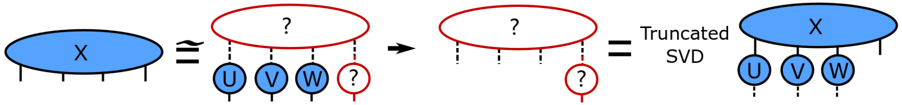
$$\ldots$$

Alternating Least Squares for Tucker Decomposition

ALS updates factor matrices in an alternating manner

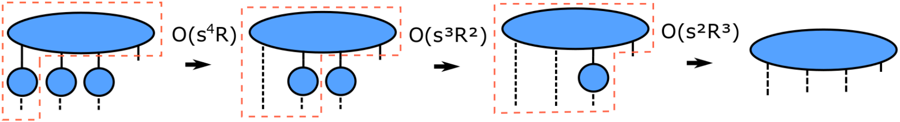


Tucker-ALS is usually solved with HOOI (Higher-Order Orthogonal Iteration)

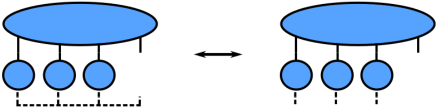


Pairwise Perturbation for Tucker Decomposition

Forming $\mathcal{Y}^{(n)}$ requires the expensive TTMc (Tensor Times Matrix-chain)



- We perform SVD on the Gram Matrix to avoid SVD of the large $\boldsymbol{Y}_{(n)}^{(n)}$
- Similar to MTTKRP in CP, Pairwise can also be applied to TTMc



	State of the art ALS	PP operator construction	PP middle steps
CP	$4s^N R$	$6s^N R$	$2Ns^2R$
Tucker	$4s^N R$	$6s^N R$	$2Ns^2R^{N-1}$

Error Analysis for Tucker: First Bound

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

- Bound relative error of $\tilde{\boldsymbol{\mathcal{Y}}}^{(3)}$ computed by PP middle step
- The ith factor matrix changed by $d\boldsymbol{A}^{(i)}$ since the first step of PP
- The spectral norm of the tensor corresponds to $||\boldsymbol{\mathcal{X}}||_2 = \sup\{||\boldsymbol{f}_{\boldsymbol{\mathcal{X}}}||_2\}$

Theorem (Error Bound with Bounded Residual)

If $||d\boldsymbol{A}^{(l)}||_2 \leq \epsilon \ll 1$ *for* $l \in \{1, 2, 3\}$ *and residual spectral norm* $\leq \frac{1}{3}||\boldsymbol{\mathcal{X}}||_2$,

$$\frac{||\tilde{\boldsymbol{\mathcal{Y}}}^{(3)} - \boldsymbol{\mathcal{Y}}^{(3)}||_2}{||\boldsymbol{\mathcal{Y}}^{(3)}||_2} = O(\epsilon^2).$$

- The error bound is independent of the input tensor conditioning

Error Analysis for Tucker: Second Bound

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

- Bound relative error of $\tilde{\boldsymbol{\mathcal{Y}}}^{(3)}$ computed by PP middle step

Theorem (Error Bound when Tucker starts with interlaced HOSVD)

If $||d\boldsymbol{A}^{(l)}||_F \leq \epsilon \ll 1$ *for* $l \in \{1, 2, 3\}$ *and*
1. interlaced HOSVD is used to initialize Tucker-ALS
2. the decomposition residual is no higher than that attained by HOSVD,

$$\frac{||\tilde{\boldsymbol{\mathcal{Y}}}^{(n)} - \boldsymbol{\mathcal{Y}}^{(n)}||_F}{||\boldsymbol{\mathcal{Y}}^{(n)}||_F} = O\Big(\epsilon^2 \Big(\frac{s}{R}\Big)^{N/2}\Big).$$

- The error bound is also independent of the input tensor conditioning

Implementation

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

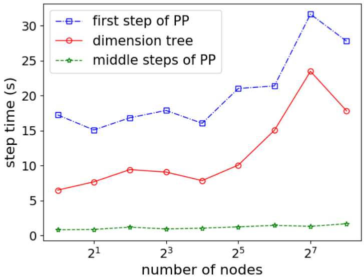
- Cyclops is a C++ library that distributes each tensor over MPI
- Used in chemistry (PySCF, QChem), quantum circuit simulation (IBM/LLNL), and graph analysis (betweenness centrality)
- Summations and contractions specified via Einstein notation

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

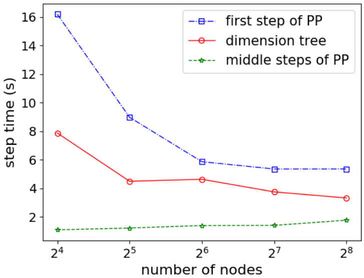
- Best distributed contraction algorithm auto-selected at runtime
- Sparse tensors supported but unused here
- Python interface, OpenMP, and GPU support present but unused
- Used interface to ScaLAPACK SVD to solve linear systems

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

Strong and Weak Scaling Microbenchmarks

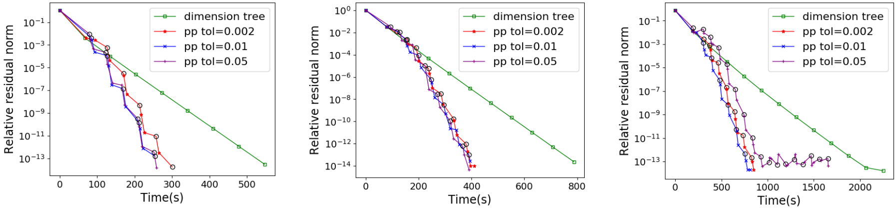


(a) Weak scaling



(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



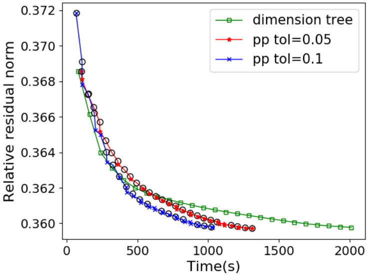
(a) Random tensor on 1 node

(b) Random tensor on 16 nodes

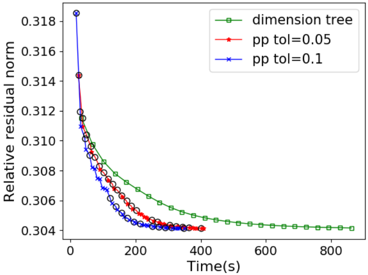
(c) Random tensor on 256 nodes

- 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 (CP)



(a) Coil Dataset



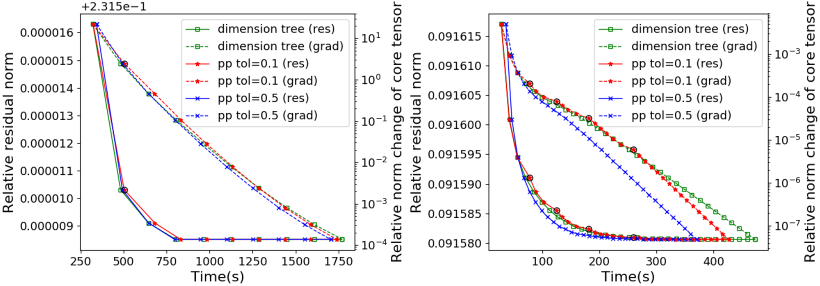
(b) Time-Lapse Dataset

- Coil Dataset[5] dimension: $128 \times 128 \times 3 \times 7200$
- Time-Lapse Dataset[6] dimension: $1024 \times 1344 \times 33 \times 9$
- Single node (KNL) execution with MPI
- Overall convergence up to 2.5X faster

[5] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-100)

[6] S. M. Nascimento, K. Amano, and D. H. Foster. Vision Research, 2016

Results for Real Tensors (Tucker)



(a) Coil Dataset (b) Time-Lapse Dataset

- Coil Dataset dimension: $128 \times 128 \times 3 \times 7200$
- Time-Lapse Dataset dimension: $1024 \times 1344 \times 33 \times 9$
- Overall convergence up to 1.3X faster
- Better performance for larger tensors

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 stronger 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[7]
- Speed-ups of about 3X for a range of problems on Stampede2 (thanks XSEDE/TACC!)
- For pseudocodes, analysis, and results, see arXiv:1811.10573

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