Inter-University Research Institute Corporation / Research Organization of Information and Systems
National Institute of Informatics

Fast Tucker Rank Reduction for Non-Negative Tensors Using Mean-Field Approximation

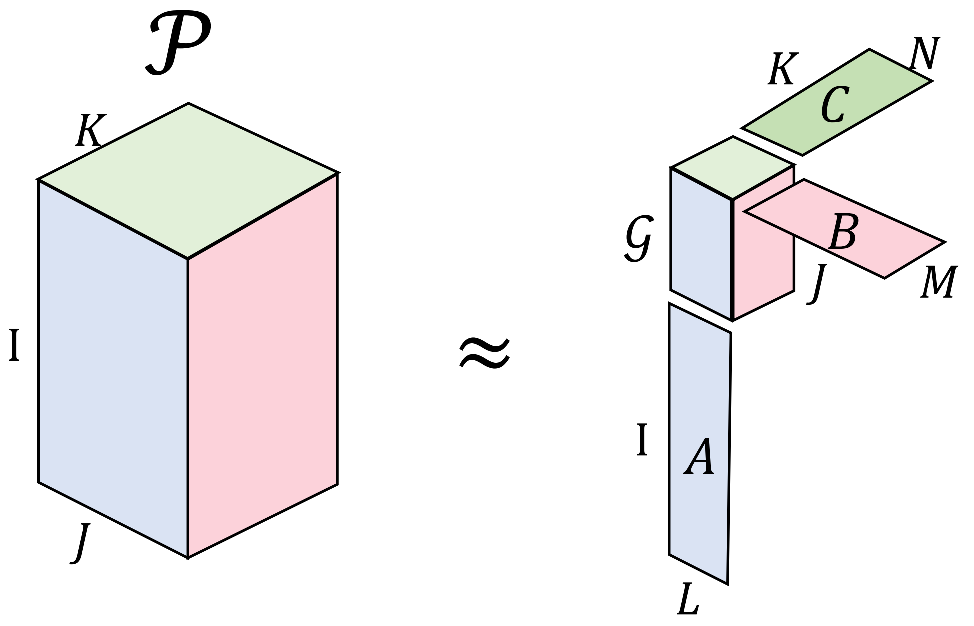
Kazu Ghalamkari[1,2], Mahito Sugiyama[1,2]

1 : The Graduate University for Advanced Studies, SOKENDAI
2 : National Institute of Informatics

github.com/gkazunii/Legendre-tucker-rank-reduction

Conference on Neural Information Processing Systems 2021

Tensor low-rank approximation



$$\mathcal{P}_{ijk} = \sum_{l=1}^{L}\sum_{m=1}^{M}\sum_{n=1}^{N} \mathcal{G}_{ijk}\, A_{il}B_{jm}C_{kn}$$

Tucker rank of $\mathcal{P}$: (L, M, N) $\quad L \leq I, M \leq J, N \leq K$

In this presentation, Tucker rank is simply referred to as Rank.

Rank (1,1,1) is called just rank 1 .

Approximating tensors with low rank tensors reduce memory requirements.

Frobenius error or KL error minimization

Many non-negative low-rank approximation methods are based on a **gradient method**.
→ It requires appropriate settings for **initial values, stopping criterion, and learning rates.** 😢😢

This study...

・Maps non-negative tensors to distributions, and derive a formula for **the best rank-1 approximation**
・Understands the rank-1 approximation of a non-negative tensor as a **mean-field approximation**
・Proposes a fast Tucker rank approximation (**LTR**) for nonnegative tensors based on the formula

Contents

- Best Rank-1 Approximation for Minimizing the KL divergence
  - Make a mapping from a tensor to a distribution
  - A rank-1 condition using parameters of the distribution
  - Rank-1 Approximation as a mean-field approximation
- Legendre Tucker Rank Reduction(LTR)
  - A fast low-rank approximation for non-negative tensors
  - Not based on a gradient method.
- Experiment
- Conclusion

Introduction of log-linear model on poset

Poset

S is a poset $\Leftrightarrow$ for all $s_1, s_2, s_3 \in S$ the following three properties are satisfied.

(1) **Reflexivity**: $s_1 \leq s_1$ (2) **Antisymmetry:** $s_1 \leq s_2, s_2 \leq s_1 \Rightarrow s_1 = s_2$ (3) **transitivity:**$s_1 \leq s_2, s_2 \leq s_3 \Rightarrow s_1 \leq s_3$

Log-linear model on poset S

We define the log-linear model on a poset S as a mapping $p: S \rightarrow (0,1)$. **Natural parameters $\boldsymbol{\theta}$** describe the model.

$$p_\theta(x) = \exp\left[\sum_{s \leq x} \theta(s)\right] \quad x \in S$$

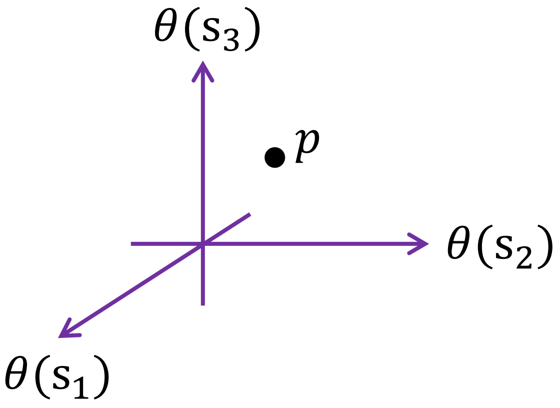
We can also describe the model by **expectation parameters η** if we use Zeta function.

$$\eta(x) = \sum_{s \geq x} p(s) \qquad p_\eta(x) = \sum_{s \in S} \mu(x,s)\eta(s)$$

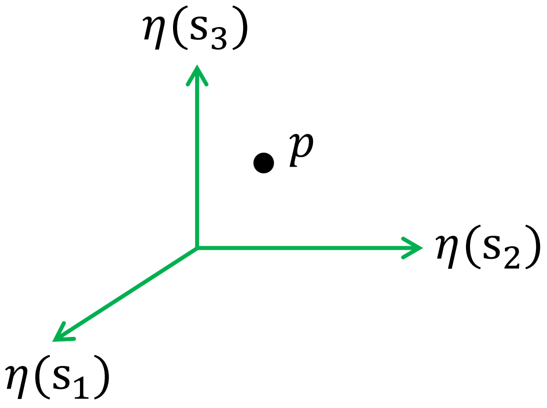
Möbius function

$$\mu(x,s) = \begin{cases} 1 & \text{if } x = y \\ -\sum_{x \leq s < y} \mu(x,s) & \text{if } x < y \\ 0 & \text{otherwise} \end{cases}$$

θ-space



η-space

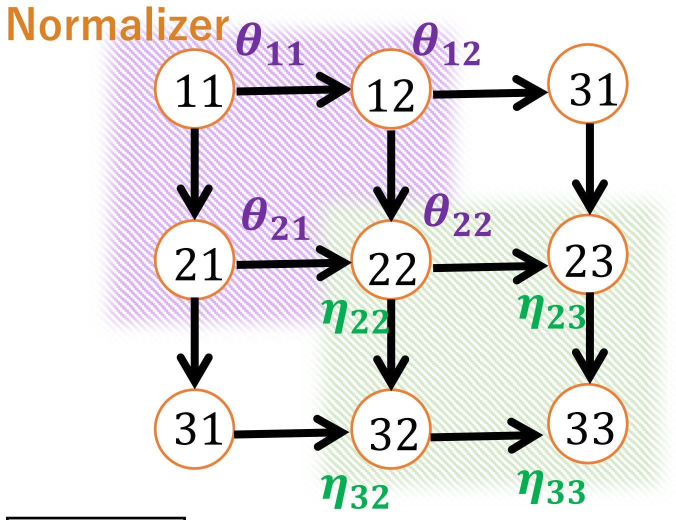


Mahito Sugiyama, Hiroyuki Nakahara and Koji Tsuda
"Tensor balancing on statistical manifold"(2017) ICML.

Mapping a poset to a non-negative matrix/tensor.

Matrix

$S = \{(i,j) | i,j = \{1,2,\cdots n\}\}$ $\quad (i_1,j_1) \le (i_2,j_2) \Leftrightarrow i_1 \le i_2$ and $j_1 \le j_2$



$$p_\theta(2,2) = \exp[\theta_{11} + \theta_{12} + \theta_{21} + \theta_{22}]$$

$$p_\eta(2,2) = \eta_{22} - \eta_{23} - \eta_{32} + \eta_{33}$$

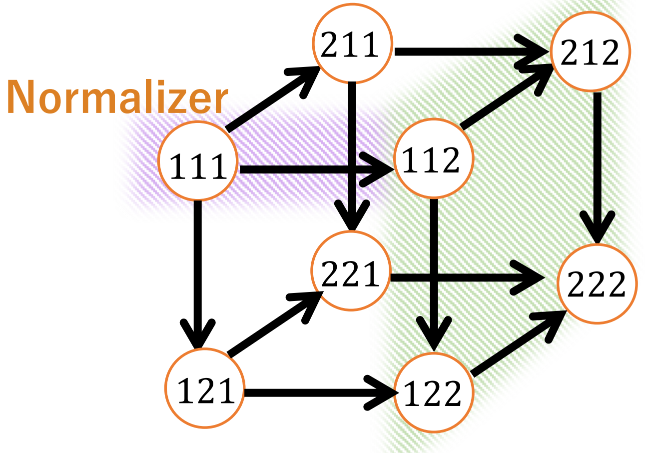
$$p_\theta(i,j) = \exp\left[\sum_{i' \le i} \sum_{j' \le j} \theta_{i'j'}\right]$$

Relation between distribution and matrix

Random Variable : index i,j
Sample space : index set
Value of the probability : element P_{ij}

Tensor

$S = \{(i,j,k) | i,j,k = \{1,2,\cdots n\}\}$ $\quad (i_1,j_1,k_1) \le (i_2,j_2,k_2) \Leftrightarrow i_1 \le i_2$ and $j_1 \le j_2$ and $k_1 \le k_2$



$$p_\theta(1,1,2) = \exp[\theta_{111} + \theta_{112}]$$

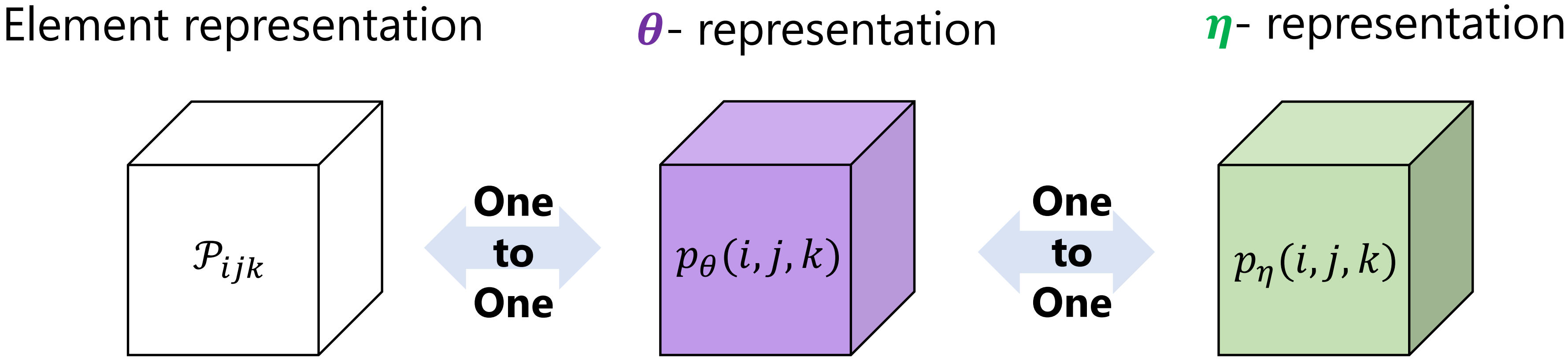
$$p_\eta(1,1,2) = \eta_{222} - \eta_{221} - \eta_{122} + \eta_{112}$$

$$p_\theta(i,j,k) = \exp\left[\sum_{i' \le i} \sum_{j' \le j} \sum_{k' \le k} \theta_{i'j'k'}\right]$$

Relation between distribution and matrix

Random Variable : index i,j,k
Sample space : index set
Value of the probability : element P_{ijk}

Various representations of a normalized tensor



We can describe matrix properties by using $\boldsymbol{\theta}$- and $\boldsymbol{\eta}$- representations.

Easier to formulate as a convex problem.

Describe the rank-1 condition of a tensor using (θ,η)

one-body parameters

$\theta_{i11}, \theta_{1j1}, \theta_{11k} \qquad \eta_{i11}, \eta_{1j1}, \eta_{11k}$

Only one index is 1.

many-body parameters

A parameter other than a one-body parameter

Rank-1 condition ($\boldsymbol{\theta}$-representation)

$\mathrm{rank}(\mathcal{P}) = 1 \Leftrightarrow$ its all many-body θ parameters are 0

$(\Longleftarrow)$

$$\mathcal{P}_{ijk} = \exp\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\theta_{ijk}\right] = \exp[\theta_{111}]\exp\left[\sum_{i=2}^{I}\theta_{i11}\right]\exp\left[\sum_{j=2}^{J}\theta_{1j1}\right]\exp\left[\sum_{k=2}^{K}\theta_{11k}\right]$$

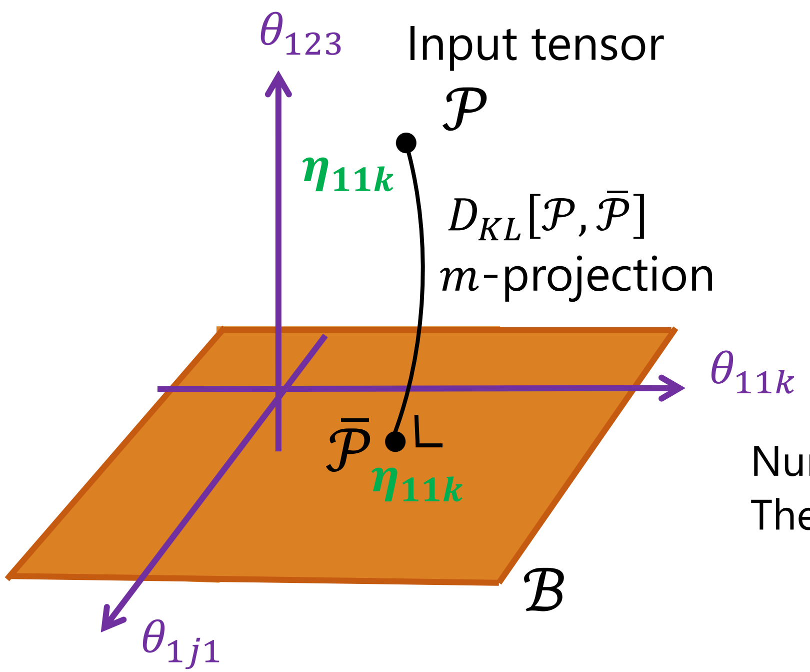
$$\mathcal{P} = e^{\theta_{111}}\begin{pmatrix} 1 \\ e^{\theta_{211}} \\ e^{\theta_{211}+\theta_{311}} \\ \vdots \\ e^{\theta_{211}+\theta_{311}+\cdots+\theta_{I11}} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ e^{\theta_{121}} \\ e^{\theta_{121}+\theta_{131}} \\ \vdots \\ e^{\theta_{121}+\theta_{131}+\cdots+\theta_{1J1}} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ e^{\theta_{211}} \\ e^{\theta_{211}+\theta_{311}} \\ \vdots \\ e^{\theta_{211}+\theta_{311}+\cdots+\theta_{11K}} \end{pmatrix}$$

The rank of the tensor that can be represented by the Kronecker product of three vectors is 1

■

Projection onto rank-1 space

The rank-1 approximation is a projection onto a subspace $\boldsymbol{\mathcal{B}}$ with all zero many-body natural parameters.



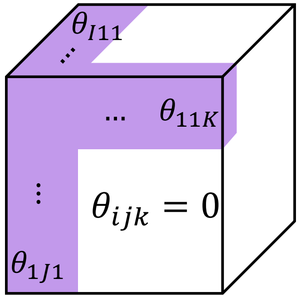
The projection from any input tensor $\mathcal{P} \in \mathbb{R}_{>0}^{I\times J\times K}$ to $\mathcal{B}$ is convex.

But!! It takes too much time to get $\bar{\mathcal{P}}$ using the gradient method.😢😢

Projection from $\mathcal{P}$ *onto* $\mathcal{B}$

Number of natural parameters to optimize is $(I+J+K)$

The computational complexity of the Newton method is $O\left((I+J+K)^3\right)$



The one-body η is invariant to this m-projection

⟺ Summation in each axial direction is invariant for rank-1 approximation

➡ Let us describe the rank-1 condition with the expectation parameter $\boldsymbol{\eta}$.

Describe the rank-1 condition using (θ,η)

Rank-1 condition ($\boldsymbol{\theta}$-representation)

$\mathrm{rank}(\mathcal{P}) = 1 \Leftrightarrow$ its all many-body θ parameters are 0

Rank-1 condition ($\boldsymbol{\eta}$-representation)

$\mathrm{rank}(\mathcal{P}) = 1 \Leftrightarrow$ its all many-body η parameters are factorizable as $\eta_{ijk} = \eta_{i11}\eta_{1j1}\eta_{11k}$

The closed-formula of the best rank 1 approximation

Rank-1 condition ($\boldsymbol{\theta}$-representation)

$\mathrm{rank}(\mathcal{P}) = 1 \Leftrightarrow$ its all many-body θ parameters are 0

Rank-1 condition ($\boldsymbol{\eta}$-representation)

$\mathrm{rank}(\mathcal{P}) = 1 \Leftrightarrow$ its all many-body η parameters are factorizable as $\eta_{ijk} = \eta_{i11}\eta_{1j1}\eta_{11k}$

We derive a solution formula of the best rank-1 approximation.

By the way,
Frobenius error
minimization
is **NP-hard**

Best rank-1 tensor formula for minimizing KL divergence ($d = 3$)

For any given positive tensor $\mathcal{P} \in \mathbb{R}_{>0}^{I\times J\times K}$, its best rank-1 approximation is

$$\bar{\mathcal{P}}_{ijk} = \left(\sum_{j'=1}^{J}\sum_{k'=1}^{K}\mathcal{P}_{ij'k'}\right)\left(\sum_{k'=1}^{K}\sum_{i'=1}^{I}\mathcal{P}_{i'jk'}\right)\left(\sum_{i'=1}^{I}\sum_{j'=1}^{J}\mathcal{P}_{i'j'k}\right),$$

that is, it is hold that

$$\bar{\mathcal{P}} = \operatorname*{argmin}_{\mathcal{Q}:\mathrm{rank}(\mathcal{Q})=1} D_{\mathrm{KL}}(\mathcal{P};\mathcal{Q}).$$

We reproduce the result in K.Huang, et al. "Kullback-Leibler principal component for tensors is not NP-hard." *ACSSC* 2017

Mean-field approximation and rank-1 approximation

Best rank-1 tensor formula for minimizing KL divergence ($d = 3$)

For any given positive tensor $\mathcal{P} \in \mathbb{R}_{>0}^{I\times J\times K}$, its best rank-1 approximation is

$$\bar{\mathcal{P}}_{ijk} = \underbrace{\left(\sum_{j'=1}^{J}\sum_{k'=1}^{K}\mathcal{P}_{ij'k'}\right)}_{\text{Normalized vector depending on only } i}\underbrace{\left(\sum_{k'=1}^{K}\sum_{i'=1}^{I}\mathcal{P}_{i'jk'}\right)}_{\text{Normalized vector depending on only } j}\underbrace{\left(\sum_{i'=1}^{I}\sum_{j'=1}^{J}\mathcal{P}_{i'j'k}\right)}_{\text{Normalized vector depending on only } k}$$

A tensor with d indices is a joint distribution with d random variables.
A vector with only 1 index is an independent distribution with only one random variable.

Rank-1 approximation approximates a joint distribution by a product of independent distributions.

Mean-field approximation **: a methodology in physics for reducing a many-body problem to a one-body problem.**

- Best Rank-1 Approximation for Minimizing the KL divergence
  - Make a mapping from a tensor to a distribution
  - A rank-1 condition using parameters of the distribution
  - Rank-1 Approximation as a mean-field approximation

- Legendre Tucker Rank Reduction (LTR)
  - A fast low-rank approximation for non-negative tensors
  - Not based on Gradient method.
  - No need to discuss learning rate, stopping criterion, or initial values

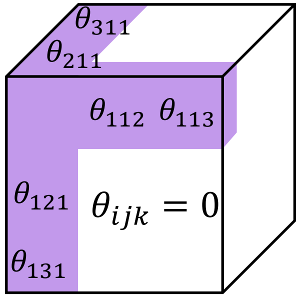
- Experiment

- Conclusion

Formulate Tucker rank reduction by relaxing the rank-1 condition

Rank-1 condition (*$\boldsymbol{\theta}$*-representation)

$\mathrm{rank}(\mathcal{P}) = 1 \Leftrightarrow$ its all many-body θ parameters are 0



Expand the tensor by focusing on the k-th axis into a rectangular matrix $\theta^{(k)}$ (mode-k expansion)

The first row and first column are the scaling factors

$$\theta^{(1)} = \begin{bmatrix} \theta_{111} & \theta_{121} & \theta_{131} & \theta_{112} & 0 & 0 & \theta_{113} & 0 & 0 \\ \theta_{211} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{311} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Two bingos

$$\theta^{(2)} = \begin{bmatrix} \theta_{111} & \theta_{211} & \theta_{311} & \theta_{112} & 0 & 0 & \theta_{311} & 0 & 0 \\ \theta_{121} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{131} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Two bingos

$$\theta^{(3)} = \begin{bmatrix} \theta_{111} & \theta_{211} & \theta_{311} & \theta_{121} & 0 & 0 & \theta_{131} & 0 & 0 \\ \theta_{112} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{113} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Two bingos

Rank (1,1,1)

The relationship between bingo and rank

$$\theta^{(1)} = \begin{bmatrix} \theta_{111} & \theta_{121} & \theta_{131} & \theta_{112} & 0 & 0 & \theta_{113} & 0 & 0 \\ \theta_{211} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{311} & \theta_{321} & \theta_{331} & \theta_{312} & \theta_{322} & \theta_{332} & \theta_{313} & \theta_{323} & \theta_{333} \end{bmatrix}$$

One bingo

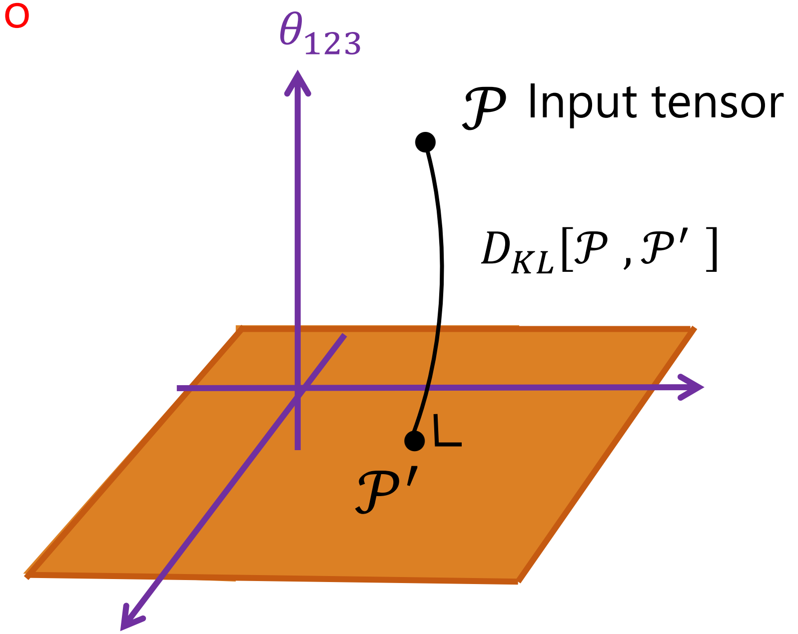
$$\theta^{(2)} = \begin{bmatrix} \theta_{111} & \theta_{211} & \theta_{311} & \theta_{112} & 0 & \theta_{312} & \theta_{311} & 0 & \theta_{313} \\ \theta_{121} & 0 & \theta_{321} & 0 & 0 & \theta_{322} & 0 & 0 & \theta_{323} \\ \theta_{131} & 0 & \theta_{331} & 0 & 0 & \theta_{332} & 0 & 0 & \theta_{333} \end{bmatrix}$$

No bingo

$$\theta^{(3)} = \begin{bmatrix} \theta_{111} & \theta_{211} & \theta_{311} & \theta_{121} & 0 & \theta_{321} & \theta_{131} & 0 & \theta_{331} \\ \theta_{112} & 0 & \theta_{312} & 0 & 0 & \theta_{322} & 0 & 0 & \theta_{332} \\ \theta_{113} & 0 & \theta_{313} & 0 & 0 & \theta_{323} & 0 & 0 & \theta_{333} \end{bmatrix}$$

No bingo

Rank (2,3,3)

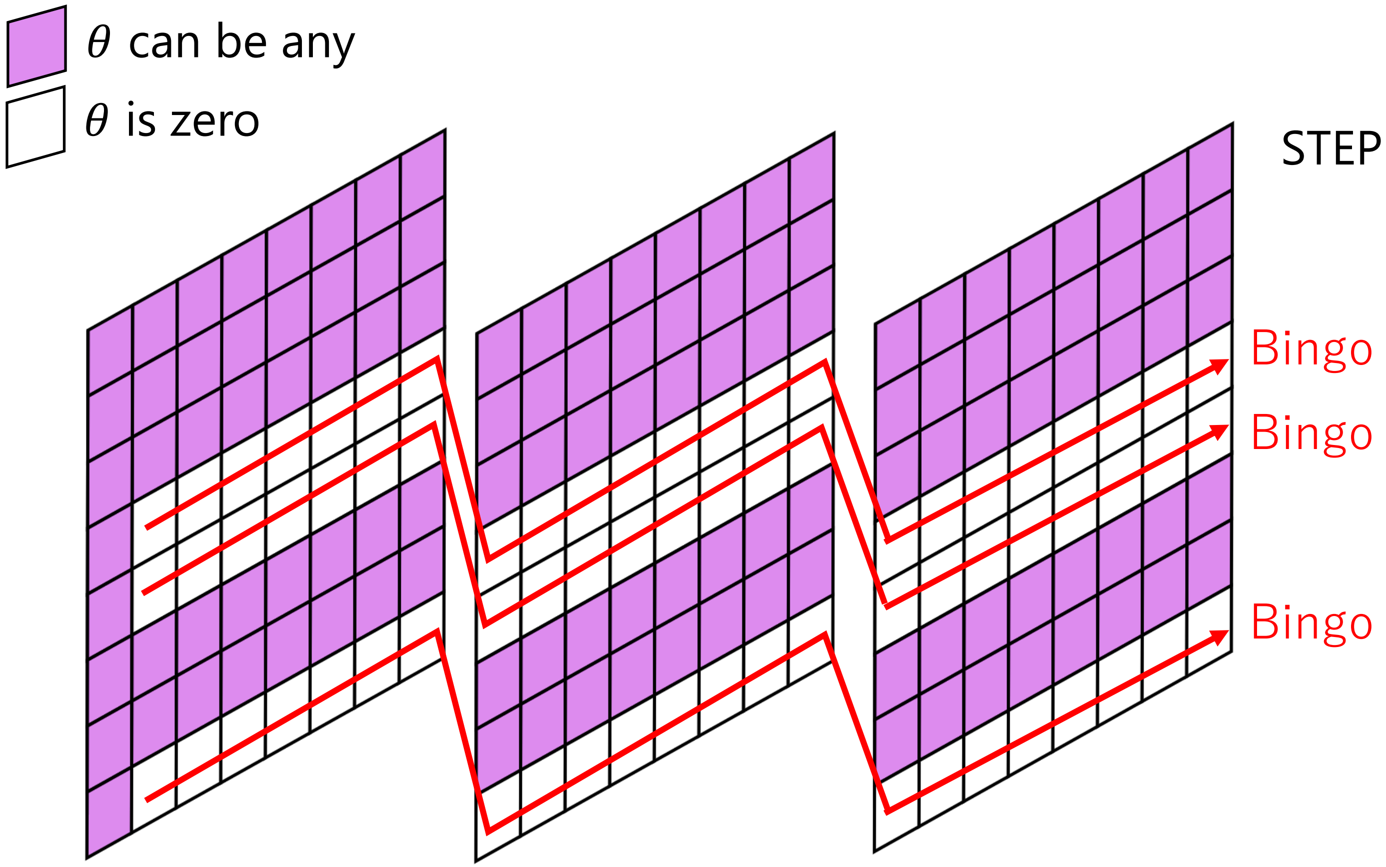


Subspace with one bingo in the mode-1 direction $\mathcal{B}^{(1)}$

Bingo rule ($d = 3$)

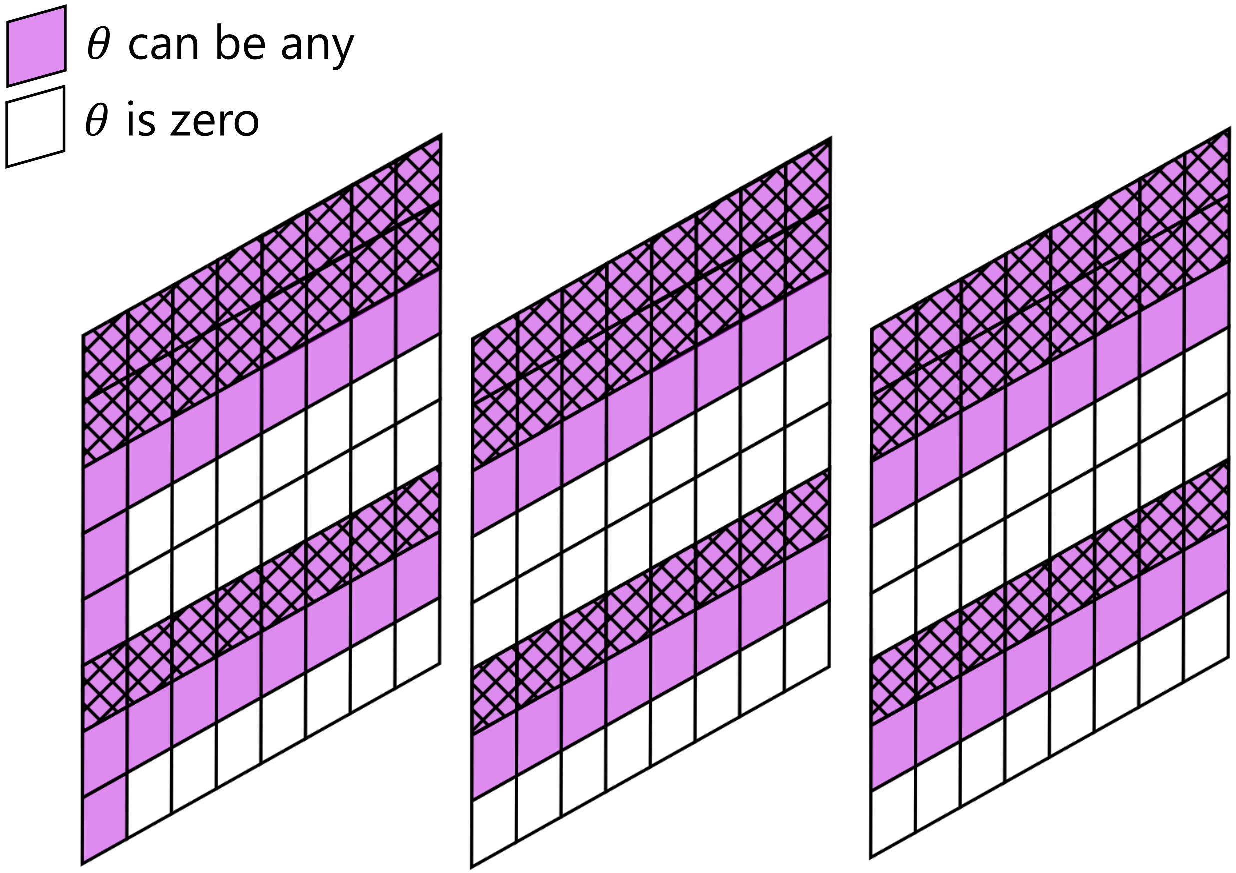
The mode-k expansion $\theta^{(k)}$ of the natural parameter of a tensor $\mathcal{P} \in \mathbb{R}_{>0}^{I_1 \times I_2 \times I_3}$ has b_k bingos
$\Rightarrow \mathrm{rank}(\mathcal{P}) \leq (I_1 - b_1, I_2 - b_2, I_3 - b_3)$

Example: Reduce the rank of (8,8,3) tensor to (**5**,8,3) or less



STEP1 : Choose a bingo location.

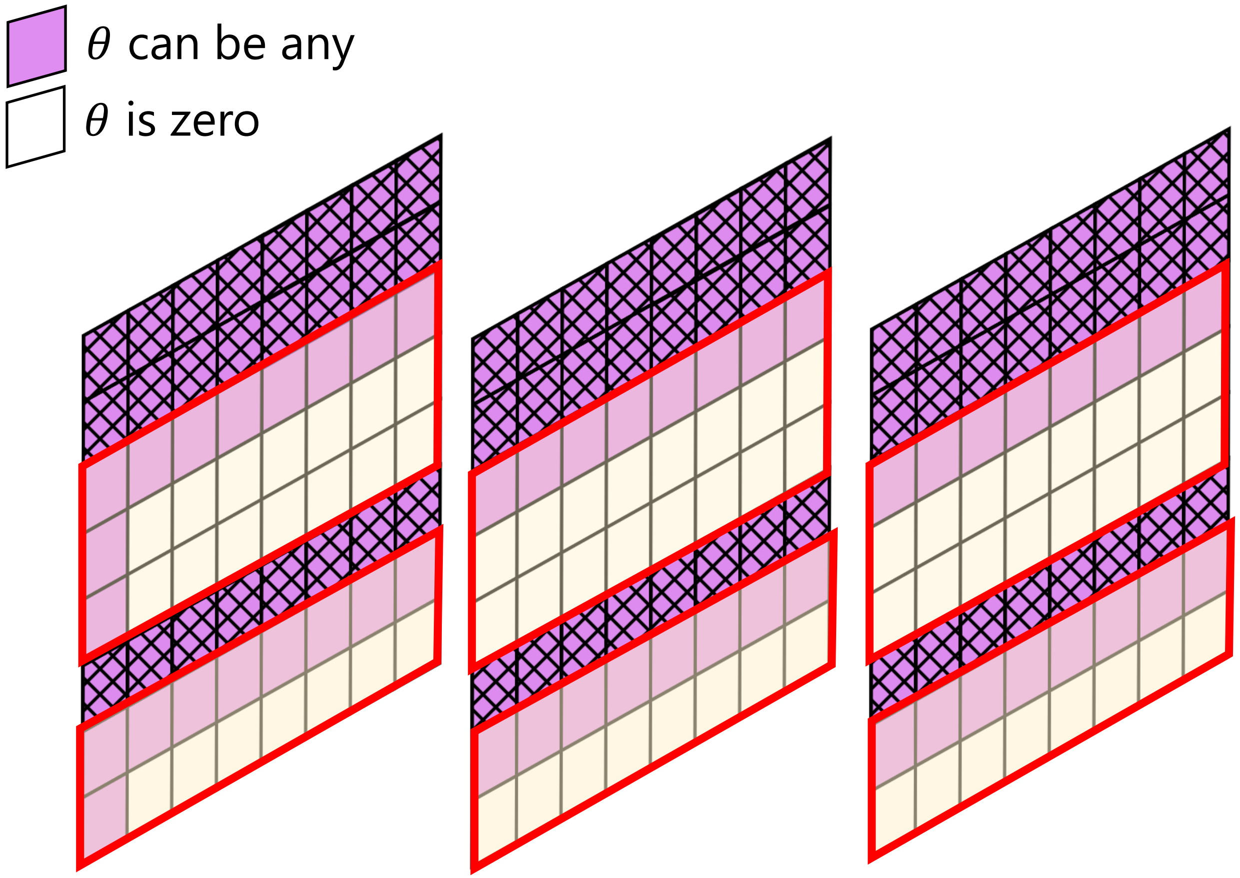
Example: Reduce the rank of (8,8,3) tensor to (**5**,8,3) or less



STEP1 : Choose a bingo location.

The shaded areas do not change their values in the projection.

Example: Reduce the rank of (8,8,3) tensor to (**5**,8,3) or less



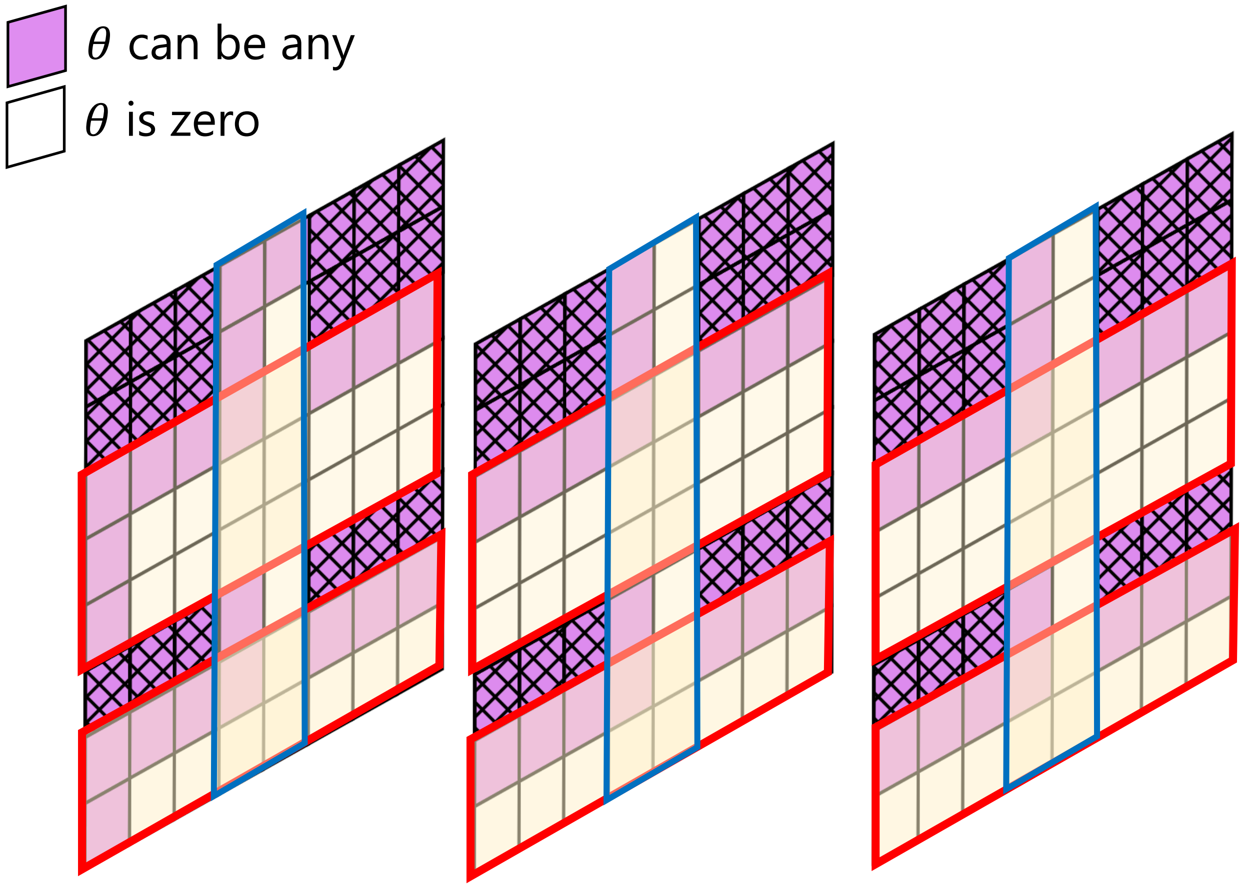
STEP1 : Choose a bingo location.

STEP2 : Replace the bingo part with the best rank-1 tensor.

Replace the partial tensor in the red box using the best rank 1 approximation formula

The best tensor is obtained in the specified bingo space. 😀
There is no guarantee that it is the best rank (5,8,3) approximation. 😢

Example: Reduce the rank of (8,8,3) tensor to (5,7,3) or less.



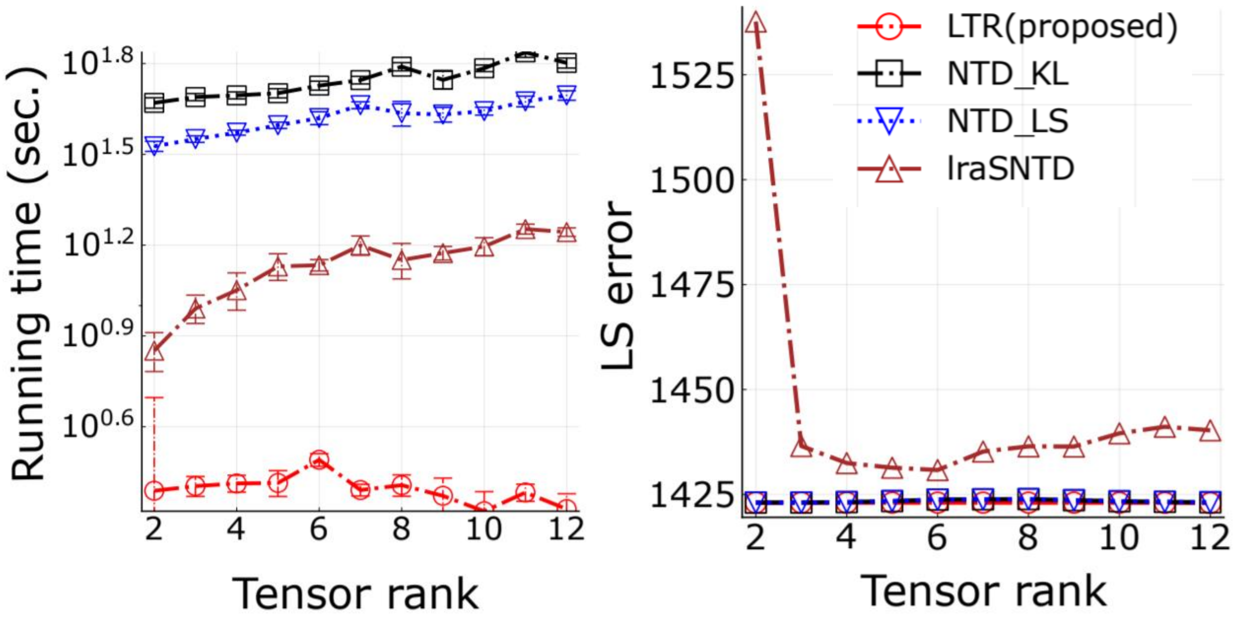
STEP1 : Choose a bingo location.

STEP2 : Replace the bingo part with the best rank-1 tensor.

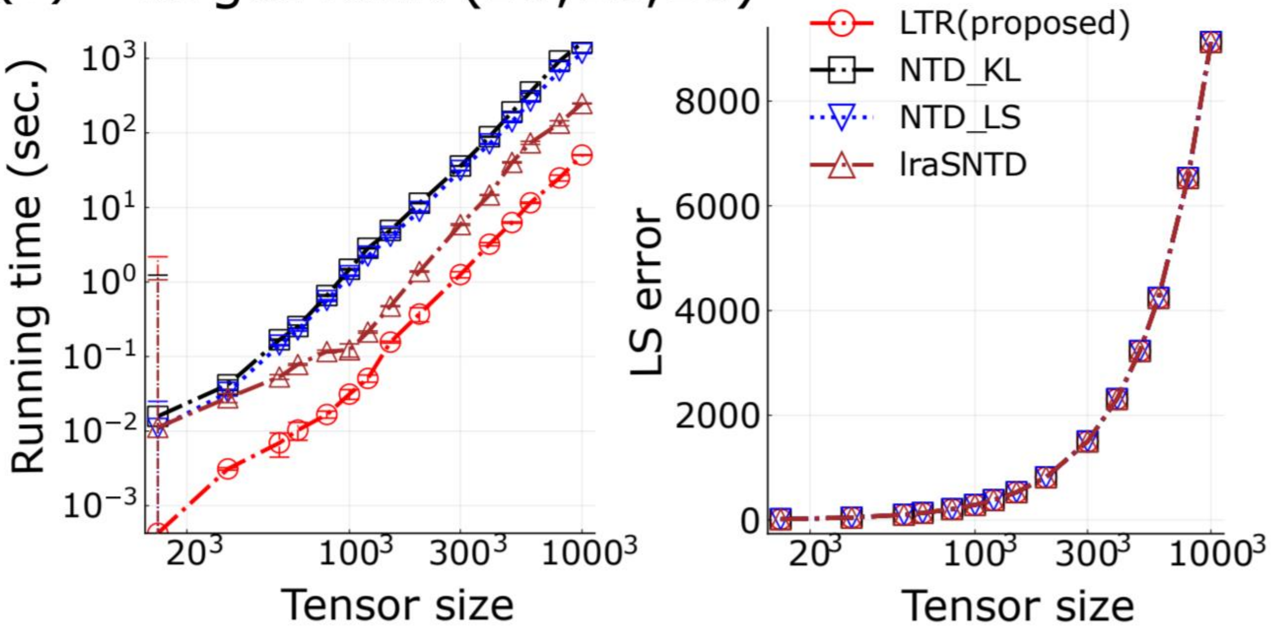
The shaded areas do not change in the projection.

Experimental results (synthetic data)

(a) Random (30,30,30,30,30) tensor

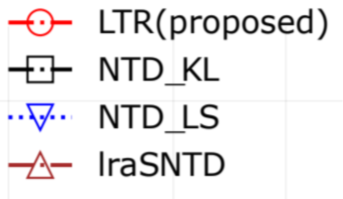


(b) Target rank (10,10,10)

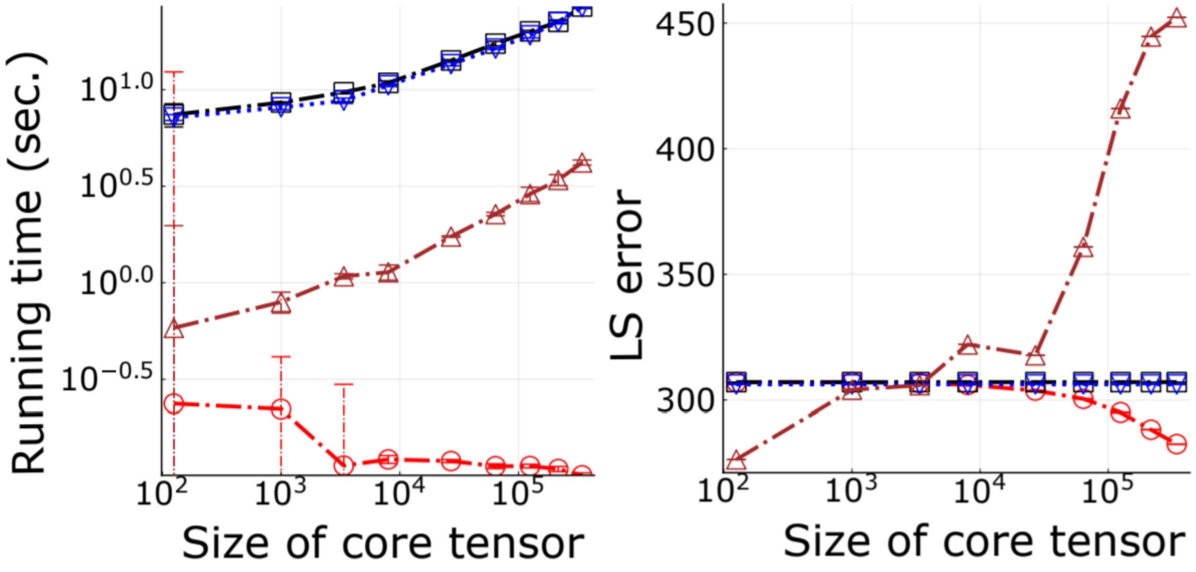


LTR is faster with the competitive approximation performance.

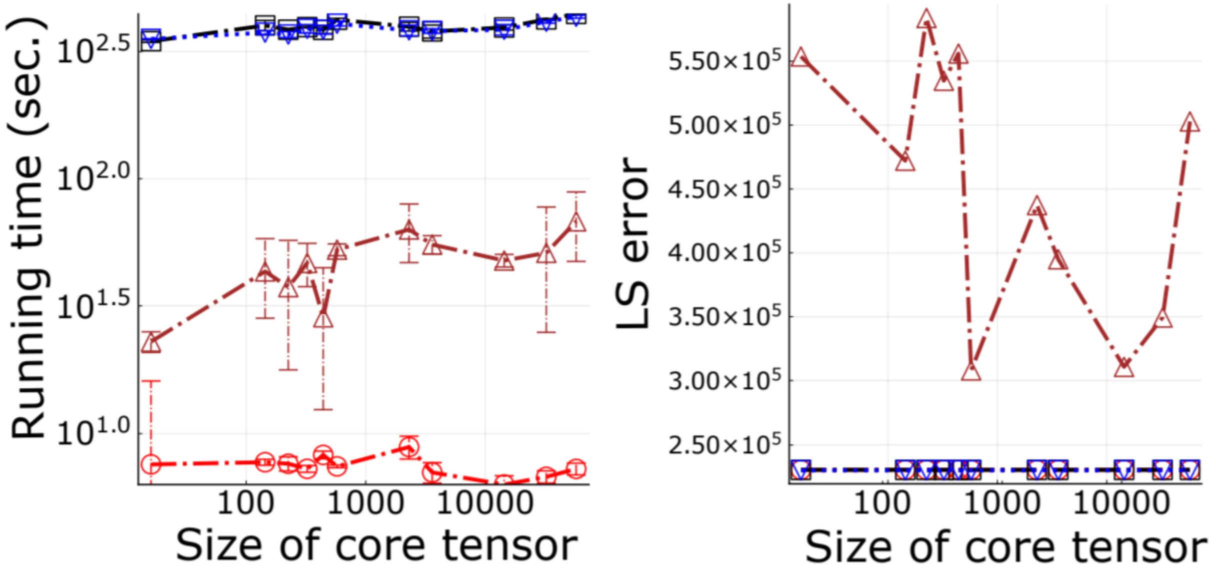
Experimental results (real data)



(c) AttFace (92, 112, 400)



(d) 4DLFD (9, 9, 512, 512, 3)



LTR is faster with the competitive approximation performance.

Conclusion

■ Describe the rank-1 condition using (θ,η)

Rank-1 condition (η-representation)

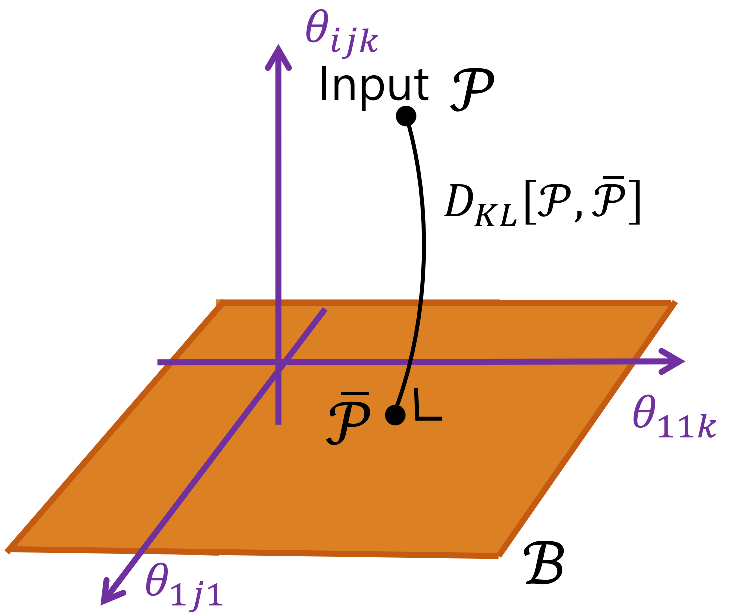
$$\bar{\eta}_{ijk} = \bar{\eta}_{i11}\bar{\eta}_{1j1}\bar{\eta}_{11k}$$

Low-rank condition (θ-representation)

Bingo reduces rank

Rank-1 condition (θ-representation)

All many body $\bar{\theta}_{ijk}$ are 0

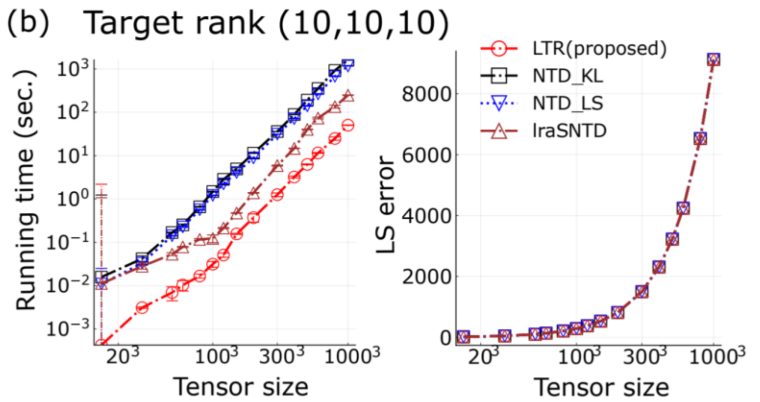


We discuss low-rank approximation as a problem of projection in (θ,η)-space

■ Best rank-1 tensor formula for minimizing KL divergence

$$\bar{\mathcal{P}}_{ijk} = \left(\sum_{j'=1}^{J}\sum_{k'=1}^{K}\mathcal{P}_{ij'k'}\right)\left(\sum_{k'=1}^{K}\sum_{i'=1}^{I}\mathcal{P}_{ij'k'}\right)\left(\sum_{i'=1}^{I}\sum_{j'=1}^{J}\mathcal{P}_{i'j'k}\right)$$

■ Legendre Tucker Rank Reduction (LTR)



- LTR is based on mean-field approximation.
- LTR is faster with the competitive approximation performance as existing methods.
- No need to discuss learning rate, stopping criteria, or initial values