Computing the best rank-(r_1, r_2, r_3) approximation of a tensor

Lars Eldén
Department of Mathematics, Linköping University
Joint work with Berkant Savas

Contents and Aim

- A very brief introduction to tensor algebra, HOSVD, best rank$-r_1, r_2, r_3$ approximation of a $3-$tensor, and an “alternating least squares algorithm”

  Tensor problems often involve heavy index-wrestling or matrization that obscure the structure. Is it possible to “algebraize” this tensor problem?

- Optimization on the Grassmann manifold

- The Newton equation for the best rank$-r_1, r_2, r_3$ optimization problem

A talk of questions and only a few answers

AIM: Develop the machinery that is needed(?) to answer the questions

Contravariant mode$-I$ multiplication of a tensor by a matrix[1]

$$\mathbf{R}^{n\times n\times n} \ni \mathcal{B} = (W)_{\{1\}}\mathcal{A}, \qquad \mathcal{B}(i,j,k) = \sum_{\nu=1}^{n} a_{\nu jk} w_{i\nu}.$$

All column vectors in the 3-tensor are multiplied by the matrix W.

When tensor-matrix multiplication is performed in all modes in the same expression, omit the subscripts:

$$(X,Y,Z)\mathcal{A}, \qquad (X_1,Y_1,Z_1)(X_2,Y_2,Z_2)\mathcal{A} = (X_1X_2, Y_1Y_2, Z_1Z_2)\mathcal{A},$$

Standard matrix multiplication of three matrices:

$$XAY^T = (X,Y)A \tag{1}$$

[1] (Lim's notation)

Covariant mode$-I$ multiplication of a tensor by a matrix

$$(\mathcal{A}(W)_{\{1\}})(i,j,k) = \sum_{\nu=1}^{n} a_{\nu jk} w_{\nu i}$$

and

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

Matrix case: $A(X,Y) = X^TAY$

Inner Product

Two tensors $\mathcal{A}$ and $\mathcal{B}$ of the same dimensions:

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i,j,k} a_{ijk} b_{ijk}, \qquad \|\mathcal{A}\| = \langle \mathcal{A}, \mathcal{A} \rangle^{1/2}.$$

Special case of contracted product of two tensors:[2]

The linear system $\sum_{j,k} k_{ijk} f_{jk} = g_i, \quad 1 \le i, j \le n,$

$$\langle \mathcal{K} \otimes F \rangle_{\{2,3;1,2\}} = g,$$

The matrix F and and the vector g are identified with tensors $\mathcal{F}$ and $\mathcal{G}$.

[2] Variant of the notation of Bader & Kolda [1].

Notation: outer and inner product

$\mathcal{A}$ and $\mathcal{B}$ are $3-$tensors of conforming dimensions

Outer product followed by a contraction: ($\mathcal{C}$ is a $4-$tensor)

$$\mathcal{C} = \langle \mathcal{A} \otimes \mathcal{B} \rangle_{\{1;1\}}, \qquad c_{ijkl} = \sum_{\mu} a_{\mu ij} b_{\mu kl}$$

Matrix multiplication: $XY = \langle X \otimes Y \rangle_{\{2;1\}}$

Inner product:

$$\langle \mathcal{A}, \mathcal{B} \rangle = \langle \mathcal{A} \otimes \mathcal{B} \rangle_{\{1:3,1:3\}} = \text{scalar}$$

Tensor SVD (HOSVD)[3]

An SVD-like of a $3-$tensor

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

where $X, Y, Z \in \mathbf{R}^{n \times n}$ are orthogonal matrices.

Core tensor $\mathcal{S}$ has the same dimensions as $\mathcal{A}$.

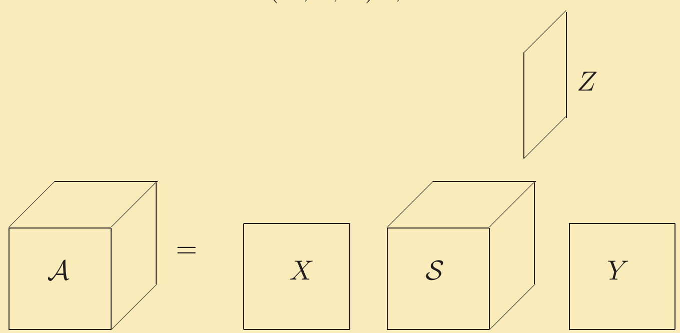
All-orthogonality: slices along any mode are orthogonal. Let $\nu \neq \mu$; then

$$\begin{aligned} & \langle \mathcal{S}(\nu,:,:), \mathcal{S}(\mu,:,:) \rangle = \langle \mathcal{S}(:,\nu,:), \mathcal{S}(:,\mu,:) \rangle \\ = \; & \langle \mathcal{S}(:,:,\nu), \mathcal{S}(:,:,\mu) \rangle = 0. \end{aligned}$$

[3] De Lathauwer et al. [4]. Related to the Tucker-3 decomposition in psychometrics and chemometrics.

HOSVD

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



Singular Values

Mode-1 singular values

$$\sigma_i^{(1)} = \|\mathcal{S}(i,:,:)\|, \qquad i = 1, \dots, n.$$

The singular values are ordered,

$$\sigma_1^{(\nu)} \geq \sigma_2^{(\nu)} \geq \cdots \geq \sigma_n^{(\nu)} \geq 0, \qquad \nu = 1, 2, 3.$$

“Energy”

The singular values are measures of the “energy” of the tensor

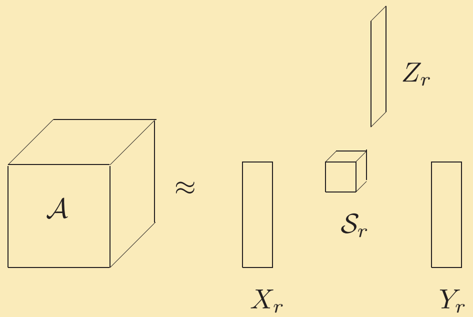
Proposition 1.

$$\|\mathcal{A}\|^2 = \|\mathcal{S}\|^2 = \sum_{i=1}^{n} \left(\sigma_i^{(1)}\right)^2 = \sum_{i=1}^{n} \left(\sigma_i^{(2)}\right)^2 = \sum_{i=1}^{n} \left(\sigma_i^{(3)}\right)^2.$$

The “energy” (mass) is concentrated at the $(1,1,1)$ corner of the tensor

We can truncate the HOSVD (in analogy to TSVD)

Truncated HOSVD



Does not give the best rank$-(r_1, r_2, r_3)$ approximation!

Questions

- How close is the truncated HOSVD to the best rank$-(r_1, r_2, r_3)$ approximation?

  Experimentally: often very close

- What mathematical structure determines the closeness?

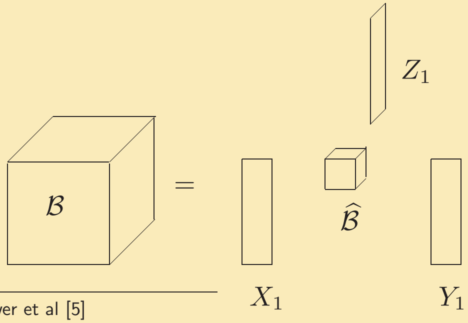
- Given a tensor one can define linear operators. Are there any tensors/linear operators with SVD=HOSVD?

  Answer: Yes, if the tensor is product-separable (Kronecker structure)

Best rank$-(r_1, r_2, r_3)$ approximation[4]

$$\min_{\mathcal{B} \in S} \|\mathcal{A} - \mathcal{B}\|_F, \qquad S = \{\mathcal{B} \mid \operatorname{rank}(\mathcal{B}) \leq (r_1, r_2, r_3)\}. \tag{2}$$

The rank constraint is to be understood: $\mathcal{B} = (X_1, Y_1, Z_1)\widehat{\mathcal{B}}$



[4] De Lathauwer et al [5]

Define three orthogonal matrices, arbitrary for now:

$$X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}, \qquad Y = \begin{pmatrix} Y_1 & Y_2 \end{pmatrix}, \qquad Z = \begin{pmatrix} Z_1 & Z_2 \end{pmatrix}.$$

In transformed coordinates, i.e., with $\widehat{\mathcal{A}} = (X^T, Y^T, Z^T)\mathcal{A}$:

$$\begin{aligned}
\min_{\mathcal{B}} \|\widehat{\mathcal{A}} \; - \; \widehat{\mathcal{B}}\|_F^2 = & \\
&= \min \sum_{i=1}^{r_1}\sum_{j=1}^{r_2}\sum_{k=1}^{r_3}(\hat{a}_{ijk} - \hat{b}_{ijk})^2 + \sum_{i=r_1+1}^{n}\sum_{j=r_2+1}^{n}\sum_{k=r_3+1}^{n}(\hat{a}_{ijk} - \hat{b}_{ijk})^2 \\
&= \min \sum_{i=1}^{r_1}\sum_{j=1}^{r_2}\sum_{k=1}^{r_3}(\hat{a}_{ijk} - \hat{b}_{ijk})^2 + \sum_{i=r_1+1}^{n}\sum_{j=r_2+1}^{n}\sum_{k=r_3+1}^{n}\hat{a}_{ijk}^2
\end{aligned}$$

Optimization Problem

Determine X_1, Y_1, and Z_1 so that

$$\|(X_1^T, Y_1^T, Z_1^T)\mathcal{A}\|_F = \|\mathcal{A}(X_1, Y_1, Z_1)\|_F$$

is maximized.

Drop subscripts, and remember that the matrices are rectangular with orthonormal columns.

Matrix case:

$$\begin{aligned}
\|A(X,Y)\|_F^2 &= \|X^T A Y\|_F^2 = \operatorname{tr}(Y^T A^T X X^T A Y) \\
&= \operatorname{tr}(W^T Y^T A^T X V V^T X^T A Y W)
\end{aligned}$$

where V and W are orthogonal.

The optimization problem

$$\max \|\mathcal{A}(X, Y, Z)\|_F, \qquad X^T X = I, \quad Y^T Y = I, \quad Z^T Z = I,$$

is not completely well-defined: Indeterminate because we may exchange

$$X \longrightarrow XV, \qquad Y \longrightarrow YW, \qquad Z \longrightarrow ZU$$

where V, W, and U are orthogonal

We are looking for subspaces rather than orthogonal matrices!

Standard method: “Alternating least squares”

Iterate until convergence

1. Fix Y, Z, solve $\max_{X^T X = I} \|\mathcal{A}(I, Y, Z)(X)_{\{1\}}\|_F$
2. Fix X, Z, solve $\max_{Y^T Y = I} \|\mathcal{A}(X, I, Z)(Y)_{\{2\}}\|_F$
3. Fix X, Y, solve $\max_{Z^T Z = I} \|\mathcal{A}(X, Y, I)(Z)_{\{3\}}\|_F$

end iterations

$\mathcal{A}(I, Y, Z)$ is a linear operator acting on X in mode 1, etc.

Solution of each subproblem given by SVD.

“Power method (alternating subspace iteration)”

Convergence may be very slow.

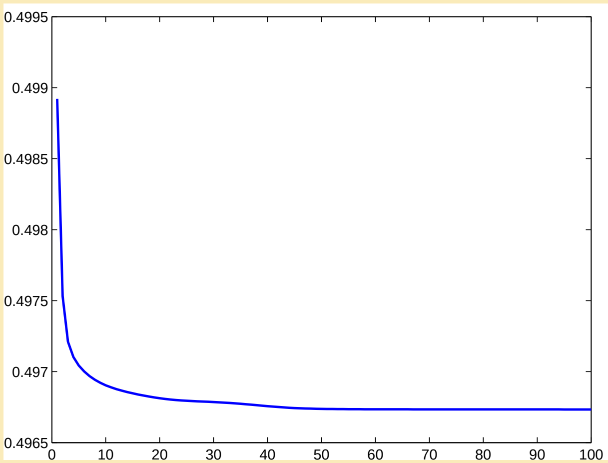
Example

```
rk=[4 4 4];
a=rand(50,50,50);
maxit=100;
a=tensor(a);
[Lam,U,err]=hopm(a,rk,maxit); % Alternating subspace iteration
                              % initialized by HOSVD
plot(err)                     % Approximation error
err(end-1)-err(end)
```

Difference in approximation error after 100 iterations:

5.3517e-08

Approximation error



Questions

- What determines the rate of convergence of the alternating subspace iteration?
- How accurately can the subspaces be computed?
- Eigenspace sensitivity depends on separation of eigenvalues. What are the corresponding quantities here?

Grassmann Manifold[5]

We want to determine subspaces rather than matrices

The Grassmann manifold of dimension r is a set of equivalence classes:

$$\mathbb{G}(n, r) = [Y], \qquad Y \in \mathbb{R}^{n \times r}, \qquad Y^T Y = I,$$

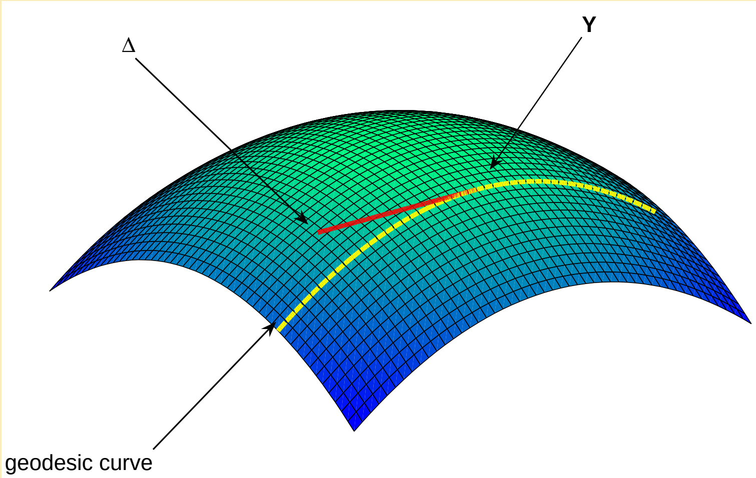
under the equivalence

$$[Y_1] = [Y_2] \quad \text{iff} \quad Y_1 = Y_2 V,$$

for some orthogonal matrix $V \in \mathbb{R}^{r \times r}$.

[5] See Edelman et al. [2].

Optimization on the Grassmann manifold



Tangent space

Newton's method operates in a vector space $\mathbb{T}_Y$: the tangent space at Y

$$\mathbb{R}^{n \times r} \ni \Delta \in \mathbb{T}_Y \quad \Longleftrightarrow \quad \Delta^T Y = 0.$$

Projection onto $\mathbb{T}_Y$:

$$\Pi = I - YY^T$$

Gradient of a function $F(Y)$

The gradient ∇F is a vector in $\mathbb{T}_Y$ such that

$$\langle \Delta, \nabla F \rangle_{\mathbb{T}_Y} = \langle \Delta, F_Y \rangle_{\mathbb{R}^{n \times r}}, \qquad \forall \Delta \in \mathbb{T}_Y$$

It follows that

$$\nabla F = \Pi F_Y, \qquad \Pi = I - YY^T$$

where F_Y is the usual Euclidean derivative

Hessian of a function $F(Y)$

The Hessian H is a vector in $\mathbb{T}_Y$:

$$H = \Pi F_{YY}(\Delta) - \Delta Y^T F_Y, \qquad \Delta \in \mathbb{T}_Y$$

F_{YY} is the usual Euclidean derivative

Grassmann Geodesic Curves

Let $\Delta \in \mathbb{T}_Y$ with thin SVD $\Delta = U\Sigma V^T$.

The geodesic curve starting from Y in the direction Δ is given by

$$Y(t) = YV\cos(\Sigma t)V^T + U\sin(\Sigma t)V^T$$

By definition:

$$\left.\frac{dY(t)}{dt}\right|_{t=0} = -YV\sin(\Sigma t)V^T + U\cos(\Sigma t)V^T\big|_{t=0} = \Delta$$

Newton-Grassmann method for $\max F(Y)$

Starting approximation Y

Iterate until convergence

1. Find the vector $\Delta \in \mathbb{T}_Y$ such that

$$H(\Delta) = -\nabla F,$$

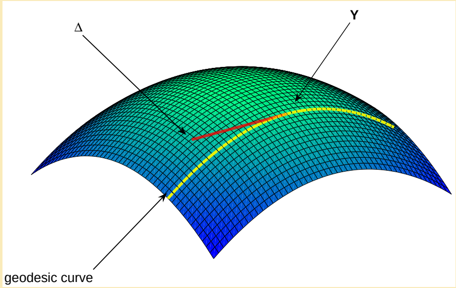
   Thin SVD: $\Delta = U\Sigma V^T$

2. Take a step along the geodesic curve of direction Δ:

$$Y := Y(1) = YV\cos(\Sigma)V^T + U\sin(\Sigma)V^T$$

end iterations

Newton's method on the Grassmann manifold



Find the direction $\Delta \in \mathbb{T}_Y$ and take a geodesic step (or `Y:=qr(Y+Δ)`)

$\boxed{\text{Well-defined optimization (correct \# d.o.f.) and quadratic convergence}}$

Newton's method

$$F(t) \approx F(0) + t\,\frac{dF}{dt}\bigg|_{t=0} + \frac{t^2}{2}\,\frac{d^2F}{dt^2}\bigg|_{t=0}$$

With $\mathbb{T}_Y$ inner product $\langle \cdot, \cdot \rangle$:

$$F(Y(1)) \approx F(Y) + \langle \Delta, \nabla F \rangle + \frac{1}{2}\langle \Delta, H(\Delta) \rangle$$

we get a Newton equation on $\mathbb{T}_Y$:

$$H(\Delta) = -\nabla F$$

Best rank$-(r, r, r)$ approximation.

For simplicity: $r_1 = r_2 = r_3 = r$. Put $\mathbb{G} := \mathbb{G}(n, r)$ and $\mathbb{G}^3 = \mathbb{G} \times \mathbb{G} \times \mathbb{G}$

$$\max_{(X,Y,Z)\in\mathbb{G}^3} F(X, Y, Z) = \max_{(X,Y,Z)\in\mathbb{G}^3} \frac{1}{2}\langle \mathcal{A}(X, Y, Z), \mathcal{A}(X, Y, Z) \rangle$$

where

$$\mathcal{A}(X, Y, Z)(i, j, k) = \sum_{\lambda,\mu,\nu} a_{\lambda\mu\nu} x_{i\lambda} y_{j\mu} z_{k\nu}$$

Can we avoid index-wrestling? Yes, almost all of it.

Differentiate along three tangent directions $\Delta_x, \Delta_y, \Delta_z$

Since

$$\frac{dX}{dt} = \Delta_x, \qquad \frac{dY}{dt} = \Delta_y, \qquad \frac{dZ}{dt} = \Delta_z,$$

and

$$\mathcal{A}(X, Y, Z)(i, j, k) = \sum_{\lambda,\mu,\nu} a_{\lambda\mu\nu} x_{i\lambda} y_{j\mu} z_{k\nu},$$

every x_{ij} etc. will be replaced by $(\Delta_x)_{ij}$ etc. in the differentiation.

Therefore

$$\frac{d\mathcal{A}(X, Y, Z)}{dt} = \mathcal{A}(\Delta_x, Y, Z) + \mathcal{A}(X, \Delta_y, Z) + \mathcal{A}(X, Y, \Delta_z)$$

Derivatives

$$\begin{aligned}\frac{dF}{dt} &= \langle \mathcal{A}(\Delta_x, Y, Z), \mathcal{A}(X, Y, Z) \rangle + \langle \mathcal{A}(X, \Delta_y, Z), \mathcal{A}(X, Y, Z) \rangle \\ &+ \langle \mathcal{A}(X, Y, \Delta_z), \mathcal{A}(X, Y, Z) \rangle\end{aligned}$$

$$\begin{aligned}\frac{d^2F}{dt^2} &= \langle \mathcal{A}(\Delta_x, Y, Z), \mathcal{A}(\Delta_x, Y, Z) \rangle - \langle \mathcal{A}(\Delta_x \Delta_x^T X, Y, Z), \mathcal{A}(X, Y, Z) \rangle \\ &+ \langle \mathcal{A}(\Delta_x, \Delta_y, Z), \mathcal{A}(X, Y, Z) \rangle + \langle \mathcal{A}(X, \Delta_y, Z), \mathcal{A}(\Delta_x, Y, Z) \rangle \\ &+ \langle \mathcal{A}(\Delta_x, Y, \Delta_z), \mathcal{A}(X, Y, Z) \rangle + \langle \mathcal{A}(X, Y, \Delta_z), \mathcal{A}(\Delta_x, Y, Z) \rangle \\ &+ Y- \text{ and } Z-\text{derivatives}\end{aligned}$$

Identify gradient and Hessian: $\langle \Delta, \nabla F \rangle + \frac{1}{2}\langle \Delta, H(\Delta) \rangle$

Tensor-matrix-products

Matrization and vectorization obcure the structure.

Basic rule: Matricize and vectorize as late as possible!

Lemma 1. *Let $\mathcal{B}$ and $\mathcal{C}$ be $3-$tensors of conforming dimensions.*

$$\begin{aligned}\langle\, \mathcal{B}(X_1)_{\{1\}}, \mathcal{C}(X_2)_{\{1\}}\,\rangle &= \langle\, X_1, \langle\, \mathcal{B}\otimes\mathcal{C}(X_2)_{\{1\}}\,\rangle_{\{2:3\}}\,\rangle \\ &= \langle\, X_1, \langle\, \mathcal{B}\otimes\mathcal{C}\,\rangle_{\{2:3\}}(X_2)_{\{1\}}\,\rangle\end{aligned}$$

Matrix factors can be "pulled out" of the inner product.

Lemma 2.

$$\langle\, \mathcal{B}(Y)_{\{2\}}\otimes\mathcal{C}\,\rangle_{\{2:3\}} = \langle\, \mathcal{D}\otimes Y\,\rangle_{\{2:4;1:2\}},$$

where the $4-$tensor $\mathcal{D}$ is defined

$$\mathcal{D} = \langle\, \mathcal{B}\otimes\mathcal{C}\,\rangle_{\{3\}}$$

$\mathcal{D}$ is a linear operator: matrix $\longrightarrow$ matrix

Grassmann gradient

$$\nabla F =$$

$$\begin{pmatrix} \langle\, \mathcal{A}(I,Y,Z)\otimes\mathcal{A}(I,Y,Z)\,\rangle_{\{2:3\}}(X)_{\{1\}} - (X)_{\{1\}}\langle\, \mathcal{A}(X,Y,Z)\otimes\mathcal{A}(X,Y,Z)\,\rangle_{\{2:3\}} \\ \langle\, \mathcal{A}(X,I,Z)\otimes\mathcal{A}(X,I,Z)\,\rangle_{\{1,3\}}(Y)_{\{2\}} - (Y)_{\{2\}}\langle\, \mathcal{A}(X,Y,Z)\otimes\mathcal{A}(X,Y,Z)\,\rangle_{\{1,3\}} \\ \langle\, \mathcal{A}(X,Y,I)\otimes\mathcal{A}(X,Y,I)\,\rangle_{\{1:2\}}(Z)_{\{3\}} - (Z)_{\{3\}}\langle\, \mathcal{A}(X,Y,Z)\otimes\mathcal{A}(X,Y,Z)\,\rangle_{\{1:2\}} \end{pmatrix}$$

The matrix elements are all inner products between slices in each mode

Cf. the subspace equation for the matrix eigenvalue problem:

$$AX = XL$$

Grassmann Hessian

$$H(\Delta) = \begin{pmatrix} (\Pi_x)_{\{1\}} & 0 & 0 \\ 0 & (\Pi_y)_{\{2\}} & 0 \\ 0 & 0 & (\Pi_z)_{\{3\}} \end{pmatrix} \begin{pmatrix} H_{xx}(\Delta_x) & H_{xy}(\Delta_y) & H_{xz}(\Delta_z) \\ H_{yx}(\Delta_x) & H_{yy}(\Delta_y) & H_{yz}(\Delta_z) \\ H_{zx}(\Delta_x) & H_{zy}(\Delta_y) & H_{zz}(\Delta_z) \end{pmatrix}$$

where the diagonal blocks are Sylvester operators:

$$\begin{aligned} H_{xx}(\Delta_x) &= (\langle\, \mathcal{A}(I,Y,Z))\otimes\mathcal{A}(I,Y,Z)\,\rangle_{\{2:3\}}(\Delta_x)_{\{1\}} \\ &\quad -(\Delta_x)_{\{1\}}(\langle\, \mathcal{A}(X,Y,Z))\otimes\mathcal{A}(X,Y,Z)\,\rangle_{\{2:3\}} \end{aligned}$$

and the off-diagonal blocks are tensor-matrix linear operators

Off-diagonal block: $\frac{\partial^2 F}{\partial X \partial Y}$

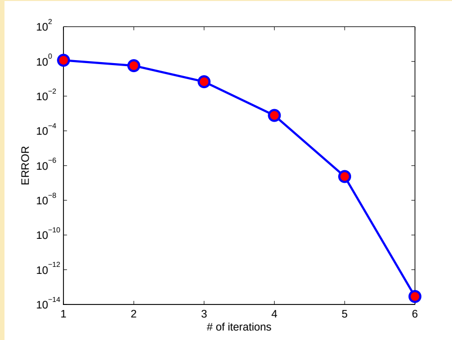
$$
\begin{aligned}
\langle \mathcal{A}(\Delta_x, \Delta_y, Z) \otimes \mathcal{A}(X, Y, Z) \rangle &= \langle \Delta_x, \langle \mathcal{A}(I, \Delta_y, Z) \otimes \mathcal{A}(X, Y, Z) \rangle_{\{2:3\}} \rangle \\
&= \langle \Delta_x, \langle \mathcal{H} \otimes \Delta_y \rangle_{\{2,4;1:2\}} \rangle,
\end{aligned}
$$

where

$$\mathcal{H} = \langle \mathcal{A}(I, I, Z) \otimes \mathcal{A}(X, Y, Z) \rangle_{\{3\}}$$

is a $4-$tensor.

Very preliminary numerical experiments



Small problem

But: the code is in a very elarly stage of development

Ongoing work

- Implementation of the tensor Newton-Grassmann method using object-oriented MATLAB:
  - tensor toolbox (Bader & Kolda)
  - homogeneous manifold optimization toolbox (home-made)
- Investigation of the theoretical properties of the best rank$-(r_1, r_2, r_3)$ approximation

References

[1] B. Bader and T. Kolda. Matlab tensor classes for fast algorithm prototyping. Technical Report SAND2004-5187, Sandia National Laboratories, Oct. 2004.

[2] A. Edelman, T. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20:303–353, 1999.

[3] L. De Lathauwer. First-order perturbation analysis of the best rank-(R_1, R_2, R_3) approximation of a tensor. *J. Chemometrics*, 18:2–11, 2004.

[4] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21:1253–1278, 2000.

[5] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-$(R_1, R_2, \ldots, R_N)$ approximation of higher-order tensor. *SIAM J. Matrix Anal. Appl.*, 21:1324–1342, 2000.

[6] T. Zhang and G.H. Golub. Rank-one approximation to higher order tensors. *SIAM J. Matrix Anal. Appl.*, 23:534–550, 2002.