Randomized and approximated algorithms for tensor decompositions

Linjian Ma and Edgar Solomonik

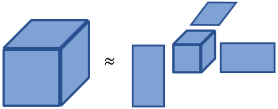
Department of Computer Science,
University of Illinois at Urbana-Champaign

May 2021

Background

Tucker decomposition

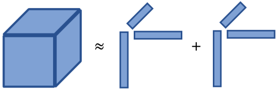
$$\boldsymbol{T} \approx \boldsymbol{X} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C}$$



- $\boldsymbol{T} \in \mathbb{R}^{s\times s\times s}$, $\boldsymbol{X} \in \mathbb{R}^{R\times R\times R}$
- $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{s\times R}$ with orthonormal columns, $R < s$

CP decomposition

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



- $\boldsymbol{T} \in \mathbb{R}^{s\times s\times s}$, $\boldsymbol{A} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_R] \in \mathbb{R}^{s\times R}$
- $R < s^2$

Higher order orthogonal iteration (HOOI)

$$\min_{\boldsymbol{A},\boldsymbol{X}} \frac{1}{2} \left\| (\boldsymbol{C} \otimes \boldsymbol{B}) \boldsymbol{X}_{(1)}^T \boldsymbol{A}^T - \boldsymbol{T}_{(1)}^T \right\|_F^2$$

CP-Alternating least squares (CP-ALS)

$$\min_{\boldsymbol{A}} \frac{1}{2} \left\| (\boldsymbol{C} \odot \boldsymbol{B}) \boldsymbol{A}^T - \boldsymbol{T}_{(1)}^T \right\|_F^2$$

Background

Higher order orthogonal iteration (HOOI)

$$\min_{\boldsymbol{A},\boldsymbol{X}} \frac{1}{2} \left\| (\boldsymbol{C} \otimes \boldsymbol{B}) \boldsymbol{X}_{(1)}^T \boldsymbol{A}^T - \boldsymbol{T}_{(1)}^T \right\|_F^2$$

- Kronecker product $\boldsymbol{C} \otimes \boldsymbol{B} \in \mathbb{R}^{s^2 \times R^2}$
- Costs $\Theta(s^3 R)$ or $\Theta(\text{nnz}(\boldsymbol{T}) R^2)$
- Fast convergence

CP-Alternating least squares (CP-ALS)

$$\min_{\boldsymbol{A}} \frac{1}{2} \left\| (\boldsymbol{C} \odot \boldsymbol{B}) \boldsymbol{A}^T - \boldsymbol{T}_{(1)}^T \right\|_F^2$$

- Khatri-Rao product $\boldsymbol{C} \odot \boldsymbol{B} \in \mathbb{R}^{s^2 \times R}$
- Costs $\Theta(s^3 R)$ or $\Theta(\text{nnz}(\boldsymbol{T}) R)$
- Slow convergence

Low rank approximation ($R \ll s$):

- Sketched HOOI for Tucker decomposition (arxiv 2104.01101)
- Overall cost with t HOOI sweeps reduced to $O\left(\text{nnz}(\boldsymbol{T}) + t\left(sR^3 + R^6\right)\right)$
- Can also accelerate CPD via performing CP-ALS on the Tucker core tensor

General rank approximation:

- Approximate ALS using pairwise perturbation (arxiv 1811.10573, 2010.12056)

Sketched HOOI for Tucker decomposition (arxiv 2104.01101)

HOOI: solve and truncate

$$\min_{\boldsymbol{P}\in\mathbb{R}^{s\times R^2}} \frac{1}{2}\left\|(\boldsymbol{C}\otimes\boldsymbol{B})\boldsymbol{P}^T - \boldsymbol{T}_{(1)}^T\right\|_F^2$$

$\boldsymbol{A}\boldsymbol{X}_{(1)} \leftarrow$ Best rank-R approximation of $\boldsymbol{P}$

Sketched HOOI: sketch, solve and truncate

$$\min_{\widehat{\boldsymbol{P}}\in\mathbb{R}^{s\times R^2}} \frac{1}{2}\left\|\boldsymbol{S}(\boldsymbol{C}\otimes\boldsymbol{B})\widehat{\boldsymbol{P}}^T - \boldsymbol{S}\boldsymbol{T}_{(1)}^T\right\|_F^2$$

$\widehat{\boldsymbol{A}}\widehat{\boldsymbol{X}}_{(1)} \leftarrow$ Best rank-R approximation of $\widehat{\boldsymbol{P}}$

- $\boldsymbol{S}\in\mathbb{R}^{m\times s^2}$ is the sketching matrix, $m < s^2$ is the sketch size
- Sketched **rank-constrained** linear least squares problem
- Sketched solution close to original solution if $\boldsymbol{S}$ satisfies some properties
- Goal: find $\boldsymbol{S}$ such that with high probability

$$\frac{1}{2}\left\|(\boldsymbol{C}\otimes\boldsymbol{B})\widehat{\boldsymbol{X}}_{(1)}^T\widehat{\boldsymbol{A}}^T - \boldsymbol{T}_{(1)}^T\right\|_F^2 \leq (1+O(\epsilon))\frac{1}{2}\left\|(\boldsymbol{C}\otimes\boldsymbol{B})\boldsymbol{X}_{(1)}^T\boldsymbol{A}^T - \boldsymbol{T}_{(1)}^T\right\|_F^2$$

Sketched HOOI for Tucker decomposition

Theorem: Sketched HOOI with accurate sketching matrix

Let $\boldsymbol{S} \in \mathbb{R}^{m \times s}$ be a $(1/2, \delta, \epsilon)$-accurate sketching matrix for the LHS $\boldsymbol{C} \otimes \boldsymbol{B}$. Then we have with probability at least $1 - \delta$,

$$\frac{1}{2}\left\|(\boldsymbol{C} \otimes \boldsymbol{B})\widehat{\boldsymbol{X}}_{(1)}^{T}\widehat{\boldsymbol{A}}^{T} - \boldsymbol{T}_{(1)}^{T}\right\|_F^2 \leq (1 + O(\epsilon))\frac{1}{2}\left\|(\boldsymbol{C} \otimes \boldsymbol{B})\boldsymbol{X}_{(1)}^{T}\boldsymbol{A}^{T} - \boldsymbol{T}_{(1)}^{T}\right\|_F^2.$$

Sketching matrices satisfying the $(1/2, \delta, \epsilon)$-accurate property

- TensorSketch (R. Pagh, TOCT 2013) with $m = O\left(R^2/\delta \cdot (R^2 + 1/\epsilon^2)\right)$
- Leverage score sampling with $m = O\left(R^2/(\epsilon^2\delta)\right)$
- Sketch size upper bounds are at most $O(1/\epsilon)$ times the upper bounds for unconstrained linear least squares problem

ALS + TensorSketch (Malik and Becker, NeurIPS 2018)

- Solving each factor matrix or the core tensor at a time
- $\min_{\boldsymbol{A}} \frac{1}{2} \left\| (\boldsymbol{C} \otimes \boldsymbol{B}) \boldsymbol{X}_{(1)}^T \boldsymbol{A}^T - \boldsymbol{T}_{(1)}^T \right\|_F^2$ or $\min_{\boldsymbol{X}} \frac{1}{2} \|(\boldsymbol{C} \otimes \boldsymbol{B} \otimes \boldsymbol{A}) \text{vec}(\boldsymbol{X}) - \text{vec}(\boldsymbol{T})\|_F^2$

Algorithm for Tucker	LS subproblem cost	Sketch size (m)
HOOI	$O(\text{nnz}(\boldsymbol{T}) R^2)$	/
ALS + TensorSketch	$\tilde{O}(msR + mR^3)$	$O(R^2/\delta \cdot (R^2 + 1/\epsilon))$
HOOI + TensorSketch	$O(msR + mR^4)$	$O(R^2/\delta \cdot (R^2 + 1/\epsilon^2))$
HOOI + leverage scores	$O(msR + mR^4)$	$O(R^2/(\epsilon^2 \delta))$

Sketched HOOI algorithm

Input: Input order N tensor $\boldsymbol{T}$, Tucker rank R, number of sweeps I_{max}, tolerance ϵ

Output: $\left\{\boldsymbol{X}, \boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(N)}\right\}$

For $n \in \{2, ..., N\}$ **do**

  $\boldsymbol{A}^{(n)} \leftarrow \texttt{Init-RRF}(\boldsymbol{T}_{(n)}, R, \epsilon)$ // Initialize with randomized range finder

Endfor

For $i \in \{1, ..., I_{max}\}$ **do**

  For $n \in \{1, ..., N\}$ **do**

    Build the sketching matrix $\boldsymbol{S}$

    $\boldsymbol{Y} \leftarrow \boldsymbol{S}\boldsymbol{T}_{(n)}$

    $\boldsymbol{Z} \leftarrow \boldsymbol{S}^{(n)}(\boldsymbol{A}^{(1)} \otimes \cdots \otimes \boldsymbol{A}^{(n-1)} \otimes \boldsymbol{A}^{(n+1)} \otimes \cdots \otimes \boldsymbol{A}^{(N)})$

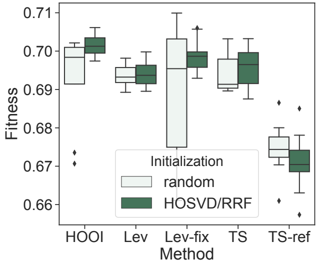
    $\boldsymbol{X}_{(n)}^{T}, \boldsymbol{A}^{(n)} \leftarrow \texttt{Solve-truncate}(\boldsymbol{Z}, \boldsymbol{Y}, R)$

  Endfor

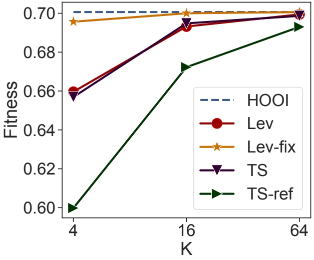
Endfor

Return $\left\{\boldsymbol{X}, \boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(N)}\right\}$

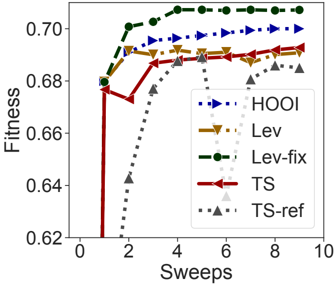
Experiments: tensors with spiked signal



(a) 5 sweeps, sample size $16R^2$



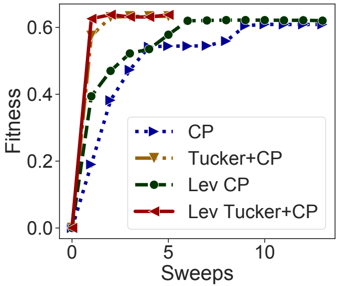
(b) 5 sweeps, sample size KR^2



(c) sample size $16R^2$

- $\boldsymbol{T} = \boldsymbol{T}_0 + \sum_{i=1}^{5} \lambda_i \boldsymbol{a}_i \circ \boldsymbol{b}_i \circ \boldsymbol{c}_i$, each $\boldsymbol{a}_i, \boldsymbol{b}_i, \boldsymbol{c}_i$ has unit 2-norm, $\lambda_i = 3\frac{\|\boldsymbol{T}_0\|_F}{i^{1.5}}$
- Leading low-rank components obey the power-law distribution
- Tensor size $200 \times 200 \times 200$, $R = 5$
- Lev-fix: leverage score deterministic sampling. TS-ref: (Malik and Becker, NeurIPS 2018)

Experiments: CP decomposition



- $\boldsymbol{T} = \sum_{i=1}^{R_{\text{true}}} \boldsymbol{a}_i \circ \boldsymbol{b}_i \circ \boldsymbol{c}_i, \; R_{\text{true}}/R = 1.2$
- Tensor size $2000 \times 2000 \times 2000$, $R = 10$, sample size $16R^2$
- Lev CP: leverage score sampling for CP-ALS (Larsen and Kolda, arXiv:2006.16438)
- Tucker+CP: Run Tucker HOOI first, then run CP-ALS on the Tucker core
- Run Tucker HOOI with 5 sweeps, CP-ALS with 25 sweeps

Accelerate CP-ALS using pairwise perturbation (arxiv 1811.10573, 2010.12056)

- 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 $O(s^3R)$)
- 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 $O(s^2R)$)

At least **1.52X** speed-ups compared to the state-of-the-art distributed parallel CP-ALS

Conclusion

Low rank approximation ($R \ll s$):

- Sketched HOOI for Tucker decomposition
- Overall cost with t HOOI sweeps reduced to $O\left(\text{nnz}(\boldsymbol{\mathcal{T}}) + t\left(sR^N + R^{3(N-1)}\right)\right)$
- Can also accelerate CPD via performing CP-ALS on the Tucker core tensor

General rank approximation:

- Approximate ALS using pairwise perturbation

References:

- Ma, L., & Solomonik, E. Fast and accurate randomized algorithms for low-rank tensor decompositions. arXiv:2104.01101.
- Ma, L., & Solomonik, E. Accelerating alternating least squares for tensor decomposition by pairwise perturbation. arXiv:1811.10573.
- Ma, L., & Solomonik, E. Efficient parallel CP decomposition with pairwise perturbation and multi-sweep dimension tree. arXiv:2010.12056 (also appear at IPDPS 2021).

Initialization with randomized range finder (RRF)

- Initialization with HOSVD is expensive
- For leverage score sampling, random initialization may results in low accuracy

Initialization with randomized range finder

Input: Matrix $\boldsymbol{T}_{(1)} \in \mathbb{R}^{s \times s^2}$, rank R, tolerance ϵ
Output: Good rank-R column subspace of $\boldsymbol{T}_{(1)}$
Initialize $\boldsymbol{S} \in \mathbb{R}^{s^2 \times k}$ with $k = O(R/\epsilon)$
$\boldsymbol{B} \leftarrow \boldsymbol{T}_{(1)}\boldsymbol{S}$
$\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V} \leftarrow \texttt{SVD}(\boldsymbol{B})$
Return $\boldsymbol{U}(:, : R)$

- $\boldsymbol{S}$ is a composite matrix, $\boldsymbol{S} = \boldsymbol{T}\boldsymbol{G}$
- $\boldsymbol{T} \in \mathbb{R}^{s^2 \times O(R^2 + R/\epsilon)}$ is a countsketch matrix
- $\boldsymbol{G} \in \mathbb{R}^{O(R^2 + R/\epsilon) \times k}$ is a random Gaussian embedding
- $\boldsymbol{S}$ is a $(1 + O(\epsilon))$-accurate best rank-R column space
- $\boldsymbol{T}_{(1)}\boldsymbol{S}$ costs $O(\text{nnz}(\boldsymbol{T}) + sR^3/\epsilon)$

Sketched HOOI for Tucker decomposition

Theorem: Sketched HOOI with accurate sketching matrix

Let $\boldsymbol{S} \in \mathbb{R}^{m \times s}$ be a $(1/2, \delta, \epsilon)$-accurate sketching matrix for the LHS $\boldsymbol{C} \otimes \boldsymbol{B}$. Then we have with probability at least $1 - \delta$,

$$\frac{1}{2}\left\|(\boldsymbol{C} \otimes \boldsymbol{B})\widehat{\boldsymbol{X}}_{(1)}^{T}\widehat{\boldsymbol{A}}^{T} - \boldsymbol{T}_{(1)}^{T}\right\|_F^2 \leq (1 + O(\epsilon))\frac{1}{2}\left\|(\boldsymbol{C} \otimes \boldsymbol{B})\boldsymbol{X}_{(1)}^{T}\boldsymbol{A}^{T} - \boldsymbol{T}_{(1)}^{T}\right\|_F^2.$$

$(1/2, \delta, \epsilon)$-accurate sketching matrix for $\boldsymbol{L}$

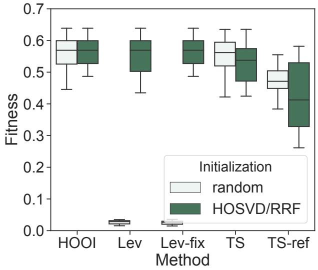
- With probability at least $1 - \delta/2$, each singular value σ of $\boldsymbol{S}\boldsymbol{Q}_L$ satisfies

$$1 - 1/2 \leq \sigma^2 \leq 1 + 1/2$$

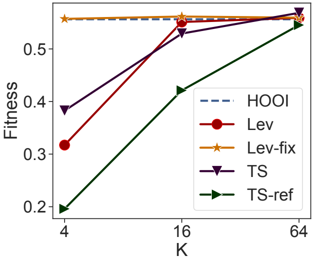
- With probability at least $1 - \delta/2$, for any fixed matrix $\boldsymbol{M}$

$$\|\boldsymbol{Q}_L^T\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{M} - \boldsymbol{Q}_L^T\boldsymbol{M}\|_F^2 \leq \epsilon^2 \cdot \|\boldsymbol{M}\|_F^2$$

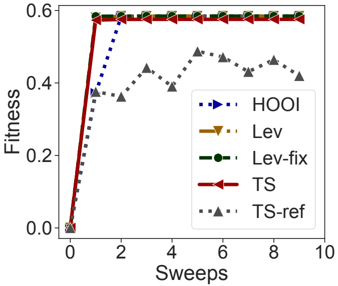
Experiments: tensors with large coherence



(a) 5 sweeps, sample size $16R^2$, n=10

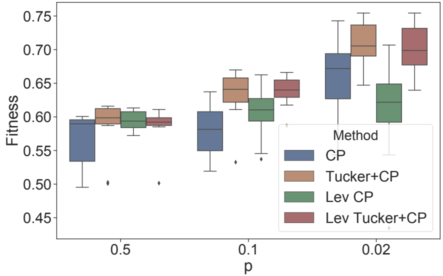


(b) 5 sweeps, sample size KR^2, n=10

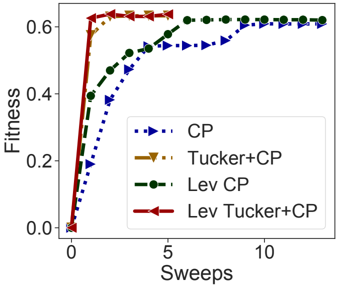


(c) sample size $16R^2$, n=10

- $\boldsymbol{T} = \boldsymbol{T}_0 + \boldsymbol{N}$, $\boldsymbol{T}_0$ uniform random tensor
- $\boldsymbol{N}$ contains $n \ll s$ elements, each with the distribution $\mathcal{N}(\|\boldsymbol{T}_0\|_F/\sqrt{n}, 1)$
- Large coherence: tensor have large variability in magnitudes
- Tensor size $1000 \times 1000 \times 1000$, $R = 5$
- RRF initialization is necessary for leverage score sampling



(a) Tensor size 2000 × 2000 × 2000, $R = 10$, sample size $16R^2$



(b)

- $\boldsymbol{T} = \sum_{i=1}^{R_{\text{true}}} \boldsymbol{a}_i \circ \boldsymbol{b}_i \circ \boldsymbol{c}_i$, $R_{\text{true}}/R = 1.2$
- Lev CP: leverage score sampling for CP-ALS (Larsen and Kolda, arXiv:2006.16438)
- Tucker+CP: Run Tucker HOOI first, then run CP-ALS on the Tucker core
- Run Tucker HOOI with 5 sweeps, CP-ALS with 25 sweeps

Cost comparison for general order N tensors

ALS + TensorSketch (Malik and Becker, NeurIPS 2018)

- Solving each factor matrix or the core tensor at a time
- $\min_{\mathbf{A}} \frac{1}{2} \left\| (\mathbf{C} \otimes \mathbf{B}) \mathbf{X}_{(1)}^T \mathbf{A}^T - \mathbf{T}_{(1)}^T \right\|_F^2$ or $\min_{\mathbf{X}} \frac{1}{2} \| (\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{A}) \text{vec}(\mathbf{X}) - \text{vec}(\mathbf{T}) \|_F^2$

Algorithm for Tucker	LS subproblem cost	Sketch size (m)
HOOI	$O(\text{nnz}(\mathbf{T}) R^{N-1})$	/
ALS + TensorSketch	$\tilde{O}(msR + mR^N)$	$O((3R)^{(N-1)}/\delta \cdot (R^{(N-1)} + 1/\epsilon))$
HOOI + TensorSketch	$O(msR + mR^{2(N-1)})$	$O((3R)^{(N-1)}/\delta \cdot (R^{(N-1)} + 1/\epsilon^2))$
HOOI + leverage scores	$O(msR + mR^{2(N-1)})$	$O(R^{(N-1)}/(\epsilon^2 \delta))$