Fast and accurate randomized algorithms for low-rank tensor decompositions

Linjian Ma and Edgar Solomonik

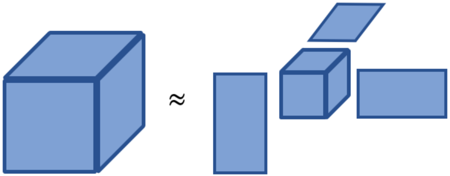
Department of Computer Science,
University of Illinois at Urbana-Champaign

NeurIPS 2021

Background

Tucker decomposition (Tucker, Psychometrika 1966)

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



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

Background

Higher order orthogonal iteration (HOOI)

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

- Update sequence: $(\boldsymbol{A},\boldsymbol{X}),(\boldsymbol{B},\boldsymbol{X}),(\boldsymbol{C},\boldsymbol{X})$
- Fast convergence

Alternating unconstrained least squares (AULS)

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

$$\min_{\boldsymbol{X}} \| (\boldsymbol{C} \otimes \boldsymbol{B} \otimes \boldsymbol{A}) \mathrm{vec}(\boldsymbol{X}) - \mathrm{vec}(\boldsymbol{T}) \|_F^2$$

- Update sequence: $(\boldsymbol{A}),(\boldsymbol{B}),(\boldsymbol{C}),(\boldsymbol{X})$
- Slow convergence

- Kronecker product $\boldsymbol{C} \otimes \boldsymbol{B} \in \mathbb{R}^{s^2 \times R^2}$
- Both algorithms cost $\Theta(s^3 R)$ (dense case) or $\Omega(\mathrm{nnz}(\boldsymbol{T})R)$ (sparse case)
- HOOI is most widely used (Andersson and Bro, 1998)

Background

Low rank Tucker decomposition ($R \ll s$):

- Previous work: sketched Tucker-AULS (Malik and Becker, NeurIPS 2018)
- Advantage: overall cost with t HOOI sweeps reduced from $\Omega(t\text{nnz}(\boldsymbol{T})R)$ to $O\left(\text{nnz}(\boldsymbol{T}) + t\left(sR^5 + R^7\right)\right)$
- Disadvantage: slow convergence since based on Tucker-AULS

Our contribution: a new sketched HOOI algorithm

- An error bound for the sketched **rank-constrained** linear least squares problem arising in Tucker
- **Efficiency**: per sweep cost comparable to sketched Tucker-AULS
- **Accuracy**: reach better decomposition accuracy compared to sketched Tucker-AULS

Sketched HOOI for Tucker decomposition

Let $\boldsymbol{L} = \boldsymbol{C} \otimes \boldsymbol{B}$, $\boldsymbol{Y} = \boldsymbol{T}_{(1)}^T$

HOOI: solve and truncate

$$\boldsymbol{P}_{\text{opt}} \leftarrow \operatorname*{argmin}_{\boldsymbol{P} \in \mathbb{R}^{s \times R^2}} \left\| \boldsymbol{L}\boldsymbol{P}^T - \boldsymbol{Y} \right\|_F^2$$

$$\boldsymbol{A}\boldsymbol{X}_{(1)} \leftarrow \boldsymbol{P}_R$$

Sketched HOOI: sketch, solve and truncate

$$\widehat{\boldsymbol{P}}_{\text{opt}} \leftarrow \operatorname*{argmin}_{\boldsymbol{P} \in \mathbb{R}^{s \times R^2}} \left\| \boldsymbol{S}\boldsymbol{L}\boldsymbol{P}^T - \boldsymbol{S}\boldsymbol{Y} \right\|_F^2$$

$$\widehat{\boldsymbol{A}}\widehat{\boldsymbol{X}}_{(1)} \leftarrow \widehat{\boldsymbol{P}}_R$$

- $\boldsymbol{P}_R, \widehat{\boldsymbol{P}}_R$: best rank-$R$ approximation of $\boldsymbol{P}_{\text{opt}}, \widehat{\boldsymbol{P}}_{\text{opt}}$
- $\boldsymbol{S} \in \mathbb{R}^{m \times s^2}$: sketching matrix, $m < s^2$ is the sketch size
- $\boldsymbol{L}$ has **orthonormal columns**
- Sketched **rank-constrained** linear least squares problem

Sketched rank-constrained linear least squares problem

Goal: find embeddings $\boldsymbol{S}$ such that with probability at least $1-\delta$

$$\left\|\boldsymbol{L}\widehat{\boldsymbol{P}}_R - \boldsymbol{Y}\right\|_F^2 \leq (1+O(\epsilon))\,\|\boldsymbol{L}\boldsymbol{P}_R - \boldsymbol{Y}\|_F^2$$

Main theorem: above inequality holds when $\boldsymbol{S}$ is a $(1/2, \delta, \epsilon)$-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 $\boldsymbol{M}$

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

Sketched rank-constrained linear least squares problem

Goal: find embeddings $\boldsymbol{S}$ such that with probability at least $1-\delta$

$$\left\|\boldsymbol{L}\widehat{\boldsymbol{P}}_R - \boldsymbol{Y}\right\|_F^2 \leq (1+O(\epsilon))\left\|\boldsymbol{L}\boldsymbol{P}_R - \boldsymbol{Y}\right\|_F^2$$

Main theorem: above inequality holds when $\boldsymbol{S}$ is a $(1/2, \delta, \epsilon)$-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 $\boldsymbol{M}$

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

Sketched rank-constrained linear least squares problem

Proof sketch: when $\boldsymbol{S}$ is a $(1/2, \delta, \epsilon)$-accurate sketching matrix

$$\|\boldsymbol{L}\boldsymbol{P}_R - \boldsymbol{Y}\|_F^2 = \left\|\boldsymbol{Y}^\perp\right\|_F^2 + \underbrace{\|\boldsymbol{P}_R - \boldsymbol{P}_{\text{opt}}\|_F^2}_{\text{low rank truncation error}}$$

$$\left\|\boldsymbol{L}\widehat{\boldsymbol{P}}_R - \boldsymbol{Y}\right\|_F^2 = \left\|\boldsymbol{Y}^\perp\right\|_F^2 + \underbrace{\left\|\widehat{\boldsymbol{P}}_{\text{opt}} - \boldsymbol{P}_{\text{opt}}\right\|_F^2}_{\text{sketched least squares error}} + \underbrace{\left\|\widehat{\boldsymbol{P}}_{\text{R}} - \widehat{\boldsymbol{P}}_{\text{opt}}\right\|_F^2 + 2\left\langle \widehat{\boldsymbol{P}}_{\text{R}} - \widehat{\boldsymbol{P}}_{\text{opt}}, \widehat{\boldsymbol{P}}_{\text{opt}} - \boldsymbol{P}_{\text{opt}}\right\rangle_F}_{\text{sketched low rank truncation error}}$$

- $\left\|\widehat{\boldsymbol{P}}_{\text{opt}} - \boldsymbol{P}_{\text{opt}}\right\|_F^2 = O(\epsilon^2)\left\|\boldsymbol{Y}^\perp\right\|_F^2$ (Drineas et al., Numerische mathematik 2011)
- $\left\|\widehat{\boldsymbol{P}}_{\text{R}} - \widehat{\boldsymbol{P}}_{\text{opt}}\right\|_F^2 = \|\boldsymbol{P}_R - \boldsymbol{P}_{\text{opt}}\|_F^2 + O(\epsilon)\|\boldsymbol{L}\boldsymbol{P}_R - \boldsymbol{Y}\|_F^2$ (Mirsky's inequality, (Mirsky, The Quarterly journal of mathematics, 1960))
- $\left\langle \widehat{\boldsymbol{P}}_{\text{R}} - \widehat{\boldsymbol{P}}_{\text{opt}}, \widehat{\boldsymbol{P}}_{\text{opt}} - \boldsymbol{P}_{\text{opt}}\right\rangle_F = O(\epsilon)\|\boldsymbol{L}\boldsymbol{P}_R - \boldsymbol{Y}\|_F^2$ (Mirsky's inequality)

Sketched rank-constrained linear least squares problem

Proof sketch: when $\boldsymbol{S}$ is a $(1/2, \delta, \epsilon)$-accurate sketching matrix

$$\|\boldsymbol{L}\boldsymbol{P}_R - \boldsymbol{Y}\|_F^2 = \left\|\boldsymbol{Y}^\perp\right\|_F^2 + \underbrace{\|\boldsymbol{P}_R - \boldsymbol{P}_{\text{opt}}\|_F^2}_{\text{low rank truncation error}}$$

$$\left\|\boldsymbol{L}\widehat{\boldsymbol{P}}_R - \boldsymbol{Y}\right\|_F^2 = \left\|\boldsymbol{Y}^\perp\right\|_F^2 + \underbrace{\left\|\widehat{\boldsymbol{P}}_{\text{opt}} - \boldsymbol{P}_{\text{opt}}\right\|_F^2}_{\text{sketched least squares error}} + \underbrace{\left\|\widehat{\boldsymbol{P}}_{\text{R}} - \widehat{\boldsymbol{P}}_{\text{opt}}\right\|_F^2 + 2\left\langle \widehat{\boldsymbol{P}}_{\text{R}} - \widehat{\boldsymbol{P}}_{\text{opt}}, \widehat{\boldsymbol{P}}_{\text{opt}} - \boldsymbol{P}_{\text{opt}} \right\rangle_F}_{\text{sketched low rank truncation error}}$$

- $\left\|\widehat{\boldsymbol{P}}_{\text{opt}} - \boldsymbol{P}_{\text{opt}}\right\|_F^2 = O(\epsilon^2)\left\|\boldsymbol{Y}^\perp\right\|_F^2$ (Drineas et al., Numerische mathematik 2011)
- $\left\|\widehat{\boldsymbol{P}}_{\text{R}} - \widehat{\boldsymbol{P}}_{\text{opt}}\right\|_F^2 = \|\boldsymbol{P}_R - \boldsymbol{P}_{\text{opt}}\|_F^2 + O(\epsilon)\|\boldsymbol{L}\boldsymbol{P}_R - \boldsymbol{Y}\|_F^2$ (Mirsky's inequality, (Mirsky, The Quarterly journal of mathematics, 1960))
- $\left\langle \widehat{\boldsymbol{P}}_{\text{R}} - \widehat{\boldsymbol{P}}_{\text{opt}}, \widehat{\boldsymbol{P}}_{\text{opt}} - \boldsymbol{P}_{\text{opt}} \right\rangle_F = O(\epsilon)\|\boldsymbol{L}\boldsymbol{P}_R - \boldsymbol{Y}\|_F^2$ (Mirsky's inequality)

Sketched rank-constrained linear least squares problem

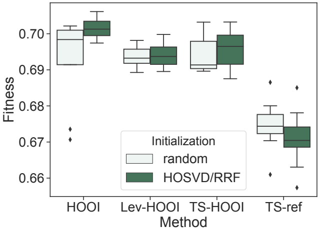
	rank-constrained LS with $\boldsymbol{L}$ having orthonormal columns	unconstrained LS
$\boldsymbol{S}$	$(1/2, \delta, \epsilon)$-accurate	$(1/2, \delta, \sqrt{\epsilon})$-accurate

- Tighter bound on ϵ for $\boldsymbol{S}$ is needed for rank-constrained LS to be $(1 + O(\epsilon))$-accurate

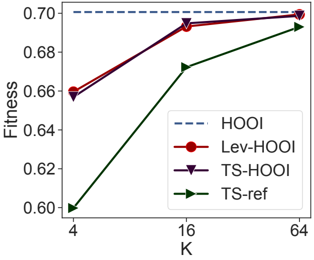
Cost comparison for order 3 tensor

Algorithm for Tucker	LS subproblem cost	Sketch size (m)
HOOI	$\Omega(\text{nnz}(\boldsymbol{T})R)$	/
AULS + 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 sampling	$O(msR + mR^4)$	$O(R^2/(\epsilon^2\delta))$

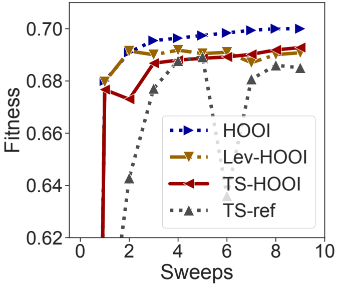
- TensorSketch: a tensorized CountSketch (Pagh, TOCT 2013)
- Leverage score sampling: Importance sampling based on the leverage score of any orthogonal space of $\boldsymbol{L}$



(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 $200 \times 200 \times 200$, $R = 5$
- TS-ref: sketched AULS with TensorSketch (Malik and Becker, NeurIPS 2018)

Conclusion

Our main contributions:

- Sketched HOOI for Tucker decomposition
- New error analysis for sketched rank-constrained linear least squares
- Overall cost with t HOOI sweeps reduced from $\Omega(t\text{nnz}(\boldsymbol{T})R)$ to $O\left(\text{nnz}(\boldsymbol{T}) + t\left(sR^N + R^{3(N-1)}\right)\right)$

Other details in the paper

- Detailed comparison of TensorSketch and leverage score sampling in terms of efficiency and accuracy
- An initialization scheme based on randomized range finder that improves the accuracy of leverage score sampling based sketching
- CP decomposition can be more efficiently calculated based on the sketched Tucker + CP method