Low Rank Approximation
Lecture 4

Daniel Kressner
Chair for Numerical Algorithms and HPC
Institute of Mathematics, EPFL
daniel.kressner@epfl.ch

The CP decomposition

CP decomposition

- Aim: Generalize concept of low rank from matrices to tensors.
- One possibility motivated by

$$\begin{aligned} X &= \big[a_1,\ a_2,\ \ldots,\ a_R\big]\big[b_1,\ b_2,\ \ldots,\ b_R\big]^T = \\ &= a_1 b_1^T + a_2 b_2^T + \cdots + a_R b_R^T. \end{aligned}$$

$\rightsquigarrow$ vectorization

$$\mathrm{vec}(X) = b_1 \otimes a_1 + b_2 \otimes a_2 + \cdots + b_R \otimes a_R.$$

Canonical **P**olyadic decomposition of tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ defined via

$$\begin{aligned} \mathrm{vec}(\mathcal{X}) &= c_1 \otimes b_1 \otimes a_1 + c_2 \otimes b_2 \otimes a_2 + \cdots + c_R \otimes b_R \otimes a_R \\ \mathcal{X} &= a_1 \circ b_1 \circ c_1 + a_2 \circ b_2 \circ c_2 + \cdots + a_R \circ b_R \circ c_R \end{aligned}$$

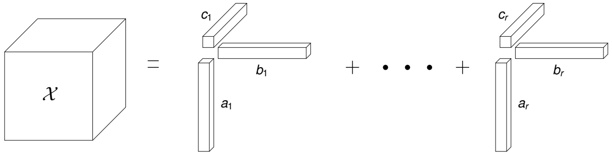
for vectors $a_j \in \mathbb{R}^{n_1}$, $b_j \in \mathbb{R}^{n_2}$, $c_j \in \mathbb{R}^{n_3}$.

CP directly corresponds to semi-separable approximation.
Tensor rank of $\mathcal{X}$ = minimal possible R

CP decomposition

Illustration of CP decomposition

$$\mathcal{X} = a_1 \circ b_1 \circ c_1 + a_2 \circ b_2 \circ c_2 + \cdots + a_R \circ b_R \circ c_R.$$



More compact notation:

$$\text{vec}(\mathcal{X}) = [\![A, B, C]\!],$$

with

$$\begin{aligned} A &= [a_1, \ldots, a_R] \in \mathbb{R}^{n_1 \times R} \\ B &= [b_1, \ldots, b_R] \in \mathbb{R}^{n_2 \times R} \\ C &= [c_1, \ldots, c_R] \in \mathbb{R}^{n_3 \times R} \end{aligned}$$

EFY. What is the complexity of computing the inner product of two tensors in CP decomposition with R terms each?

CP decomposition and matricizations

Elementwise expression of CP decomposition:

$$\mathcal{X}_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr}, \quad i_\mu = 1, \dots, n_\mu$$

Shows that every matricization has rank at most R.

EFY. Develop a necessary and sufficient condition on the matricizations for a tensor $\mathcal{X}$ to have tensor rank 1.

For the 1-mode matricization:

$$X^{(1)} = AY,$$

with rth row of Y given by Kronecker prod of rth columns of B and C.

Definition. Let $V = [v_1, \dots, v_k]$ be $m \times k$ and $W = [w_1, \dots, w_k]$ be $n \times k$. Then the Kathri-Rao product of V and W is the $mn \times k$ matrix given by

$$V \odot W = \begin{bmatrix} v_1 \otimes w_1 & \cdots & v_k \otimes w_k \end{bmatrix}$$

EFY. Show the relations

$$(V \odot W) \odot Z = V \odot (W \odot Z), \quad (V \odot W)^T (V \odot W) = V^T V * W^T W,$$

where $*$ denotes the elementwise product and one assumes that all involved matrix sizes are suitably chosen.

CP decomposition and matricizations

We have

$$X^{(1)} = A(C \odot B)^T, \; X^{(2)} = B(C \odot A)^T, \; X^{(3)} = C(B \odot A)^T.$$

Corollary. Let R_μ be rank of $X^{(\mu)}$. Then

$$\max\{R_1, R_2, R_3\} \leq \operatorname{rank}(\mathcal{X}) \leq \min\{R_2 R_3, R_1 R_3, R_1 R_2\}.$$

Proof. Lower bound obvious, upper bound EFY.

- Except for very special situations (e.g., $n_3 = 2$) there are no tighter upper bounds known.
- A real $2 \times 2 \times 2$ tensor has rank at most 3 [Kruskal'1989], has ranks 2 or 3 with positive probability, and ranks 0 or 1 with zero probability.
- A real $3 \times 3 \times 3$ tensor has rank at most 5 [Kruskal'1989] and has rank 4 with probability 1 [ten Berge et al.'2004].
- Computing the tensor rank is an NP hard problem.

See [Kolda/Bader'2009] for more.

CP decomposition and uniqueness

The relation $X^{(1)} = A(C \odot B)^T$ can be written as

$$X^{(1)} = \left[A \cdot \mathrm{diag}(C(1,:)) \cdot B^T \quad \cdots \quad A \cdot \mathrm{diag}(C(n_3,:)) \cdot B^T \right].$$

$\rightsquigarrow$ simultaneous diagonalizations of all frontal slices of $\mathcal{X}$:

$$X(:,:,k) = A \cdot \mathrm{diag}(C(k,:)) \cdot B^T, \quad k = 1, \dots, n_3.$$

For $n_3 = 1$, significant degrees of freedom. For $n_3 > 1$, CP decomposition often unique (up to permutations and scalings).

Definition. The *k-rank* of a matrix A is the maximum value k such that any k columns are linearly independent.

Theorem (Kruskal'1977). The CP decomposition is unique (up to permutations and scalings) if

$$k\text{-rank}(A) + k\text{-rank}(B) + k\text{-rank}(C) \geq 2R + 2.$$

EFY. Construct an example which satisfies k-rank(A) + k-rank(B) + k-rank(C) = $2R + 1$ and uniqueness does not hold.

Properties of tensor vs. matrix rank

Given a real matrix A, its (matrix) rank is the same over $\mathbb{R}$ and $\mathbb{C}$.

This is not true, in general, for tensor rank.

Consider $2 \times 2 \times 2$ tensor $\mathcal{X}$ given by 1-mode matricization

$$X^{(1)} = \left[\begin{array}{cc|cc} 1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 \end{array}\right]$$

Considered as complex tensor $\mathcal{X} \in \mathbb{C}^{2\times2\times2}$ this tensor has rank 2 because one can write

$$\mathcal{X} = [\![A, B, C]\!],$$

with

$$A = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -\mathrm{i} & \mathrm{i} \end{bmatrix}, \quad B = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{bmatrix}$$

and it is easy to see that $\mathcal{X}$ does not have rank 1.

Properties of tensor vs. matrix rank

Given a real matrix A, its (matrix) rank is the same over $\mathbb{R}$ and $\mathbb{C}$.

This is not true, in general, for tensor rank.

Consider $2 \times 2 \times 2$ tensor $\mathcal{X}$ given by 1-mode matricization

$$X^{(1)} = \left[\begin{array}{cc|cc} 1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 \end{array} \right]$$

Considered as real tensor $\mathcal{X} \in \mathbb{R}^{2\times 2\times 2}$ this tensor has rank 3 because one can write

$$\mathcal{X} = [\![A, B, C]\!],$$

with

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix}$$

and the tensor can be shown to not have rank 2 or smaller, using techniques by ten Berge [1991].

Properties of tensor vs. matrix rank

The matrix rank is a lower semi-continuous function.

Meaning: A converging sequence of matrices of rank r always converges to a matrix of rank *at most* r.

Proof: Equivalently one has to show that for any $m \times n$ matrix A of rank r there is an open neighbourhood $\mathcal{U}$ such that all matrices $B \in \mathcal{U}$ have rank r or larger.

Use fact that a matrix A has rank at least r if and only if each of its $r \times r$ submatrices is invertible. For submatrix there is a neighbourhood of A such that this submatrix stays invertible in the neighbourhood. Take $\mathcal{U}$ to be the intersection of all these neighbourhoods.

Properties of tensor vs. matrix rank

The matrix rank is a lower semi-continuous function.

This is not true, in general, for tensor rank.

Consider $2 \times 2 \times 2$ tensor

$$\mathcal{X} = e_1 \circ e_1 \circ e_2 + e_1 \circ e_2 \circ e_1 + e_2 \circ e_1 \circ e_1.$$

with unit vectors e_1, e_2. Its rank can be shown to be 3.

Consider parametrized rank-2 tensor

$$\mathcal{X}_\alpha = \alpha\left(e_1 + \frac{1}{\alpha}e_2\right) \circ \left(e_1 + \frac{1}{\alpha}e_2\right) \circ \left(e_1 + \frac{1}{\alpha}e_2\right) - \alpha e_1 \circ e_1 \circ e_1.$$

This tensor has rank 2 but, as $X_\alpha \stackrel{\alpha\to\infty}{\to} X$, it is arbitrarily close to a tensor of rank 3.

A tensor $\mathcal{X}$ is called degenerate if it is arbitrarily close to a tensor of lower tensor rank.

See [Silva/Lim'2008] for more details on degenerate tensors.

Properties of tensor vs. matrix rank

The matrix rank is a lower semi-continuous function.

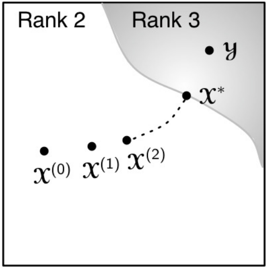
This is not true, in general, for tensor rank.

For tensors of order $d \geq 3$:

- tensor rank R is **not** upper semi-continuous $\rightsquigarrow$

  lack of closedness

- successive rank-1 approximations fail



Picture taken from [Kolda/Bader'2009].

To avoid degeneracies, can use smoothed version of tensor rank.

The border rank of a tensor $\mathcal{X}$ is defined as the smallest integer r such that for any $\varepsilon > 0$ there exists $\mathcal{E}$ with $\|\mathcal{E}\| = \varepsilon$ and $\mathcal{X} + \mathcal{E}$ has tensor rank r.

Block matrix multiplication and tensor rank

Consider 2×2 block matrix product

$$\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

Standard block matrix multiplication:

$$\begin{aligned} C_{11} &= A_{11}B_{11} + A_{12}B_{21} \\ C_{21} &= A_{21}B_{11} + A_{22}B_{21} \\ C_{12} &= A_{11}B_{12} + A_{12}B_{22} \\ C_{22} &= A_{21}B_{12} + A_{22}B_{22} \end{aligned}$$

Requires 8 block matrix multiplications!

Can be reduced via tensors + CP decomposition.

Block matrix multiplication and tensor rank

More general view of matrix multiplication algorithms:

- Map multiindex into single index:

$$\begin{pmatrix}(1,1)\\(2,1)\\(1,2)\\(2,2)\end{pmatrix} \to \begin{pmatrix}1\\2\\3\\4\end{pmatrix}$$

- Translate each operation into a rank-one tensor.
  Compute $A_{11}B_{11}$ and add to $(1,1)$ entry of C becomes:

$$e_{(1,1)} \circ e_{(1,1)} \circ e_{(1,1)} = \begin{pmatrix}1\\0\\0\\0\end{pmatrix} \circ \begin{pmatrix}1\\0\\0\\0\end{pmatrix} \circ \begin{pmatrix}1\\0\\0\\0\end{pmatrix}$$

  Compute $A_{21}B_{12}$ and add to $(2,2)$ entry of C becomes:

$$e_{(2,1)} \circ e_{(1,2)} \circ e_{(2,2)} = \begin{pmatrix}0\\1\\0\\0\end{pmatrix} \circ \begin{pmatrix}0\\0\\1\\0\end{pmatrix} \circ \begin{pmatrix}0\\0\\0\\1\end{pmatrix}$$

Block matrix multiplication and tensor rank

2×2 matrix multiplication tensor:

$$
\begin{aligned}
\mathcal{T}_2 &= e_{(1,1)} \circ e_{(1,1)} \circ e_{(1,1)} + e_{(1,2)} \circ e_{(2,1)} \circ e_{(1,1)} \\
&+ e_{(2,1)} \circ e_{(1,1)} \circ e_{(2,1)} + e_{(2,2)} \circ e_{(2,1)} \circ e_{(2,1)} \\
&+ \cdots
\end{aligned}
$$

Strassen discovered 1969 that this tensor has rank less than 8 by providing an explicit CP decomposition with 8 terms:

$$
\begin{aligned}
\mathcal{T}_2 &= (e_{(1,1)} + e_{(2,2)}) \circ (e_{(1,1)} + e_{(2,2)}) \circ (e_{(1,1)} + e_{(2,2)}) \\
&+ (e_{(2,1)} + e_{(2,2)}) \circ e_{(1,1)} \circ (e_{(2,1)} - e_{(2,2)}) \\
&+ e_{(1,1)} \circ (e_{(1,2)} - e_{(2,2)}) \circ (e_{(1,2)} + e_{(2,2)}) \\
&+ e_{(2,2)} \circ (e_{(2,1)} - e_{(1,1)}) \circ (e_{(1,1)} + e_{(2,1)}) \\
&+ (e_{(1,1)} + e_{(1,2)}) \circ e_{(2,2)} \circ (-e_{(1,1)} + e_{(1,2)}) \\
&+ (e_{(2,1)} - e_{(1,1)}) \circ (e_{(1,1)} + e_{(1,2)}) \circ e_{(2,2)} \\
&+ (e_{(1,2)} - e_{(2,2)}) \circ (e_{(2,1)} + e_{(2,2)}) \circ e_{(1,1)}
\end{aligned}
$$

Block matrix multiplication and tensor rank

First compute seven matrix products:

$$\begin{aligned}
M_1 &:= (A_{11} + A_{22})(B_{11} + B_{22}) \\
M_2 &:= (A_{21} + A_{22})B_{1,1} \\
M_3 &:= A_{11}(B_{12} - B_{22}) \\
M_4 &:= A_{22}(B_{21} - B_{11}) \\
M_5 &:= (A_{11} + A_{12})B_{22} \\
M_6 &:= (A_{21} - A_{11})(B_{11} + B_{12}) \\
M_7 &:= (A_{12} - A_{22})(B_{21} + B_{22})
\end{aligned}$$

Then compute entries of C:

$$\begin{aligned}
C_{11} &= M_1 + M_4 - M_5 + M_7 \\
C_{12} &= M_3 + M_5 \\
C_{21} &= M_2 + M_4 \\
C_{22} &= M_1 - M_2 + M_3 + M_6
\end{aligned}$$

Reduces complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^{\log_2 7}) = \mathcal{O}(n^{2.807\cdots})$

Block matrix multiplication and tensor rank

- $\mathcal{T}_2$ has tensor and border rank 7 [Strassen, Winograd, ...].
- $\mathcal{T}_3$ has tensor rank between 19 and 23; border rank between 15 and 21.
- [Bläser'1999]:$\mathcal{T}_n$ has tensor rank at least $\frac{5}{2}n^2 - 3n$.
- State-of-the-art constructions of fast matrix multiplication all tensor based.
- Record based on the so called laser method for tensors: $n^{2.3728639}$ [Le Gall'2014]. See also `https://simons.berkeley.edu/sites/default/files/docs/2438/slideslegall.pdf`.

Alternating least squares (ALS)

Fitting the CP decomposition to a tensor $\mathcal{X}$ requires the solution of the optimization problem

$$\min\left\{\|\mathcal{X} - [\![A, B, C]\!]\|^2 : A \in \mathbb{R}^{n_1\times R}, B \in \mathbb{R}^{n_2\times R}, C \in \mathbb{R}^{n_3\times R}\right\}.$$

EFY. Show that for $R = 1$, the solution of this optimization problem can be obtained from the solution of

$$\max\{u^T \circ_1 v^T \circ_2 w^T \circ_3 \mathcal{X} : u \in \mathbb{R}^{n_1}, v \in \mathbb{R}^{n_2}, w \in \mathbb{R}^{n_3}, u^T u = v^T v = w^T w = 1\}.$$

Note: This quantity is called spectral norm of tensor and for $d = 2$ it coincides with the spectral norm for matrices.

Difficulties:

- Scaling (and permutation) indeterminacy.
- target function convex in each factor but *not* jointly convex $\rightsquigarrow$ potentially many local minima.

Idea: Normalize all but one factor and exploit multilinearity of the format.

Alternating least squares (ALS)

Let B, C be fixed and with unit norm columns. Optimize for A only:

$$\min\left\{\|\mathcal{X} - [\![A, B, C]\!]\|^2 : A \in \mathbb{R}^{n_1 \times R}\right\}.$$

This is a standard linear least-squares problem (in an unusual form). Recall that 1-mode matricization of $[\![A, B, C]\!]$ given by $A(C \odot B)^T$:

$$\min\left\{\|X^{(1)} - A(C \odot B)^T\|_F^2 : A \in \mathbb{R}^{n_1 \times R}\right\}.$$

Solution via normal equations given by

$$\begin{aligned} A &= X^{(1)}(C \odot B)((C \odot B)^T(C \odot B))^{-1} \\ &= X^{(1)}(C \odot B)(C^TC * B^TB)^{-1} \end{aligned}$$

Remarks:

- This assumes that $C \odot B$ has full row rank. If not, replace $(C^TC * B^TB)^{-1}$ by pseudo-inverse.

Alternating least squares (ALS)

Input: $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, starting factors $B \in \mathbb{R}^{n_2 \times R}$,$C \in \mathbb{R}^{n_3 \times R}$.
Output: CP approximation $[\![A, B, C]\!]$.

```
Normalize columns of B.
while not converged do
    Normalize columns of C.
    Set A ← X^(1)(C ⊙ B)(C^T C ∗ B^T B)^(−1).
    Normalize columns of A.
    Set B ← X^(2)(C ⊙ A)(C^T C ∗ A^T A)^(−1).
    Normalize columns of B.
    Set C ← X^(3)(B ⊙ A)(B^T B ∗ A^T A)^(−1).
end while
```

EFY.

- Implement this procedure using, e.g., the tensor toolboxes in Matlab, Python, or Julia.
- Compare its convergence for a random tensor (of full generic rank) and a random tensor of exact rank R (generated by $[\![A, B, C]\!]$ for random matrices A, B, C).
- Apply this procedure to the Strassen tensor with $R = 7$. How close do you have to start at the exact CP decomposition in order to recover it?

EFY. Write down the algorithm for the special case $R = 1$, $d = 2$. What do you obtain?

The Tucker decomposition

Tucker decomposition

- Alternative rank concept for tensors motivated by

$$A = U \cdot \Sigma \cdot V^T, \qquad U \in \mathbb{R}^{n_1 \times r}, \quad V \in \mathbb{R}^{n_2 \times r}, \quad \Sigma \in \mathbb{R}^{r \times r}.$$

$\leadsto$ vectorization

$$\mathrm{vec}(X) = \big(V \otimes U\big) \cdot \mathrm{vec}(\Sigma).$$

Ignore diagonal structure of Σ and call it C.

Tucker decomposition of tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ defined via

$$\mathrm{vec}(\mathcal{X}) = \big(W \otimes V \otimes U\big) \cdot \mathrm{vec}(\mathcal{C})$$

with $U \in \mathbb{R}^{n_1 \times r_1}$, $V \in \mathbb{R}^{n_2 \times r_2}$, $W \in \mathbb{R}^{n_3 \times r_3}$,
and core tensor $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$.

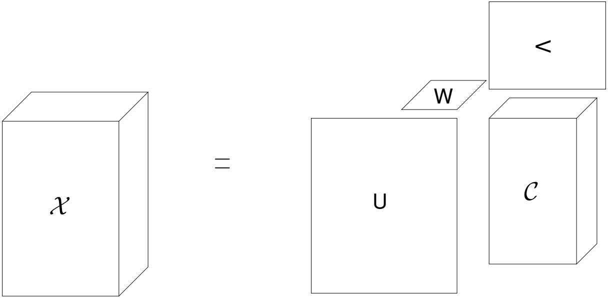
In terms of μ-mode matrix products:

$$\mathcal{X} = U \circ_1 V \circ_2 W \circ_3 \mathcal{C} =: (U, V, W) \circ \mathcal{C}.$$

Tucker decomposition

Illustration of Tucker decomposition

$$\mathcal{X} = (U, V, W) \circ \mathcal{C}$$



Tucker decomposition

Consider all three matricizations:

$$
\begin{aligned}
X^{(1)} &= U \cdot C^{(1)} \cdot (W \otimes V)^T, \\
X^{(2)} &= V \cdot C^{(2)} \cdot (W \otimes U)^T, \\
X^{(3)} &= W \cdot C^{(3)} \cdot (V \otimes U)^T.
\end{aligned}
$$

These are low rank decompositions $\rightsquigarrow$

$$
\operatorname{rank}\big(X^{(1)}\big) \le r_1, \quad \operatorname{rank}\big(X^{(2)}\big) \le r_2, \quad \operatorname{rank}\big(X^{(3)}\big) \le r_3.
$$

Multilinear rank of tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ defined by tuple

$$
(r_1, r_2, r_3), \quad \text{with} \quad r_i = \operatorname{rank}\big(X^{(i)}\big).
$$

EFY. Show that $r_1 \le r_2 r_3$, $r_2 \le r_1 r_3$, $r_3 \le r_1 r_2$. Are these inequalities tight?

Higher-order SVD (HOSVD)

Goal: Approximate given tensor $\mathcal{X}$ by Tucker decomposition with prescribed multilinear rank (r_1, r_2, r_3).

1. Calculate SVD of matricizations:
$$X^{(\mu)} = \widetilde{U}_\mu \widetilde{\Sigma}_\mu \widetilde{V}_\mu^T \quad \text{for } \mu = 1, 2, 3.$$
2. Truncate basis matrices:
$$U_\mu := \widetilde{U}_\mu(:, 1 : r_\mu) \quad \text{for } \mu = 1, 2, 3.$$
3. Form core tensor:
$$\mathcal{C} := U_1^T \circ_1 U_2^T \circ_2 U_3^T \circ_3 \mathcal{X}.$$

Truncated tensor produced by HOSVD [Lathauwer/De Moor/Vandewalle'2000]:

$$\widetilde{\mathcal{X}} := U_1 \circ_1 U_2 \circ_2 U_3 \circ_3 \mathcal{C}.$$

Remark:
Orthogonal projection $\widetilde{\mathcal{X}} := \big(\pi_1 \circ \pi_2 \circ \pi_3\big)\mathcal{X}$ with $\pi_\mu \mathcal{X} := U_\mu U_\mu^T \circ_\mu \mathcal{X}$.

Higher-order SVD (HOSVD)

Theorem. Tensor $\widetilde{\mathcal{X}}$ resulting from HOSVD satisfies quasi-optimality condition

$$\|\mathcal{X} - \widetilde{\mathcal{X}}\| \leq \sqrt{d}\|\mathcal{X} - \mathcal{X}_{\text{best}}\|,$$

where $\mathcal{X}_{\text{best}}$ is best approximation of $\mathcal{X}$ with multilinear ranks $(r_1, \dots, r_d)$.

Proof:

$$\begin{aligned}
\|\mathcal{X} - \widetilde{\mathcal{X}}\|^2 &= \|\mathcal{X} - (\pi_1 \circ \pi_2 \circ \pi_3)\mathcal{X}\|^2 \\
&= \|\mathcal{X} - \pi_1\mathcal{X}\|^2 + \|\pi_1\mathcal{X} - (\pi_1 \circ \pi_2)\mathcal{X}\|^2 + \cdots \\
&\quad \cdots + \|(\pi_1 \circ \pi_2)\mathcal{X} - (\pi_1 \circ \pi_2 \circ \pi_3)\mathcal{X}\|^2 \\
&\leq \|\mathcal{X} - \pi_1\mathcal{X}\|^2 + \|\mathcal{X} - \pi_2\mathcal{X}\|^2 + \|\mathcal{X} - \pi_3\mathcal{X}\|^2
\end{aligned}$$

Using

$$\|\mathcal{X} - \pi_\mu\mathcal{X}\| \leq \|\mathcal{X} - \mathcal{X}_{\text{best}}\| \quad \text{for } \mu = 1, 2, 3$$

leads to

$$\|\mathcal{X} - \widetilde{\mathcal{X}}\|^2 \leq 3 \cdot \|\mathcal{X} - \mathcal{X}_{\text{best}}\|^2.$$

Approximation error obtained from HOSVD

Another direct consequence of the proof:

Corollary. Let $\sigma_k^{(\mu)}$ denote the kth singular of $X^{(\mu)}$. Then the approximation $\widetilde{\mathcal{X}}$ obtained from the HOSVD satisfies

$$\|\mathcal{X} - \widetilde{\mathcal{X}}\|^2 \le \sum_{\mu=1}^{3} \sum_{k=r_\mu+1}^{n_\mu} (\sigma_k^{(\mu)})^2.$$

EFY. Provide a lower bound for $\|\mathcal{X} - \mathcal{X}_{\text{best}}\|$ in terms of the singular values of the matricizations of $\mathcal{X}$.

EFY. There is a more efficient variant, called the Sequentially Truncated HOSVD (STHOSVD), which proceeds as follows:

1. Calculate SVD $X^{(1)} = \widetilde{U}_1 \widetilde{\Sigma}_1 \widetilde{V}_1^T$. Truncate $U_1 := \widetilde{U}_1(:, 1:r_1)$ and update $\mathcal{X} \leftarrow U_1^T \circ_1 \mathcal{X}$.
2. Calculate SVD $X^{(2)} = \widetilde{U}_2 \widetilde{\Sigma}_2 \widetilde{V}_2^T$. Truncate $U_2 := \widetilde{U}_2(:, 1:r_2)$ and update $\mathcal{X} \leftarrow U_2^T \circ_2 \mathcal{X}$.
3. Calculate SVD $X^{(3)} = \widetilde{U}_3 \widetilde{\Sigma}_3 \widetilde{V}_3^T$. Truncate $U_3 := \widetilde{U}_3(:, 1:r_3)$ and update $\mathcal{X} \leftarrow U_3^T \circ_3 \mathcal{X}$.
4. Set $\mathcal{C} = \mathcal{X}$.

Any order of truncation is possible. It is usually most efficient to proceed from the largest to the smallest mode size.
Show that the bound of the theorem also holds for the STHOSVD.

- SVD can be replaced by any low-rank approximation technique discussed in this course. By triangular inequality, bound of Corollary still holds with an extra term accounting for the inexact SVD.

HOOI

Aim at finding best approximation:

$$\|\mathcal{X} - U_1 \circ_1 U_2 \circ_2 U_3 \circ_3 \mathcal{C}\| = \text{min!}$$

Given ONB U_1, U_2, U_3, the core tensor $\mathcal{C} := U_1^T \circ_1 U_2^T \circ_2 U_3^T \circ_3 \mathcal{X}$ is the optimal choice.

With the choice of core tensor above, we have

$$\|\mathcal{X} - U_1 \circ_1 U_2 \circ_2 U_3 \circ_3 \mathcal{C}\|^2 = \|\mathcal{X}\|^2 - \|U_1^T \circ_1 U_2^T \circ_2 U_3^T \circ_3 \mathcal{X}\|^2.$$

EFY. Prove this relation.

In turn, minimization problem equivalent to

$$\|U_1^T \circ_1 U_2^T \circ_2 U_3^T \circ_3 \mathcal{X}\| = \text{max!}$$

s.t. U_1, U_2, U_3 ONB.

HOOI

ALS for minimization problem: Let U_2, U_3 be fixed and maximize wrt $U_1 \rightsquigarrow$

$$\max_{U_1^T U_1 = I} \| U_1^T \circ_1 U_2^T \circ_2 U_3^T \circ_3 \mathcal{X} \| = \max_{U_1^T U_1 = I} \| U_1^T X^{(1)} (U_3 \otimes U_2) \|_F.$$

Solution obtained by setting U_1 to the r_1 dominant left singular vectors of $X^{(1)}(U_3 \otimes U_2)$.

Higher-order orthogonal iteration (HOOI):

1: **while** not converged **do**
2: Set U_1 to the r_1 dominant left singular vectors of $X^{(1)}(U_3 \otimes U_2)$.
3: Set U_2 to the r_2 dominant left singular vectors of $X^{(2)}(U_3 \otimes U_1)$.
4: Set U_3 to the r_3 dominant left singular vectors of $X^{(3)}(U_2 \otimes U_1)$.
5: **end while**

- Convergence often observed to be good but there are example, where it does not even converge to a stationary point.
- HOOI best initialized with factors from the (ST)HOSVD.
- Again, SVD can be replaced by other methods but little to no analysis exists that would guide such a choice.
- See [Kolda/Bader'09] for more details.

Tucker decomposition – Summary

For general tensors:

- multilinear rank r is upper semi-continuous $\rightsquigarrow$ closedness property.
- HOSVD – simple and robust algorithm to obtain quasi-optimal low-rank approximation.
- quasi-optimality good enough for most applications in scientific computing.
- robust black-box algorithms/software available (e.g., Tensor Toolbox).

Drawback:

Storage of core tensor $\sim r^d$
$\rightsquigarrow$ curse of dimensionality