Efficient parallel CP decomposition with pairwise perturbation and multi-sweep dimension tree

Linjian Ma and Edgar Solomonik

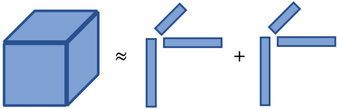
Department of Computer Science,
University of Illinois at Urbana-Champaign

May 2021

Background

CP tensor decompositions

$$\boldsymbol{X} \approx \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r,$$



- Consider $\boldsymbol{X} \in \mathbb{R}^{s\times s\times s}$, and the CP rank $R \ll s$

Alternating least squares (**CP-ALS**) updates factors in an alternating manner

$$\min_{\boldsymbol{a}_1,\dots,\boldsymbol{a}_R} \frac{1}{2}\left\|\sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r - \boldsymbol{X}\right\|_F^2 \quad \rightarrow \quad \min_{\boldsymbol{A}} \frac{1}{2}\left\|(\boldsymbol{C}\odot\boldsymbol{B})\boldsymbol{A}^T - \boldsymbol{X}_{(1)}^T\right\|_F^2$$

$$\min_{\boldsymbol{b}_1,\dots,\boldsymbol{b}_R} \frac{1}{2}\left\|\sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r - \boldsymbol{X}\right\|_F^2 \quad \rightarrow \quad \min_{\boldsymbol{B}} \frac{1}{2}\left\|(\boldsymbol{C}\odot\boldsymbol{A})\boldsymbol{B}^T - \boldsymbol{X}_{(2)}^T\right\|_F^2$$

$$\min_{\boldsymbol{c}_1,\dots,\boldsymbol{c}_R} \frac{1}{2}\left\|\sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r - \boldsymbol{X}\right\|_F^2 \quad \rightarrow \quad \min_{\boldsymbol{C}} \frac{1}{2}\left\|(\boldsymbol{B}\odot\boldsymbol{A})\boldsymbol{C}^T - \boldsymbol{X}_{(3)}^T\right\|_F^2$$

Background of CP-ALS

$$\min_{\mathbf{A}} \frac{1}{2}\left\|(\mathbf{C}\odot\mathbf{B})\mathbf{A}^T - \mathbf{X}_{(1)}^T\right\|_F^2 \quad\Longrightarrow\quad \mathbf{A} \leftarrow \mathbf{X}_{(1)}(\mathbf{C}\odot\mathbf{B})\Big((\mathbf{C}^T\mathbf{C}) * (\mathbf{B}^T\mathbf{B})\Big)^{-1}$$

- $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_R] \in \mathbb{R}^{s\times R}$, similar for $\mathbf{B}$, $\mathbf{C}$
- $\mathbf{X}_{(1)} \in \mathbb{R}^{s\times s^2}$ is the first mode matricization of $\mathbf{X}$
- Matricization in the first mode maps tensor element (i_1, i_2, i_3) to matrix element $(i_1, (i_2-1)s + i_3)$
- $\mathbf{C}\odot\mathbf{B} \in \mathbb{R}^{s^2\times R}$ is the Khatri-Rao Product (KRP) of $\mathbf{C}, \mathbf{B}$
- $\mathbf{C}\odot\mathbf{B} := [\mathbf{c}_1\otimes\mathbf{b}_1, \ldots, \mathbf{c}_R\otimes\mathbf{b}_R]$
- $\mathbf{X}_{(1)}(\mathbf{C}\odot\mathbf{B})$: MTTKRP, main bottleneck, costs $4s^N R$ flops

Our contributions

- We introduce a new **multi-sweep dimension tree (MSDT)** algorithm to reduce the leading computational cost
- We introduce a communication-efficient distributed parallel **pairwise perturbation (PP)** algorithm (focus of the talk)
- At least **1.52X** speed-ups compared to the state-of-the-art CP-ALS

	Computational cost	Horizontal communication cost
ALS	$4s^N R \cdot \gamma$	$\mathcal{O}(N\log(P)\cdot\alpha + NsR/P^{\frac{1}{N}}\cdot\beta)$
MSDT	$\frac{2N}{N-1}s^N R \cdot \gamma$	$\mathcal{O}(N\log(P)\cdot\alpha + NsR/P^{\frac{1}{N}}\cdot\beta)$
PP-init	$4s^N R \cdot \gamma$	/
PP-approx	$2N^2(s^2R + R^2)\cdot\gamma$	$\mathcal{O}(N\log(P)\cdot\alpha + NsR/P^{\frac{1}{N}}\cdot\beta)$

Cost based on the BSP $\alpha - \beta - \gamma$ model

Pairwise perturbation algorithm (L. Ma et al, arxiv 1811.10573)

- Main idea of the PP algorithm: approximate the MTTKRP $\boldsymbol{M}^{(1)} = \boldsymbol{X}_{(1)}\,(\boldsymbol{B} \odot \boldsymbol{C})$
- Let $\boldsymbol{B}_p$ denote the $\boldsymbol{B}$ calculated at some iteration prior to the current one
- $\boldsymbol{B} = \boldsymbol{B}_p + d\boldsymbol{B}$, $\boldsymbol{C} = \boldsymbol{C}_p + d\boldsymbol{C}$

$$
\begin{aligned}
\boldsymbol{M}^{(1)} &= \boldsymbol{X}_{(1)}\Big((\boldsymbol{B}_p + d\boldsymbol{B}) \odot (\boldsymbol{C}_p + d\boldsymbol{C})\Big) \\
&= \boldsymbol{X}_{(1)}(\boldsymbol{B}_p \odot \boldsymbol{C}_p) + \boldsymbol{X}_{(1)}(\boldsymbol{B}_p \odot d\boldsymbol{C}) + \boldsymbol{X}_{(1)}(d\boldsymbol{B} \odot \boldsymbol{C}_p) + \boldsymbol{X}_{(1)}(d\boldsymbol{B} \odot d\boldsymbol{C}) \\
&\approx \boldsymbol{X}_{(1)}(\boldsymbol{B}_p \odot \boldsymbol{C}_p) + \boldsymbol{X}_{(1)}(\boldsymbol{B}_p \odot d\boldsymbol{C}) + \boldsymbol{X}_{(1)}(d\boldsymbol{B} \odot \boldsymbol{C}_p) := \widetilde{\boldsymbol{M}}^{(1)}
\end{aligned}
$$

Pairwise perturbation contains two steps:

- Initialization step: calculates $\boldsymbol{X}_{(1)}(\boldsymbol{B}_p \odot \boldsymbol{C}_p)$, $\boldsymbol{X}_{(1,3)}\boldsymbol{B}_p$, $\boldsymbol{X}_{(1,2)}\boldsymbol{C}_p$ (overall cost $O(s^N R)$)
- Approximated step: finish the calculation of $\boldsymbol{X}_{(1)}(\boldsymbol{B}_p \odot d\boldsymbol{C}), \boldsymbol{X}_{(1)}(d\boldsymbol{B} \odot \boldsymbol{C}_p)$ (overall cost $O(Ns^2R)$)

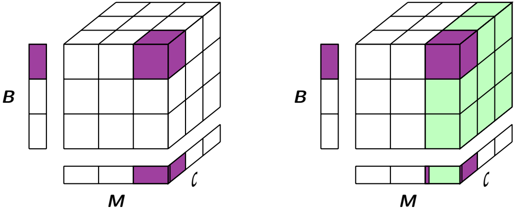
Parallel MTTKRP (G. Ballard et al, HiPC 2018)

Main idea: decompose MTTKRP into local MTTKRPs to reduce the communication cost

- Input tensor uniformly distributed across the processor grid $\mathcal{P}$
- $\boldsymbol{X}_{\mathcal{P}(i,j,k)}$ denotes the local data on the processor $\mathcal{P}(i,j,k)$
- Perform local MTTKRP on each processor

$$\boldsymbol{M}_{\mathcal{P}(i,j,k)} = \text{Local-MTTKRP}(\boldsymbol{X}_{\mathcal{P}(i,j,k)}, \boldsymbol{B}_{\mathcal{P}(i,j,k)}, \boldsymbol{C}_{\mathcal{P}(i,j,k)}),$$

- Perform reduce-scatter on $\boldsymbol{M}_{\mathcal{P}(i,j,k)}$ to uniformly distribute $\boldsymbol{M}$

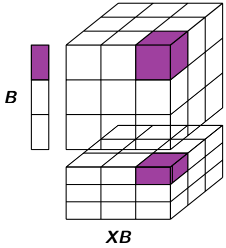


(a) Compute local MTTKRP (b) Perform reduce-scatter

Parallel pairwise perturbation

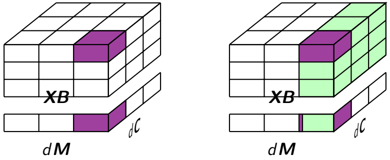
PP initialization step

- Perform local partial MTTKRP contraction $\boldsymbol{XB}_{\mathcal{P}(i,j,k)} = \boldsymbol{X}_{(1,3)\mathcal{P}(i,j,k)}\boldsymbol{B}_{p\mathcal{P}(i,j,k)}$
- No communication needed
- More memory needed



PP approximated step

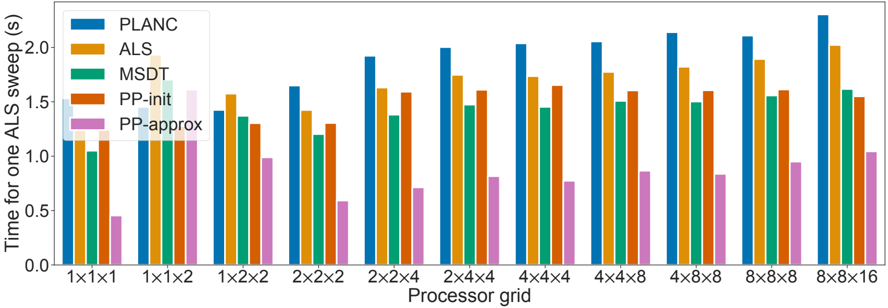
- Computing local MTTKRP updates
- Communication cost same as parallel MTTKRP



(a) Local MTTKRP update

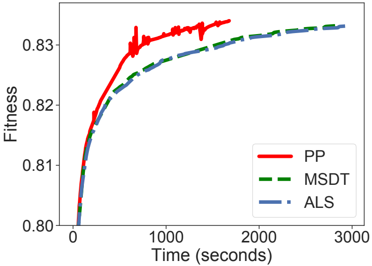
(b) Perform reduce-scatter

Weak scaling benchmark results

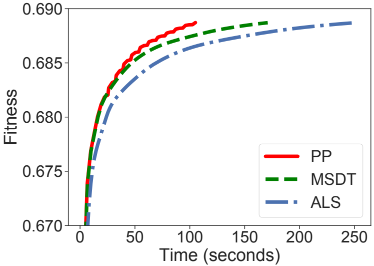


- Rank $R = 400$, and the local tensor size $400 \times 400 \times 400$
- PP approximated step has up to 2X speed-up compared to state-of-art ALS (S. Eswar et al, arxiv 1909.01149)

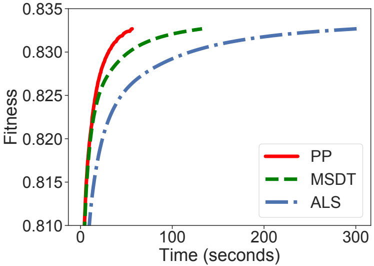
Performance on real data



(a) Quantum chemistry tensor



(b) Coil dataset



(c) Timelapse dataset

- Tensor size: (a) $4520 \times 280 \times 280$ (b) $128 \times 128 \times 3 \times 7200$ (c) $1024 \times 1344 \times 33 \times 9$
- CP rank: (a) $R = 1000$ (b) $R = 20$ (c) $R = 50$
- Up to 5.4X speed-up

Conclusion

Contributions

- We introduce a communication-efficient distributed parallel **pairwise perturbation (PP)** algorithm
- We introduce a new **multi-sweep dimension tree (MSDT)** algorithm to reduce the leading computational cost
- At least **1.52X** speed-ups compared to the state-of-the-art CP-ALS

Other details

- PP and MSDT are vertical communication/memory bandwidth bound