Low rank tensor approximation

Krylov-type methods and perturbation analysis

Berkant Savas
Department of Mathematics
Linköping University

Workshop on Tensor Approximation in High Dimension
Hausdorff Institute for Mathematics, Universität Bonn
August 2, 2011

Joint work with Lars Eldén, Linköping University

A few multilinear algebraic definitions I

- Inner product

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i,j,k} a_{ijk} b_{ijk}$$

- Frobenius norm

$$\|\mathcal{A}\|_F = \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle}$$

- Contracted tensor products

$$\mathcal{C} = \langle \mathcal{A}, \mathcal{B} \rangle_{2,3} = \langle \mathcal{A}, \mathcal{B} \rangle_{-1} \qquad c_{\alpha\beta} = \sum_{j,k} a_{\alpha jk} b_{\beta jk}$$

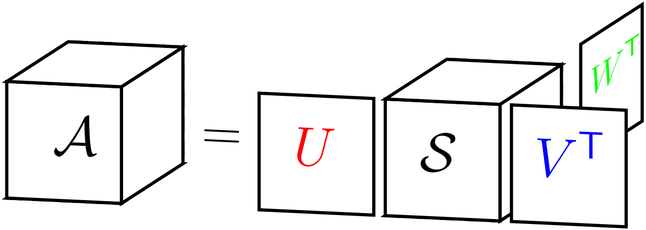
- Multilinear rank

$$\text{rank}\,(\mathcal{A}) = (r_1, r_2, r_3)$$

$$r_1 = \text{rank}(A^{(1)}) \qquad r_2 = \text{rank}(A^{(2)}) \qquad r_3 = \text{rank}(A^{(3)})$$

A few multilinear algebraic definitions II

- Multilinear tensor-matrix multiplication: U, V, W matrices



$$\mathcal{A} = (U, V, W) \cdot \mathcal{S} \qquad a_{ijk} = \sum_{\lambda,\mu,\nu} u_{i\lambda} v_{j\mu} w_{k\nu} s_{\lambda\mu\nu}$$

- convention

$$\left(U^{\mathsf{T}}, V^{\mathsf{T}}, W^{\mathsf{T}}\right) \cdot \mathcal{A} = \mathcal{A} \cdot (U, V, W)$$

- special cases

$$(U, V) \cdot A = UAV^{\mathsf{T}} \qquad \mathcal{A} \cdot (I, I, W) = \mathcal{A} \cdot (W)_3$$

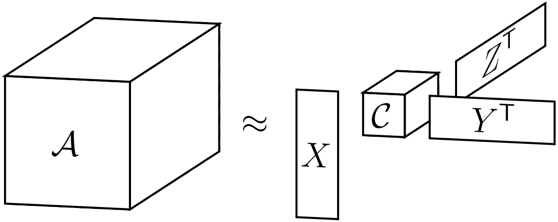
Multilinear low rank approximation

$$\min_{\mathcal{B}} \|\mathcal{A} - \mathcal{B}\| \qquad \text{rank}(\mathcal{B}) \leq (r_1, r_2, r_3)$$

With

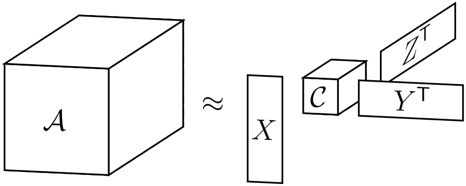
$$\mathcal{B} = (X, Y, Z) \cdot \mathcal{C}$$

$\mathcal{C}$ has dimensions $r_1 \times r_2 \times r_3$



Multilinear low rank approximation

$$\min \|\mathcal{A} - (X, Y, Z) \cdot \mathcal{C}\| \qquad \min \left\| A - XCY^{\mathsf{T}} \right\|$$



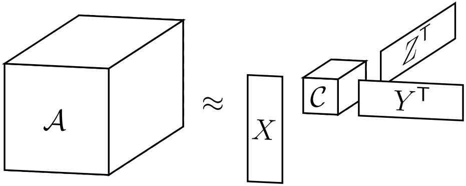
Problem overparameterized: $\mathcal{C}$ can be eliminated
Problem equivalent to

$$\max \|\mathcal{A} \cdot (X, Y, Z)\| \qquad \max \|X^{\mathsf{T}} A Y\|$$

$$X^{\mathsf{T}} X = I \qquad Y^{\mathsf{T}} Y = I \qquad Z^{\mathsf{T}} Z = I$$

Multilinear low rank approximation

$$\min \|\mathcal{A} - (X, Y, Z) \cdot \mathcal{C}\| \qquad \min \left\| A - XCY^{\mathsf{T}} \right\|$$



In addition to $\quad X^{\mathsf{T}}X = I \qquad Y^{\mathsf{T}}Y = I \qquad Z^{\mathsf{T}}Z = I$

$$\|\mathcal{A} \cdot (X, Y, Z)\| = \|\mathcal{A} \cdot (XU, YV, ZW)\|$$
$$\left\| X^{\mathsf{T}}AY \right\| = \left\| U^{\mathsf{T}}X^{\mathsf{T}}AYV \right\|$$

Grassmann manifold problem!

Computing the core $\mathcal{C}$ and the low rank approximation

Given orthonormal X, Y, Z

$$\min_{\mathcal{C}} \|\mathcal{A} - (X, Y, Z) \cdot \mathcal{C}\| \qquad \min_{C} \left\| A - XCY^{\mathsf{T}} \right\|$$

Solution

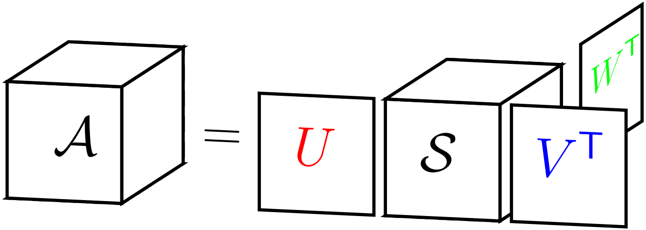
$$\mathcal{C} = \mathcal{A} \cdot (X, Y, Z) \qquad C = X^{\mathsf{T}} A Y$$

and the low rank approximation is

$$\hat{\mathcal{A}} = (X, Y, Z) \cdot \mathcal{C} = \left(XX^{\mathsf{T}}, YY^{\mathsf{T}}, ZZ^{\mathsf{T}} \right) \cdot \mathcal{A}$$

$$\hat{A} = XCY^{\mathsf{T}} = XX^{\mathsf{T}} A YY^{\mathsf{T}} = \left(XX^{\mathsf{T}}, YY^{\mathsf{T}} \right) \cdot A$$

HOSVD [De Lathauwer et al. 2000]



$$\mathcal{A} = (U, V, W) \cdot \mathcal{S} \qquad a_{ijk} = \sum_{\lambda,\mu,\nu} u_{i\lambda} v_{j\mu} w_{k\nu} s_{\lambda\mu\nu}$$

- U, V, W are orthogonal
- $\mathcal{S}$ is all-orthogonal

Methods for best low rank approximation

- Alternating minimization
  1. HOOI (Kroonenberg, De Lathauwer)
  2. "Trace maximization" (Regalia)
- Grassmann manifold approach
  1. Newton (Eldén, Savas)
  2. Trust region/Newton (Ishteva, De Lathauwer et al.)
  3. BFGS quasi-Newton (Savas, Lim)
  4. Limited memory BFGS (Savas, Lim)
  5. Symmetric tensors, without explicit representation (Morton)
- In this talk we will consider Krylov-type tensor computations.

OBS: This approach does not solve the best low rank problem

Krylov-type tensor computations

Krylov subspaces for matrices

Given $A \in \mathbb{R}^{m \times m}$ and starting vector $v \in \mathbb{R}^m$

$$\mathcal{K}_k(A, v) = \text{span}\{v, Av, A^2v, \ldots, A^{k-1}v\}$$

If $v_1 = v$

$$\mathcal{K}_k(A, v) = \text{span}\{v_1, v_2, v_3, \ldots, v_k\}$$

$$v_{i+1} = Av_i, \qquad i = 1, \ldots, k-1$$

Specifically useful for large and sparse problems

- Systems of linear equations
- Eigenvalues and eigenvectors
- Singular values and singular vectors
- Approximation of matrices and functions of matrices

A general square—The Arnoldi process

for $k = 1, 2, \ldots$ **do**

1 $h_k = U_k^{\mathsf{T}} A u_k$

2 $v = A u_k - U_k h_k$

3 $\beta_k = h_{k+1,k} = \|v\|_2$

4 $u_{k+1} = v / \beta_k$

5 $\hat{H}_k = \begin{pmatrix} \hat{H}_{k-1} & h_k \\ 0 & h_{k+1,k} \end{pmatrix}$

end for

- The Arnoldi decomposition:

$$A U_k = U_{k+1} \hat{H}_k, \qquad \hat{H}_k \sim (k+1) \times k \quad \text{Hessenberg}$$

- If A is symmetric $\Rightarrow$ Lanczos recurrence.

$A \in \mathbb{R}^{m\times n}$—Golub-Kahan bidiagonalization

- $\beta_1 v_1, \qquad u_0 = 0$

  for $j = 1, 2, \ldots, k$ **do**

  1 $\alpha_j u_j = A^{\mathsf{T}} v_j - \beta_j u_{j-1}$

  2 $\beta_{j+1} v_{j+1} = A u_j - \alpha_j v_j$

  end for

  α_j, β_j are chosen to normalize u_j, v_j
- $U_k = [u_1, \ldots, u_k]$ and $V_{k+1} = [v_1, \ldots, v_{k+1}]$ we have

$$AU_k = V_{k+1} B_{k+1}, \qquad V_k^{\mathsf{T}} V_k = I, \quad U_{k+1}^{\mathsf{T}} U_{k+1} = I$$

  and B_{k+1} is bidiagonal,
- U_k and V_k orthonormal basis for

$$\mathcal{K}_k(A^{\mathsf{T}}A, u) = \mathrm{span}\{u, (A^{\mathsf{T}}A)u, (A^{\mathsf{T}}A)^2 u, \ldots, (A^{\mathsf{T}}A)^{k-1} u\}$$
$$\mathcal{K}_k(AA^{\mathsf{T}}, v) = \mathrm{span}\{v, (AA^{\mathsf{T}})v, (AA^{\mathsf{T}})^2 v, \ldots, (AA^{\mathsf{T}})^{k-1} v\}$$

Matrix-vector and tensor-vector products

- $A \in \mathbb{R}^{m \times n}, \quad v \in \mathbb{R}^n$

$$Av = A \cdot (v)_2 \in \mathbb{R}^m$$

- $\mathcal{A} \in \mathbb{R}^{l \times m \times n}, \quad v \in \mathbb{R}^m, \quad w \in \mathbb{R}^n$

$$\mathcal{A} \cdot (w)_3 \in \mathbb{R}^{l \times m} \qquad [\mathcal{A} \cdot (w)_3]_{ij} = \sum_k a_{ijk} w_k$$

$$\mathcal{A} \cdot (v, w)_{2,3} \in \mathbb{R}^l \qquad [\mathcal{A} \cdot (v, w)_{2,3}]_i = \sum_{jk} a_{ijk} v_j w_k$$

- similarly

$$\mathcal{A} \cdot (u, w)_{1,3} \in \mathbb{R}^m \qquad \mathcal{A} \cdot (u, v)_{1,2} \in \mathbb{R}^n$$

Minimal Krylov method I

$\mathcal{A} \in \mathbb{R}^{l \times m \times n}$ and starting vectors with norm one
$u_1 \in \mathbb{R}^l$, $v_1 \in \mathbb{R}^m$, $w_1 \in \mathbb{R}^n$

$$\begin{aligned} u_{i+1} &= \mathcal{A} \cdot (v_i, w_i)_{2,3} & i &= 1, \ldots, k-1 \\ v_{i+1} &= \mathcal{A} \cdot (u_i, w_i)_{1,3} & i &= 1, \ldots, k-1 \\ w_{i+1} &= \mathcal{A} \cdot (u_i, v_i)_{1,2} & i &= 1, \ldots, k-1 \end{aligned}$$

Set

$$U_k = [u_1 \ u_2 \ \ldots \ u_k] \quad V_k = [v_1 \ v_2 \ \ldots \ v_k] \quad W_k = [w_1 \ w_2 \ \ldots \ w_k]$$

orthogonalize explicitly using Gram–Schmidt on each sequence

Minimal Krylov method II

$\mathcal{A} \in \mathbb{R}^{l \times m \times n}$ and starting vectors with norm one
$u_1 \in \mathbb{R}^l$, $v_1 \in \mathbb{R}^m$, $w_1 \in \mathbb{R}^n$

$$\begin{aligned} u_{i+1} &= \mathcal{A} \cdot (v_i, w_i)_{2,3} && i = 1, \dots, k-1 \\ v_{i+1} &= \mathcal{A} \cdot (u_{i+1}, w_i)_{1,3} && i = 1, \dots, k-1 \\ w_{i+1} &= \mathcal{A} \cdot (u_{i+1}, v_{i+1})_{1,2} && i = 1, \dots, k-1 \end{aligned}$$

Orthogonalize, set

$$U_k = [u_1 \ u_2 \ \dots \ u_k] \quad V_k = [v_1 \ v_2 \ \dots \ v_k] \quad W_k = [w_1 \ w_2 \ \dots \ w_k]$$

and approximate

$$\mathcal{A} \approx \left(U_k U_k^{\mathsf{T}}, V_k V_k^{\mathsf{T}}, W_k W_k^{\mathsf{T}} \right) \cdot \mathcal{A}$$

Krylov method on a low rank matrix

Let $A \in \mathbb{R}^{n \times n}$ with $\operatorname{rank}(A) = k$.

$$A = U_k \Sigma_k V_k^\mathsf{T}$$

then

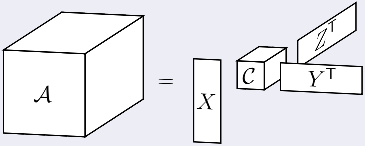
$$\mathcal{K}_k(A, u) = \mathcal{K}_{k+p}(A, u) \qquad p \geq 1$$

We only need to do k steps of Arnoldi.

Low rank tensor and minimal Krylov method

Theorem

Let $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$ *with* $\text{rank}(\mathcal{A}) = (p, q, r)$ *and assume* $p \leq q \leq r$. *Then*



$$\mathcal{A} = (X, Y, Z) \cdot \mathcal{C}$$

- *In p steps we get* U_p*, s.t.* $\text{span}(U_p) = \text{span}(X)$
- *In q steps we get* V_q*, s.t.* $\text{span}(V_q) = \text{span}(Y)$
- *In r steps we get* W_r*, s.t.* $\text{span}(W_r) = \text{span}(Z)$

using a "modified" minimal Krylov recursion.

Low rank tensors + noise and minimal Krylov method

Theorem

Let $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$ *with* $\operatorname{rank}(\mathcal{A}) = (p, q, r)$ *and add noise* $\rho\mathcal{E}$, *again assume* $p \leq q \leq r$. *Then*

$$\mathcal{A} = (X, Y, Z) \cdot \mathcal{C} + \rho\mathcal{E}$$

- *In* p *steps we get* U_p, *s.t.* $\operatorname{span}(U_p) \approx \operatorname{span}(X)$ *within level of noise*
- *In* q *steps we get* V_q, *s.t.* $\operatorname{span}(V_p) \approx \operatorname{span}(Y)$ *within level of noise*
- *In* r *steps we get* W_r, *s.t.* $\operatorname{span}(W_p) \approx \operatorname{span}(Z)$ *within level of noise*

using a "modified" minimal Krylov recursion.

Maximal Krylov Method

Generate all possible combinations at each step

$$
\begin{aligned}
\{u_1\} \times \{v_1\} &\longrightarrow w_1 \\
\{v_1\} \times \{w_1\} &\longrightarrow u_2 \\
\{u_1, u_2\} \times \{w_1\} &\longrightarrow \{v_2\ v_3\} \\
\{u_1, u_2\} \times \{v_1, v_2, v_3\} &\longrightarrow \{w_1, w_2, w_3, w_4, w_5, w_6\} \\
\{v_1, v_2, v_3\} \times \{w_1, w_2, \ldots, w_6\} &\longrightarrow \{u_2, u_3, \ldots, u_{19}\} \\
\{u_1, u_2, \ldots, u_{19}\} \times \{w_1, w_2, \ldots, w_6\} &\longrightarrow \{v_2, v_3, v_4, \ldots, v_{115}\}
\end{aligned}
$$

Krylov factorization for maximal recursion

Theorem (Tensor Krylov factorizations)

After a complete u-loop:

$$\mathcal{A} \cdot (V_k, W_l)_{2,3} = (U_j)_1 \cdot \mathcal{H}_{jkl}.$$

After a complete v-loop:

$$\mathcal{A} \cdot (U_j, W_l)_{1,3} = (V_m)_2 \cdot \mathcal{H}_{jml}.$$

After a complete w-loop:

$$\mathcal{A} \cdot (U_j, V_m)_{1,2} = (W_n)_3 \cdot \mathcal{H}_{jmn}.$$

Example of a maximal Krylov method

1. $\{u_1\} \times \{v_1\} \longrightarrow w_1 \Rightarrow \mathcal{A} \cdot (u_1, v_1)_{1,2} = (w_1)_3 \cdot \mathcal{H}_{111}$
2. $\{v_1\} \times \{w_1\} \longrightarrow u_2 \Rightarrow \mathcal{A} \cdot (v_1, w_1)_{2,3} = ([u_1 \ u_2])_1 \cdot \mathcal{H}_{211}$
3. $\{u_1, u_2\} \times \{w_1\} \longrightarrow \{v_2 \ v_3\}$

   $$\Rightarrow \mathcal{A} \cdot ([u_1 \ u_2], w_1)_{1,3} = ([v_1 \ v_2 \ v_3])_2 \cdot \mathcal{H}_{231}$$

4. $\{u_1, u_2\} \times \{v_1, v_2, v_3\} \longrightarrow \{w_1, w_2, w_3, w_4, w_5, w_6\}$

   $$\Rightarrow \mathcal{A} \cdot ([u_1 \ u_2], [v_1 \ v_2 \ v_3])_{1,2} = ([w_1 \ \cdots \ w_6])_2 \cdot \mathcal{H}_{236}$$

5. ....
6. the bad: dimension of subspaces explode

Krylov subspaces of contracted tensor products

Recall: for a matrix $A \in \mathbb{R}^{m\times n}$ we have

$$AA^\mathsf{T} = \langle A, A\rangle_{-1} \in \mathbb{R}^{m\times m}, \qquad A^\mathsf{T}A = \langle A, A\rangle_{-2} \in \mathbb{R}^{n\times n}$$

$\mathcal{A} \in \mathbb{R}^{m\times n\times l}$, $u \in \mathbb{R}^m, v \in \mathbb{R}^n, w \in \mathbb{R}^l$ and consider

$$\begin{aligned}
\langle \mathcal{A}, \mathcal{A}\rangle_{-1} &= A^{(1)}(A^{(1)})^\mathsf{T} \in \mathbb{R}^{m\times m} & \qquad \mathcal{K}_k(\langle \mathcal{A}, \mathcal{A}\rangle_{-1}, u)\\
\langle \mathcal{A}, \mathcal{A}\rangle_{-2} &= A^{(2)}(A^{(2)})^\mathsf{T} \in \mathbb{R}^{n\times n} & \qquad \mathcal{K}_k(\langle \mathcal{A}, \mathcal{A}\rangle_{-2}, v)\\
\langle \mathcal{A}, \mathcal{A}\rangle_{-3} &= A^{(3)}(A^{(3)})^\mathsf{T} \in \mathbb{R}^{l\times l} & \qquad \mathcal{K}_k(\langle \mathcal{A}, \mathcal{A}\rangle_{-3}, w)
\end{aligned}$$

Symmetric matrices! Apply the Lanczos recurrence.

All computations are implemented using

$$A \cdot (v, w)_{2,3} \qquad A \cdot (u, w)_{1,3} \qquad A \cdot (u, v)_{1,2}$$

Optimal subspaces give truncated HOSVD

Optimized tensor-Krylov approach

Let $U_i = [u_1 \ \cdots \ u_i]$, $V_i = [v_1 \ \cdots \ v_i]$, and $W_{i-1} = [w_1 \ \cdots \ w_{i-1}]$.
Find θ and η that give optimal $\widehat{w}$

$$\widehat{w} = \mathcal{A} \cdot (U_i\theta, V_i\eta)_{1,2} \, ,$$

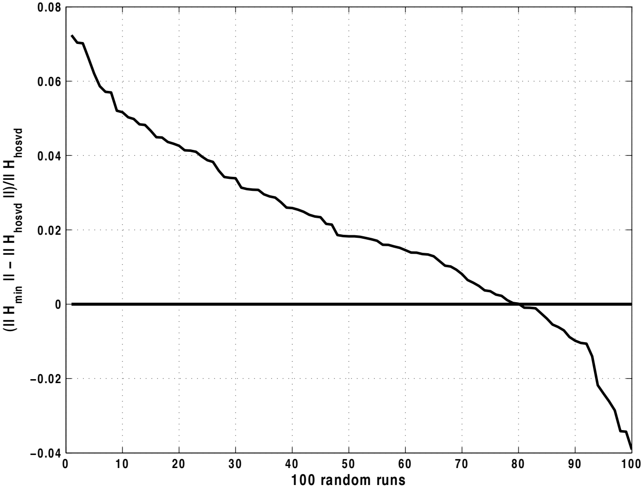
$$\max_{\theta,\eta} \|\widehat{w}\|, \quad \text{s.t.} \quad \widehat{w} \perp W_{i-1}, \|\theta\| = \|\eta\| = 1, \quad \theta, \eta \in \mathbb{R}^i.$$

Solution: best rank-$(1,1,1)$ approximation

$$(\theta, \eta, \omega) \cdot \mathcal{S}_{111} \approx A \cdot \left(U_i, V_i, I - W_{i-1} W_{i-1}^{\mathsf{T}} \right) .$$

[Goreinov, Oseledets and Savostyanov 2010] and [Savas and Eldén 2010]

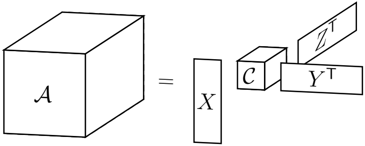
Experiments: Minimal Krylov vs truncated HOSVD



Difference between $\|\hat{\mathcal{A}}_{\text{min}}\|$ and $\|\hat{\mathcal{A}}_{\text{hosvd}}\|$, of a $50 \times 60 \times 40$ tensor $\mathcal{A}$. Rank of approximation is $(10, 10, 10)$.

HOSVD using the minimal tensor Krylov method

Let $\mathcal{A}$ have exact low rank



with HOSVD $\mathcal{A} = (U, V, W) \cdot \mathcal{S}$

Alternative I

1. SVD of $A^{(1)}$ gives U
2. SVD of $A^{(1)}$ gives V
3. SVD of $A^{(1)}$ gives W
4. Compute $S = A \cdot (U, V, W)$

Alternative II

1. Minimal Krylov method on $\mathcal{A}$ gives U_p, V_q, and W_r
2. Compute $\mathcal{C} = \mathcal{A} \cdot (U_p, V_q, W_r)$
3. Compite HOSVD of $\mathcal{C} = \left(\bar{U}, \bar{V}, \bar{W}\right) \cdot \mathcal{S}$
4. Change basis: $U = U_p \bar{U}$, $V = V_q \bar{V}$, $W = W_r \bar{W}$

Experiments: Applied on the Natflix data

Tensor of size

- User mode: 480189 — number of users
- Movie mode: 17770 — number of movies
- Time mode: 2243 — number of days

Time for computing a rank-$(100, 100, 100)$ approximation: 17 hours
Bottleneck:

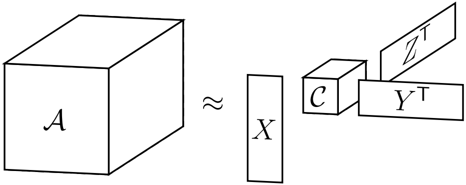
$$u = \mathcal{A} \cdot (v, w)_{2,3}$$

- Efficient storage schemes are needed
- Do not compute $\mathcal{A} \cdot (w)_3$ as it will be dense

Optimality conditions

Recall: Multilinear low rank approximation

$$\min \|\mathcal{A} - (X, Y, Z) \cdot \mathcal{C}\|$$



Problem overparameterized, nonlinear, and equivalent to

$$\max f(X, Y, Z), \quad \text{s.t.} \quad X^\mathsf{T} X = I \qquad Y^\mathsf{T} Y = I \qquad Z^\mathsf{T} Z = I$$

where $f(X, Y, Z) = \|\mathcal{A} \cdot (X, Y, Z)\|_F^2$

First order optimality conditions

- Objective function $f(X, Y, Z) = \|A \cdot (X, Y, Z)\|_F^2$
- At a stationary point (X, Y, Z) we have $\nabla f = 0$

$$\nabla f_x = \langle \mathcal{A} \cdot (X, Y, Z), \mathcal{A} \cdot (X_\perp, Y, Z) \rangle_{-1} = 0$$
$$\nabla f_y = \langle \mathcal{A} \cdot (X, Y, Z), \mathcal{A} \cdot (X, Y_\perp, Z) \rangle_{-2} = 0$$
$$\nabla f_z = \langle \mathcal{A} \cdot (X, Y, Z), \mathcal{A} \cdot (X, Y, Z_\perp) \rangle_{-3} = 0$$

- $[X \ X_\perp]$, $[Y \ Y_\perp]$, $[Z \ Z_\perp]$ are orthogonal
- For the matrices we have $f(X, Y) = \|X^{\mathsf{T}} A Y\|_F^2$

$$\langle X^{\mathsf{T}} A X, X_\perp^{\mathsf{T}} A Y \rangle_{-1} = X^{\mathsf{T}} A X (X_\perp^{\mathsf{T}} A Y)^{\mathsf{T}} = 0$$

- Let $[U_k \ U_\perp]$ and $[V_k \ V_\perp]$ contain left and right singular vectors, then

$$[U_k \ U_\perp]^{\mathsf{T}} A [V_k \ V_\perp] = \begin{bmatrix} U_k^{\mathsf{T}} A V_k & U_k^{\mathsf{T}} A V_\perp \\ U_\perp^{\mathsf{T}} A V_k & U_\perp^{\mathsf{T}} A V_\perp \end{bmatrix} = \begin{bmatrix} \Sigma_k & 0 \\ 0 & \Sigma_\perp \end{bmatrix}$$

WOOW..., what is that??

$$\langle \mathcal{A} \cdot (X, Y, Z), \mathcal{A} \cdot (X_\perp, Y, Z) \rangle_{-1} = 0$$

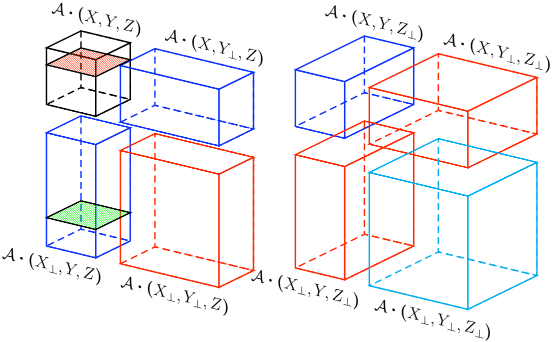


Figure visualizes the generalization of

$$[U_k \ U_\perp]^\mathsf{T} A [V_k \ V_\perp] = \begin{bmatrix} U_k^\mathsf{T} A V_k & U_k^\mathsf{T} A V_\perp \\ U_\perp^\mathsf{T} A V_k & U_\perp^\mathsf{T} A V_\perp \end{bmatrix} = \begin{bmatrix} \Sigma_k & 0 \\ 0 & \Sigma_\perp \end{bmatrix}$$

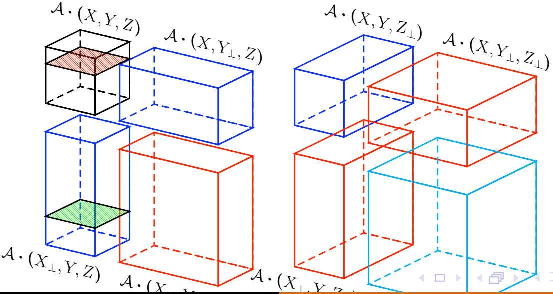
Second order optimality conditions ⇒ ordering

- (X, Y, Z) is a local max of $f(X,Y,Z) = \|\mathcal{A} \cdot (X, Y, Z)\|_F^2$ if the Hessian is negative definite.

Theorem

Let (X, Y, Z) be a local min and* $\mathcal{B} = \begin{bmatrix} \mathcal{A} \cdot (X, Y, Z) \\ \mathcal{A} \cdot (X_\perp, Y, Z) \end{bmatrix}$ *then*

$$\|\mathcal{B}(1,:,:)\| \geq \cdots \geq \|\mathcal{B}(r_1,:,:)\| > \|\mathcal{B}(r_1+1,:,:)\| \geq \cdots \geq \|\mathcal{B}(I,:,:)\|$$



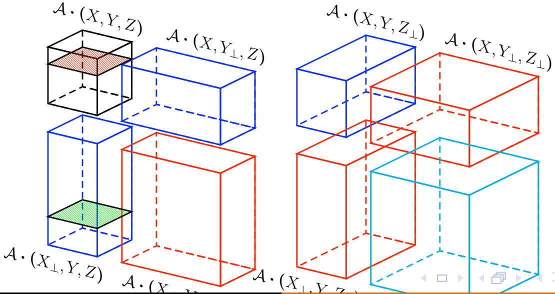
We can also get orthogonality

- (X, Y, Z) is a local max of $f(X, Y, Z) = \| \mathcal{A} \cdot (X, Y, Z) \|_F^2$ if the Hessian is negative definite.

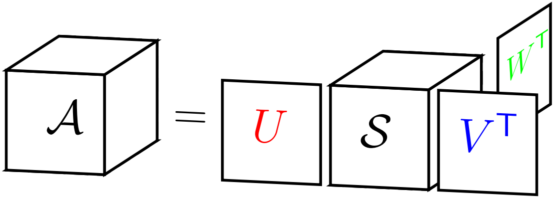
Theorem

Let (X, Y, Z) *be a local min** *and* $\mathcal{B} = \begin{bmatrix} \mathcal{A} \cdot (X, Y, Z) \\ \mathcal{A} \cdot (X_\perp, Y, Z) \end{bmatrix}$ *then*

$$\langle \mathcal{B}(i, :, :), \mathcal{B}(j, :, :) \rangle = 0, \quad i \neq j$$



Compare with the HOSVD ...



- U, V, W are orthogonal
- $\mathcal{S}$ is all-orthogonal:

$$\langle \mathcal{S}(:,:,i), \mathcal{S}(:,:,j) \rangle = 0 \quad i \neq j$$

$$\|\mathcal{S}(:,:,1)\| \geq \|\mathcal{S}(:,:,2)\| \geq \|\mathcal{S}(:,:,3)\| \geq \cdots$$

- Properties hold in all three modes simultaniously

Perturbation theory and concept of "gap" for tensors

Perturbation analysis: setup

- (X, Y, Z) representative of a stationary point for

$$\|\mathcal{A} - (X, Y, Z) \cdot \mathcal{C}\|_F^2$$

- Now perturbe $\mathcal{A}$ with a small $\mathcal{E}$

$$\tilde{\mathcal{A}} = \mathcal{A} + \mathcal{E}$$

- What are the first order perturbation $(\tilde{X}, \tilde{Y}, \tilde{Z})$ of the stationary point?

$$\tilde{X} = X + \delta X \qquad \tilde{Y} = Y + \delta Y \qquad \tilde{Z} = Z + \delta Z$$

- We want to bound δX, δY, and δZ in terms of some properties of $\mathcal{A}$

First order optimality condition

- The perturbed point $(\tilde{X}, \tilde{Y}, \tilde{Z})$ has to satisfy the first order optimality conditions, i.e. $\nabla f = (\nabla f_x, \nabla f_y, \nabla f_z) = 0$

$$\nabla f_x = \left\langle \widetilde{\mathcal{A}} \cdot \left(\widetilde{X}, \widetilde{Y}, \widetilde{Z}\right), \widetilde{\mathcal{A}} \cdot \left(\widetilde{X}_\perp, \widetilde{Y}, \widetilde{Z}\right) \right\rangle_{-1} = 0$$

$$\nabla f_y = \left\langle \widetilde{\mathcal{A}} \cdot \left(\widetilde{X}, \widetilde{Y}, \widetilde{Z}\right), \widetilde{\mathcal{A}} \cdot \left(\widetilde{X}, \widetilde{Y}_\perp, \widetilde{Z}\right) \right\rangle_{-2} = 0$$

$$\nabla f_z = \left\langle \widetilde{\mathcal{A}} \cdot \left(\widetilde{X}, \widetilde{Y}, \widetilde{Z}\right), \widetilde{\mathcal{A}} \cdot \left(\widetilde{X}, \widetilde{Y}, \widetilde{Z}_\perp\right) \right\rangle_{-3} = 0$$

First order optimality condition: "shake the gradient"...

- ..., and we get the Hessian! In operator form we have

$$\begin{bmatrix} H_{xx} & H_{xy} & H_{xz} \\ H_{yx} & H_{yy} & H_{yz} \\ H_{zx} & H_{zy} & H_{zz} \end{bmatrix} \begin{bmatrix} \Delta_x \\ \Delta_y \\ \Delta_z \end{bmatrix}$$

- The Hessian in the LHS and some "other things" on the RHS

$$\begin{aligned} \mathcal{H}_{xx}(\delta X) + \mathcal{H}_{xy}(\delta Y) + \mathcal{H}_{xz}(\delta Z) &= -(\langle \mathcal{F}_{\perp}^{x}, \mathcal{E}_0 \rangle_{-1} + \langle \mathcal{E}_x, \mathcal{F} \rangle_{-1}) \\ \mathcal{H}_{yx}(\delta X) + \mathcal{H}_{yy}(\delta Y) + \mathcal{H}_{yz}(\delta Z) &= -\left(\langle \mathcal{F}_{\perp}^{y}, \mathcal{E}_0 \rangle_{-2} + \langle \mathcal{E}_y, \mathcal{F} \rangle_{-2}\right) \\ \mathcal{H}_{zx}(\delta X) + \mathcal{H}_{zy}(\delta Y) + \mathcal{H}_{zz}(\delta Z) &= -(\langle \mathcal{F}_{\perp}^{z}, \mathcal{E}_0 \rangle_{-3} + \langle \mathcal{E}_z, \mathcal{F} \rangle_{-3}) \end{aligned}$$

- This looks messy....
- BUT: these equations will give us the "gap"!

Example 1: Matrix case

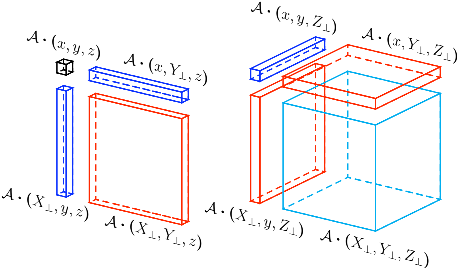
1. $$\begin{bmatrix} -(\Sigma_1^2 \otimes I) & \Sigma_1 \otimes \Sigma_2 \\ \Sigma_1 \otimes \Sigma_2 & -(\Sigma_1^2 \otimes I) \end{bmatrix} \begin{bmatrix} \text{vec}(\delta X) \\ \text{vec}(\delta Y) \end{bmatrix} = - \begin{bmatrix} (\Sigma_1 \otimes I)\,\text{vec}(E_x) \\ (\Sigma_1 \otimes I)\,\text{vec}(E_y) \end{bmatrix}$$
2. We can uncouple into 2×2 block systems. The worst is
$$\begin{bmatrix} -\sigma_r^2 & \sigma_r \sigma_{r+1} \\ \sigma_r \sigma_{r+1} & -\sigma_r^2 \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} = - \begin{bmatrix} \sigma_r e_x \\ \sigma_r e_y \end{bmatrix}$$
3. Solution:
$$\begin{bmatrix} \delta x \\ \delta y \end{bmatrix} = \frac{1}{(\sigma_r - \sigma_{r+1})(\sigma_r + \sigma_{r+1})} \begin{bmatrix} \sigma_r e_x + \sigma_{r+1} e_y \\ \sigma_{r+1} e_x + \sigma_r e_y \end{bmatrix}$$
4. Bounding gives:
$$|\delta x| \leq \frac{1}{\sigma_r - \sigma_{r+1}}(|e_x| + |e_y|) \qquad |\delta y| \leq \frac{1}{\sigma_r - \sigma_{r+1}}(|e_x| + |e_y|)$$
5. The quantity $(\sigma_r - \sigma_{r+1})$ is called the gap!

Example 2: Tensor case with rank-$(1, 1, 1)$ approximation

- Let (x, y, z) be a stationary point of $f(x, y, z) = \|\mathcal{A} \cdot (x, y, z)\|_F^2$
- With $c = \mathcal{A} \cdot (x, y, z)$ the perturbation equations become

$$\begin{bmatrix} -c^2 I & cF_\perp^{xy} & cF_\perp^{xz} \\ cF_\perp^{yx} & -c^2 I & cF_\perp^{yz} \\ cF_\perp^{zx} & cF_\perp^{zy} & -c^2 I \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix} = \begin{bmatrix} ce_x \\ ce_y \\ ce_z \end{bmatrix}$$

- Where $F_\perp^{xy} = \mathcal{A} \cdot (X_\perp, Y_\perp, z)$, ...



- Due to orthogonality blue vectors are zero!

Example 2: Tensor case with rank-$(1, 1, 1)$ approximation

- The perturbation satisfy

$$\begin{bmatrix} -c^2 I & cF_\perp^{xy} & cF_\perp^{xz} \\ cF_\perp^{yx} & -c^2 I & cF_\perp^{yz} \\ cF_\perp^{zx} & cF_\perp^{zy} & -c^2 I \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix} = \begin{bmatrix} ce_x \\ ce_y \\ ce_z \end{bmatrix}$$

- We get the bound

$$\left\| \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix} \right\| \leq \|G^{-1}\| \cdot \left\| \begin{bmatrix} e_x \\ e_y \\ e_z \end{bmatrix} \right\| \qquad G = -c \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} + \begin{bmatrix} 0 & F_\perp^{xy} & F_\perp^{xz} \\ F_\perp^{yx} & 0 & F_\perp^{yz} \\ F_\perp^{zx} & F_\perp^{zy} & 0 \end{bmatrix}$$

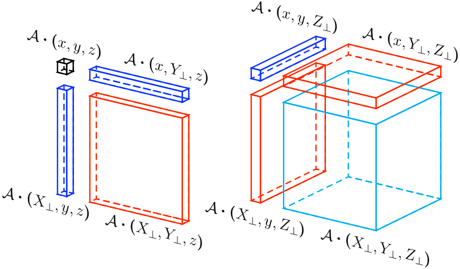
- What can we say about $\|G^{-1}\|$?

Example 2: Tensor case with rank-$(1,1,1)$ approximation

- We have that $\|G^{-1}\| \leq \frac{1}{|\lambda_{\min}(G)|}$

$$G = -c \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} + \begin{bmatrix} 0 & F_\perp^{xy} & F_\perp^{xz} \\ F_\perp^{yx} & 0 & F_\perp^{yz} \\ F_\perp^{zx} & F_\perp^{zy} & 0 \end{bmatrix} \equiv -cI + F$$

- The "gap" becomes $|\lambda_{\min}(G)| = c - \mu_{\max}$ where $\mu_{\max}$ is the largest eigenvalue of F

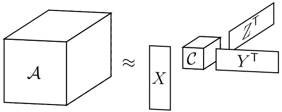


$$\begin{bmatrix} \sigma_1 & 0 \\ 0 & \Sigma_\perp \end{bmatrix}$$

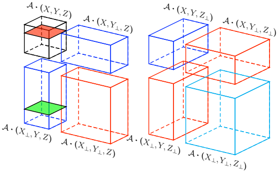
- Compare with the matrix case

Summary

1. We considered the multilinear low rank approximation of a tensor



2. Presented several Krylov-type procedures that generate low rank approximations
3. Interpreted first and second order optimality conditions: ordering and orthogonality



4. Generalized the "gap" from matrix sensitivity analysis to tensors

Thank you for your time!

berkant.savas@liu.se