Low Rank Orthogonal Approximation of Tensors

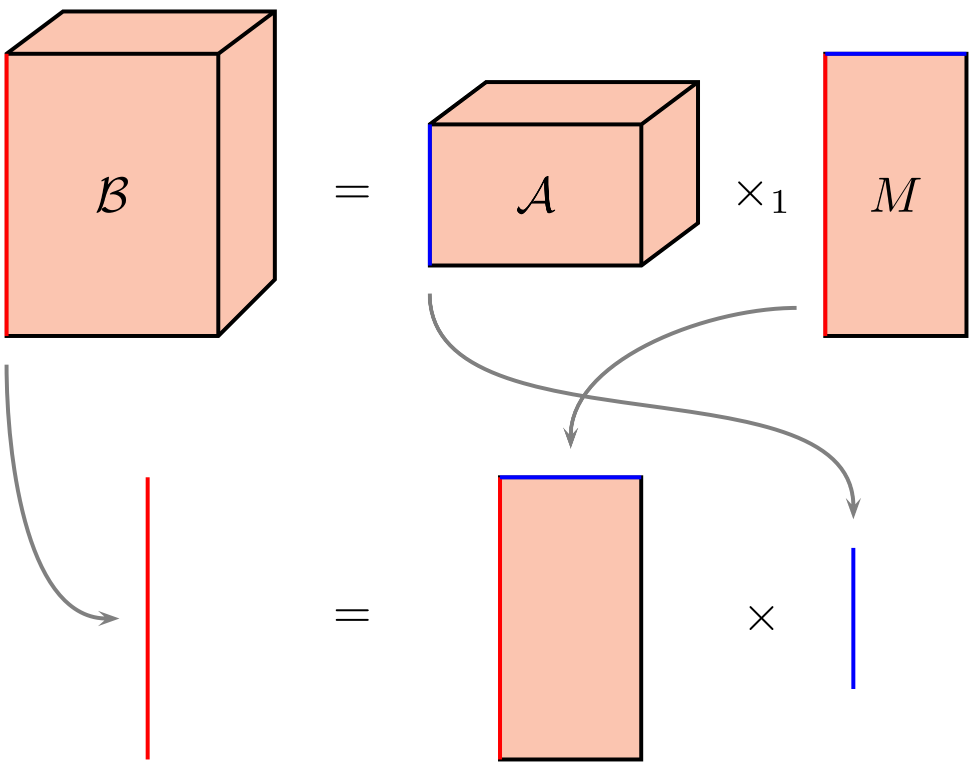
Jie Chen and Yousef Saad

Department of Computer Science and Engineering
University of Minnesota

Mode-n Product

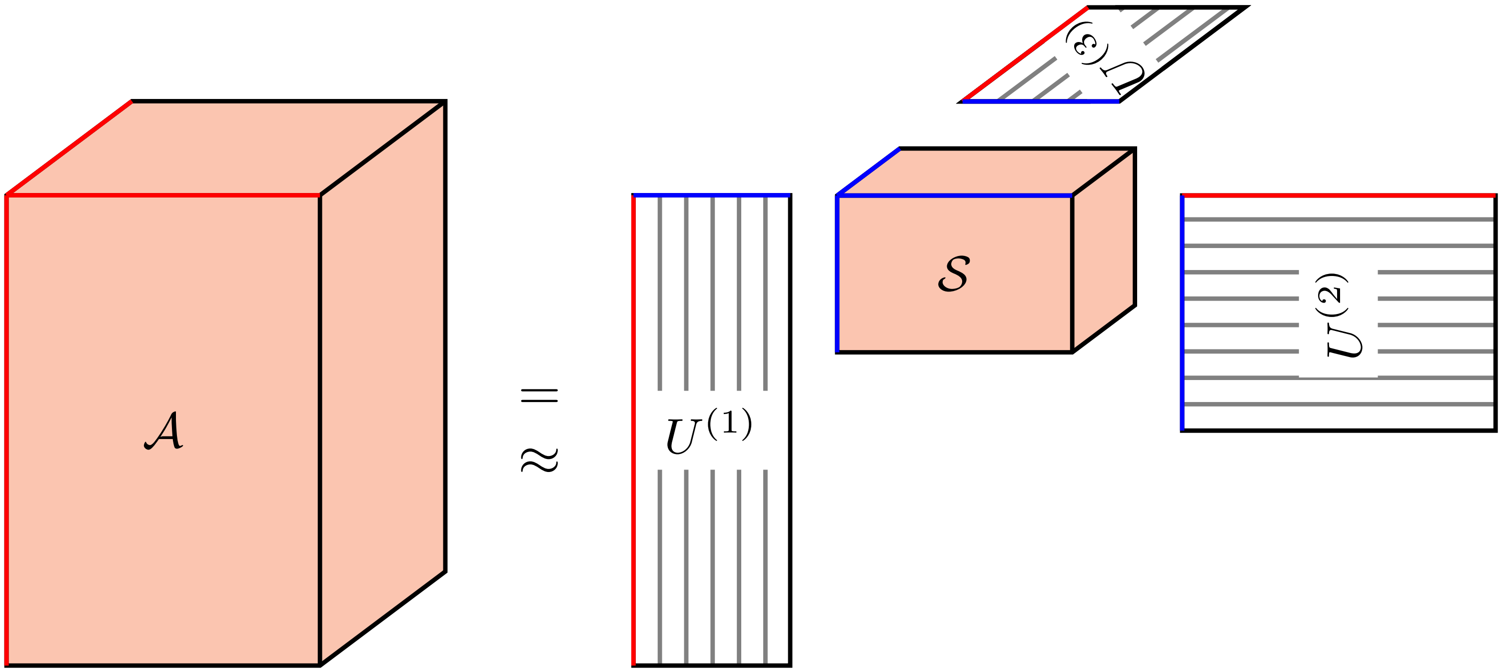
Mode-n product $\mathcal{B} = \mathcal{A} \times_n M$

$$\mathcal{B}(\dots, i_{n-1}, j_n, i_{n+1}, \dots) = \sum_{i_n} \mathcal{A}(\dots, i_{n-1}, i_n, i_{n+1}, \dots) M(j_n, i_n)$$



Tensor Factorizations/Approximations

$$\mathcal{A} = \text{or} \approx \mathcal{S} \times_1 U^{(1)} \times_2 U^{(2)} \times \cdots \times_N U^{(N)}$$



When $\mathcal{S}$ is diagonal, $\sum_{i=1}^{r} s_{ii\ldots i}\, u_i^{(1)} \otimes u_i^{(2)} \otimes \cdots \otimes u_i^{(N)}$.

Tensor Factorizations/Approximations

$$\mathcal{A} = \mathcal{S} \times_1 U^{(1)} \times_2 U^{(2)} \times \cdots \times_N U^{(N)}$$

What can be done and what cannot be done?

Strict equality:

- $\mathcal{S}$ diagonal: Always.
- $\mathcal{S}$ diagonal with a minimum number (r) of nonzeros: Tensor rank problem. NP-complete over a finite field, and NP-hard for rational numbers. [Håstad 1990]
- Above case, and $\mathcal{A}$ symmetric: Homogeneous polynomial decomposition. [Comon et al 2008; Brachat et al 2009]
- Each $U^{(n)}$ orthogonal: HOSVD. [De Lathauwer et al 2000]
- $\mathcal{S}$ diagonal, each $U^{(n)}$ orthogonal: Rare.

Tensor Factorizations/Approximations

$$\mathcal{A} \approx \mathcal{S} \times_1 U^{(1)} \times_2 U^{(2)} \times \cdots \times_N U^{(N)}$$

What can be done and what cannot be done?

Approximation in an optimal sense:

- $\mathcal{S}$ diagonal: CANDECOMP/PARAFAC. Optimum may not exist; ill-posed. [Harshman 1970; Carroll and Chang 1970; De Silva and Lim 2008]
- Each $U^{(n)}$ orthogonal: Tucker/HOOI. [Tucker 1966; De Lathauwer et al 2000]
- $\mathcal{S}$ scalar, each $U^{(n)}$ a vector: Optimal rank-1 approximation. [De Lathauwer et al 2000; Zhang and Golub 2001; Kofidis and Regalia 2001]
- $\mathcal{S}$ diagonal, each $U^{(n)}$ orthogonal: LROAT.

LROAT

Low Rank Orthogonal Approximation of Tensor $\mathcal{A}$:

$$\min \quad E = \left\| \mathcal{A} - \textstyle\sum_{i=1}^{r} \sigma_i u_i^{(1)} \otimes u_i^{(2)} \otimes \cdots \otimes u_i^{(N)} \right\|_F$$

$$\text{s.t.} \quad \left\langle u_j^{(n)}, u_k^{(n)} \right\rangle = \delta_{jk}, \quad \text{for } n = 1, 2, \ldots, N.$$

In mode-n product form, this is

$$\min \left\| \mathcal{A} - \mathcal{S} \times_1 U^{(1)} \times_2 U^{(2)} \times \cdots \times_N U^{(N)} \right\|_F ,$$

where

- $\mathcal{S}$ is diagonal (with diagonal entries σ_i's),
- $u_i^{(n)}$'s are the orthonormal columns of $U^{(n)}$.

LROAT: Properties

$$\underbrace{\mathcal{A}}_{\text{size } d_1 \times d_2 \times \cdots \times d_N} - \sum_{i=1}^{r} \sigma_i \underbrace{\boxed{u_i^{(1)} \otimes u_i^{(2)} \otimes \cdots \otimes u_i^{(N)}}}_{\mathcal{T}_i}$$

- rank$(\sum_{i=1}^{r} \sigma_i \mathcal{T}_i) = r$.
  (Guaranteed by the linear independence of the $u_i^{(n)}$ vectors.)
- At optimality,

$$\sigma_i = \langle \mathcal{A}, \mathcal{T}_i \rangle_F = \mathcal{A} \times_1 u_i^{(1)^T} \times_2 u_i^{(2)^T} \times \cdots \times_N u_i^{(N)^T},$$

$$\|\mathcal{A} - \sum_{i=1}^{r} \sigma_i \mathcal{T}_i\|_F^2 = \|\mathcal{A}\|_F^2 - \sum_{i=1}^{r} \sigma_i^2.$$

  Hence, $\min \|\mathcal{A} - \sum_{i=1}^{r} \sigma_i \mathcal{T}_i\|_F^2 = \max \sum_{i=1}^{r} \sigma_i^2$.
- The global optimum exists for all $r \leq \min\{d_1, \ldots, d_N\}$.
  (Because the feasible region is compact.)

LROAT: Maximal Diagonal

From a different perspective, let

$$\tilde{U}^{(n)} = \left[U^{(n)},\ {U^{(n)}}^{\perp} \right] \quad \text{square orthogonal matrix.}$$

Define

$$\tilde{\mathcal{S}} = \mathcal{A} \times_1 {\tilde{U}^{(1)}}^T \times_2 {\tilde{U}^{(2)}}^T \times \cdots \times_N {\tilde{U}^{(N)}}^T.$$

Then

$$\mathcal{A} = \tilde{\mathcal{S}} \times_1 \tilde{U}^{(1)} \times_2 \tilde{U}^{(2)} \times \cdots \times_N \tilde{U}^{(N)}.$$

The diagonal entries of $\tilde{\mathcal{S}}$ are nothing but σ_i's.

- LROAT is equivalent to the maximal diagonality problem.

LROAT: Equivalent Problem

What to be presented, is an iterative algorithm to solve:

$$
\begin{array}{ll}
\max & E = \sum_{i=1}^{r} \left(\mathcal{A} \times_1 {u_i^{(1)}}^T \times_2 {u_i^{(2)}}^T \times \cdots \times_N {u_i^{(N)}}^T \right)^2 \\
\text{s.t.} & \left\langle u_j^{(n)}, u_k^{(n)} \right\rangle = \delta_{jk}, \quad \text{for } n = 1, 2, \ldots, N.
\end{array}
$$

Bare a few questions in mind:

1. Is the optimization problem well-posed? (Yes.)
2. Does the algorithm converge?
3. Where does it converge to?
4. How fast does it converge?

LROAT: First Order Condition

Lagrangian:

$$L = \sum_{i=1}^{r} \sigma_i^2 - \sum_{j,k=1}^{r} \sum_{n=1}^{N} \mu_{j,k}^{n} \left(\left\langle u_j^{(n)}, u_k^{(n)} \right\rangle - \delta_{jk} \right).$$

Define

$$v_i^{(n)} = \mathcal{A} \times_1 u_i^{(1)^T} \times \cdots \times_{n-1} u_i^{(n-1)^T} \times_{n+1} u_i^{(n+1)^T} \times \cdots \times_N u_i^{(N)^T} \in \mathbb{R}^{d_n}.$$

Set the gradient of Lagrangian to zero:

$$\frac{\partial L}{\partial u_i^{(n)}} = 2\sigma_i v_i^{(n)} - \sum_{j=1}^{r} \mu_{j,i}^{n} u_j^{(n)} - \sum_{k=1}^{r} \mu_{i,k}^{n} u_k^{(n)} = 0.$$

LROAT: First Order Condition

In matrix form,

$$\begin{bmatrix} v_1^{(n)} \cdots v_r^{(n)} \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} = \begin{bmatrix} u_1^{(n)} \cdots u_r^{(n)} \end{bmatrix} \begin{bmatrix} \frac{\mu_{1,1}^n + \mu_{1,1}^n}{2} & \cdots & \frac{\mu_{1,r}^n + \mu_{r,1}^n}{2} \\ \vdots & \ddots & \vdots \\ \frac{\mu_{r,1}^n + \mu_{1,r}^n}{2} & \cdots & \frac{\mu_{r,r}^n + \mu_{r,r}^n}{2} \end{bmatrix}$$

$$\boxed{V^{(n)}\Sigma = U^{(n)}M^{(n)}}$$

Interpret $U^{(n)}M^{(n)}$ as the polar factoraization of the matrix $V^{(n)}\Sigma$.

LROAT: Algorithm

$$\boxed{V^{(n)}\Sigma = U^{(n)}M^{(n)}}$$

Algorithm:

1: Initialize each $U^{(n)}$
2: **repeat**
3: **for** $n \leftarrow 1, \ldots, N$ **do**
4: Compute $V^{(n)}$
5: Compute Σ
6: Update $U^{(n)} \leftarrow$ polar-factor$(V^{(n)}\Sigma)$
7: **end for**
8: **until** convergence

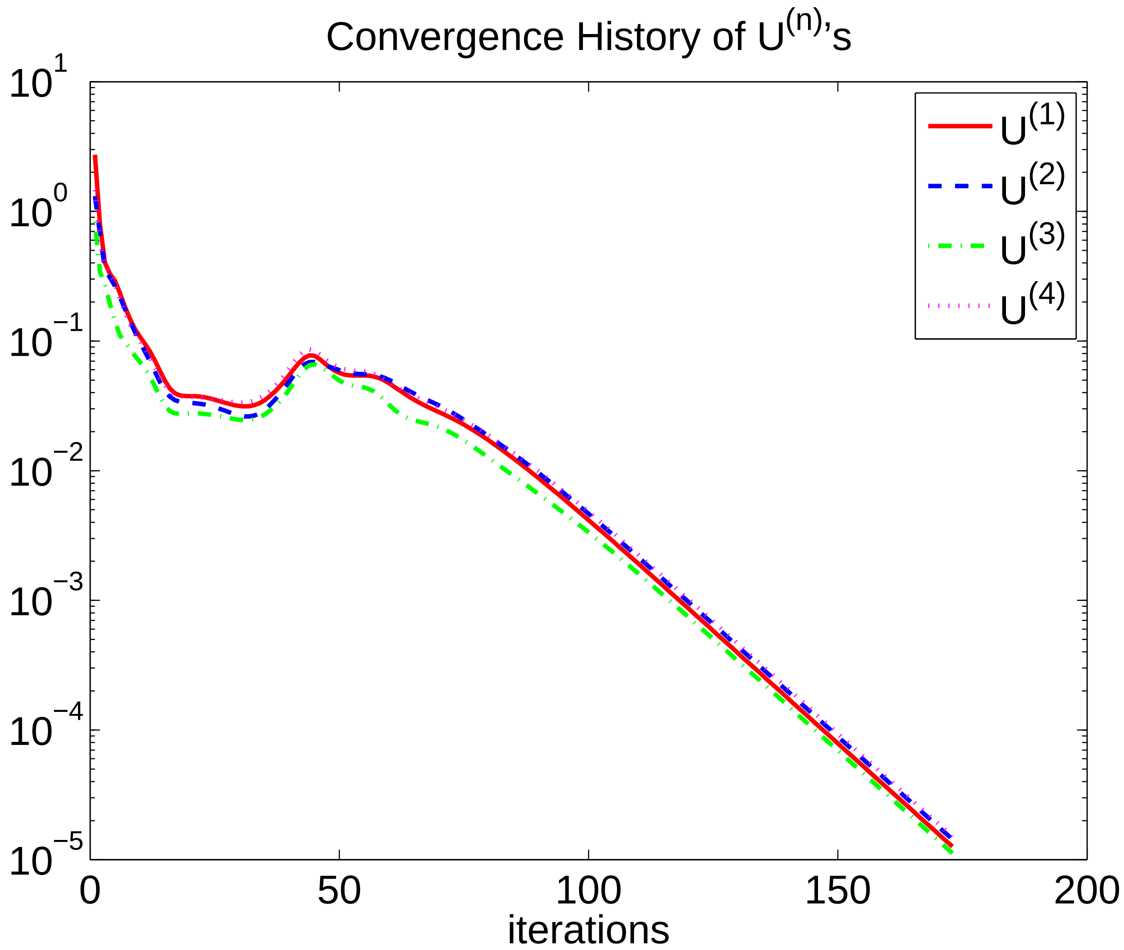
LROAT: Convergence

- Each update increases the objective $\sum_{i=1}^{r} \sigma_i^2$.
  - Use the fact that $\sum_{i=1}^{r} \sigma_i^2 = \mathrm{tr}\left({U^{(n)}}^T V^{(n)} \Sigma \right)$.
  - This implies at least the objective function value converges.
- Every limit point of the parameter ($u_i^{(n)}$'s) sequence is stationary, i.e., in the limit, $V^{(n)}\Sigma = U^{(n)} M^{(n)}$ is satisfied for $n = 1, \ldots, N$.
  - Proved by a fixed point lemma, applied to the fixed point mapping $V^{(n)}\Sigma \to U^{(n)}$.
  - Requires that $V^{(n)}$ does not become rank deficient during iterations.

LROAT: A Few Notes

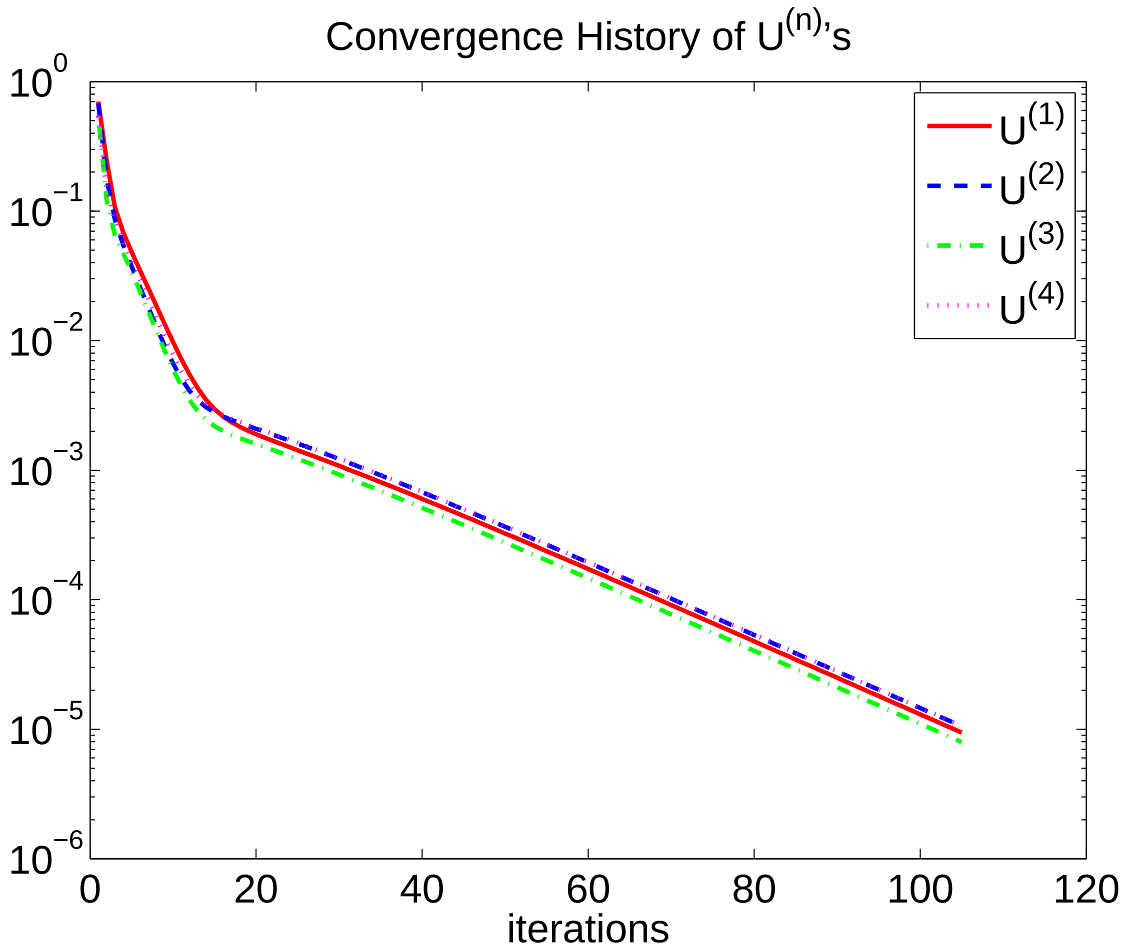
- The method is in general not an alternating least squares / block coordinate descent method. (Each update does not maximize $\sum_{i=1}^{r} \sigma_i^2$.)
- When $r = 1$, the method boils down to alternating least squares / higher-order power method, for computing the optimal rank-1 approximation.
- Symmetric LROAT?
  - Case: $\mathcal{A}$ is symmetric. Requires: $U^{(n)}$ be the same for all n.
  - First order condition: $V\Sigma = UM$.
  - Similar update: $U \leftarrow$ polar-factor$(V\Sigma)$.
  - However, not always converge! (The objective $\sum_{i=1}^{r} \sigma_i^2$ no longer monotonically increases.)

LROAT: Rate of Convergence



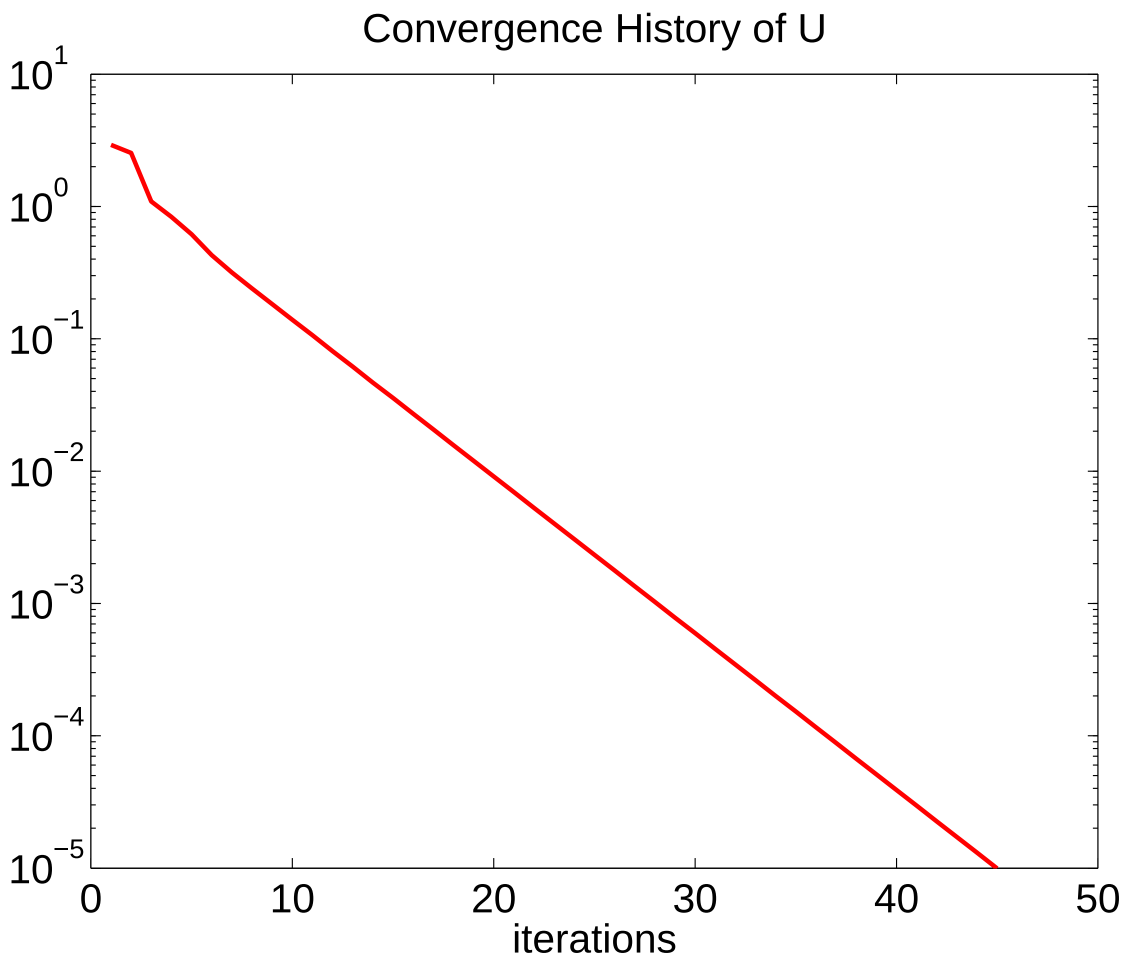
Random tensor, $20 \times 16 \times 10 \times 32$, $r = 5$. LROAT

LROAT: Rate of Convergence



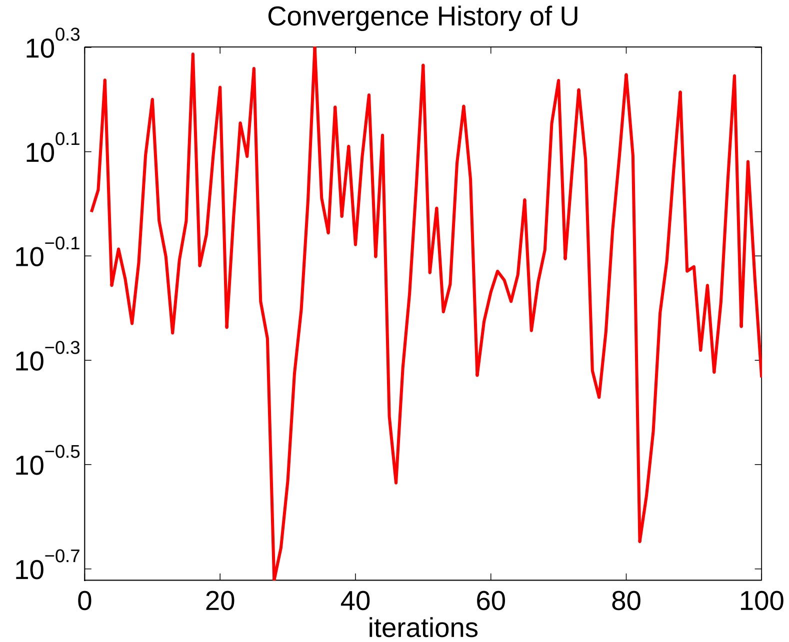
rank-5 + noise, $20 \times 16 \times 10 \times 32$, $r = 5$. LROAT

LROAT: Rate of Convergence

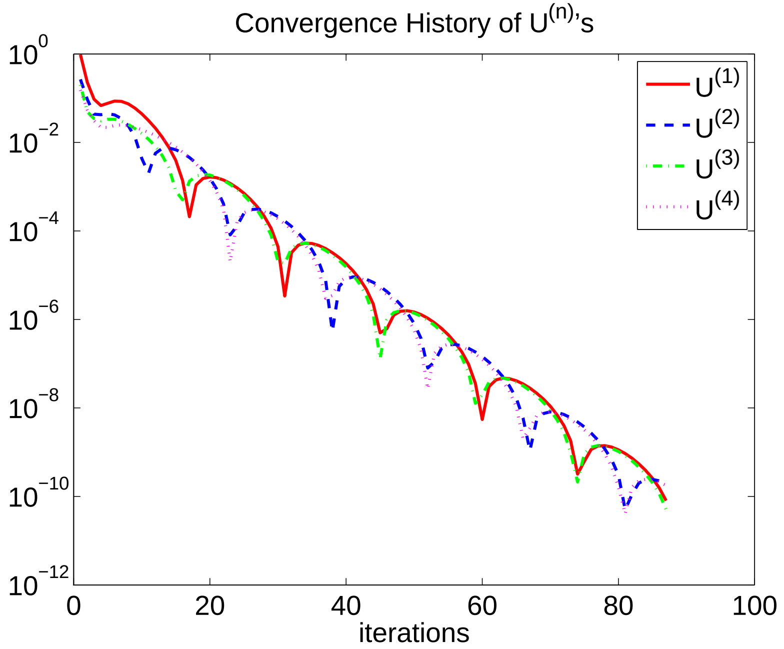


$\mathcal{A}(i,j,k) = \frac{1}{i^2+j^2+k^2}$, $10 \times 10 \times 10$, $r = 5$. Symmetric LROAT

LROAT: Rate of Convergence



Symmetric LROAT



LROAT

Symmetric tensor, $3 \times 3 \times 3 \times 3$, $r = 2$.

Comparison: LROAT, Tucker and PARAFAC

- Generate $\mathcal{A} = \mathcal{C} + \rho\mathcal{D} \in \mathbb{R}^{10\times10\times10\times10}$, where
  - $\mathcal{C} = \sum_{i=1}^{5} \sigma_i u_i \otimes u_i \otimes u_i \otimes u_i \quad (u_i$ orthonormal),
  - $\mathcal{D}$ is a symmetric random tensor,
  - $\rho = 0.05 \, \|\mathcal{C}\|_F \, / \, \|\mathcal{D}\|_F$.
- Approximate $\mathcal{A}$ by three models: LROAT, Tucker, PARAFAC.
- Compare for each approximated tensor:
  - Difference between the factor matrix and $U = [u_1, \ldots, u_5]$.
  - How large is the residual?
- The algorithm for Tucker and PARAFAC: alternating least squares.

Comparison: LROAT, Tucker and PARAFAC

$$U = \begin{bmatrix} -.16 & +.05 & \\ -.61 & -.33 & \\ +.05 & +.22 & \\ +.11 & -.79 & \cdots \\ -.42 & +.00 & \\ +.44 & +.01 & \\ +.44 & -.34 & \\ & \vdots & \end{bmatrix} \quad U_{\mathsf{LROAT}} = \begin{bmatrix} -.16 & +.05 & \\ -.62 & -.34 & \\ +.04 & +.22 & \\ +.11 & -.79 & \cdots \\ -.42 & +.00 & \\ +.44 & +.01 & \\ +.44 & -.34 & \\ & \vdots & \end{bmatrix} \quad U_{\mathsf{Tucker}} = \begin{bmatrix} -.17 & -.04 & \\ -.33 & -.62 & \\ -.09 & +.21 & \\ +.51 & -.60 & \cdots \\ -.36 & -.22 & \\ +.37 & +.25 & \\ +.55 & -.05 & \\ & \vdots & \end{bmatrix} \quad U_{\mathsf{PARAFAC}} = \begin{bmatrix} -.16 & +.05 & \\ -.61 & -.33 & \\ +.04 & +.22 & \\ +.11 & -.79 & \cdots \\ -.42 & +.00 & \\ +.44 & +.01 & \\ +.44 & -.34 & \\ & \vdots & \end{bmatrix}$$

$$\|U - U_{\mathsf{LROAT}}\| = 0.0535, \quad \|U - U_{\mathsf{Tucker}}\| = 0.5711, \quad \|U - U_{\mathsf{PARAFAC}}\| = 0.0680.$$

$$\frac{\|\mathcal{A} - \mathcal{A}_{\mathsf{LROAT}}\|_F}{\|\mathcal{A}\|_F} = 9.69\%, \quad \frac{\|\mathcal{A} - \mathcal{A}_{\mathsf{Tucker}}\|_F}{\|\mathcal{A}\|_F} = 9.42\%, \quad \frac{\|\mathcal{A} - \mathcal{A}_{\mathsf{PARAFAC}}\|_F}{\|\mathcal{A}\|_F} = 9.65\%.$$

Conclusion

- A new approximation model—diagonal core, orthogonal factor matrices.
- An algorithm that guarantees convergence.
- Empirically linear convergence.
- Can be used to maximize the diagonal of the core.
- Can be useful for certain applications where the tensor itself has orthogonal factors.

References

- J. Chen and Y. Saad. On the Tensor SVD and the Optimal Low Rank Orthogonal Approximation of Tensors. SIMAX, 2009. (LROAT)
- T. G. Kolda and B. W. Bader. Tensor Decompositions and Applications. SIAM Review, to appear. (a survey)
- J. Håstad. Tensor rank is NP-complete. J. Algorithms, 1990. (tensor rank)
- V. de Silva and L.-H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. SIMAX, 2008. (tensor rank)
- P. Comon et al. Symmetric Tensors and Symmetric Tensor Rank. SIMAX, 2008. (tensor rank)
- J. Brachat et al. Symmetric tensor decomposition. arXiv:0901.3706v2. (homogeneous polynomial decomposition)
- E. Kofidis and P. A. Regalia. On the Best Rank-1 Approximation of Higher-Order Supersymmetric Tensors. SIMAX, 2001. (symmetric power iteration, oscillating behavior)

References

- L. R. Tucker. Some mathematical notes on three-mode factor analysis. Psychometrika, 1966. (Tucker)
- J. D. Carroll and J.-J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. Psychometrika, 1970. (CANDECOMP)
- R. A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. UCLA Working Papers in Phonetics, 1970. (PARAFAC)
- L. De Lathauwer et al. A Multilinear Singular Value Decomposition, SIMAX, 2000. (HOSVD, maximal diagonal)
- L. De Lathauwer et al. On the Best Rank-1 and Rank-$(R_1, R_2, \ldots, R_N)$ Approximation of Higher-Order Tensors. SIMAX, 2000. (optimal rank-1 approximation)
- T. Zhang and G. H. Golub. Rank-One Approximation to High Order Tensors. SIMAX, 2001. (optimal rank-1 approximation)