Tensor decompositions: old, new, and beyond

Ryota Tomioka

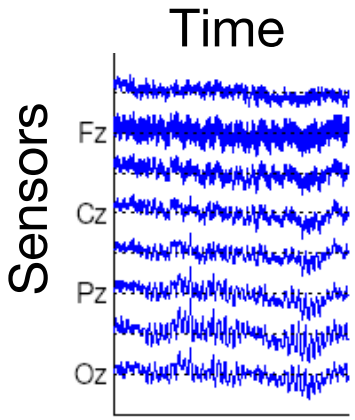
@MLSS, Kyoto 2015

Toyota Technological Institute at Chicago

Tensors appear everywhere

Matrices

Multivariate time-series



Collaborative filtering

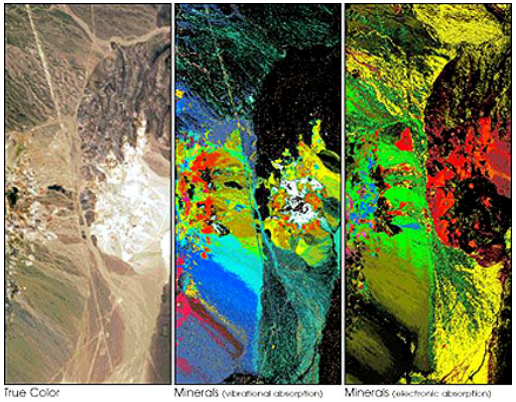
Movies

Users

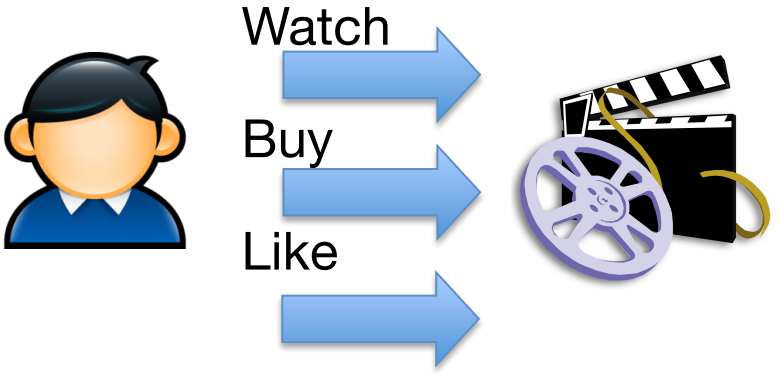
	Star Wars	Titanic	Blade Runner
User 1	5	2	4
User 2	1	4	2
User 3	5	?	?

Tensors

Spatio-temoral data



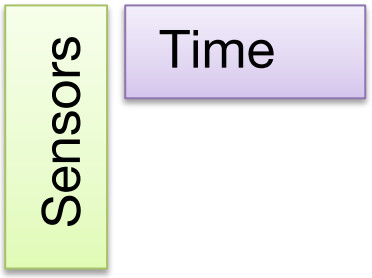
Multiple relations



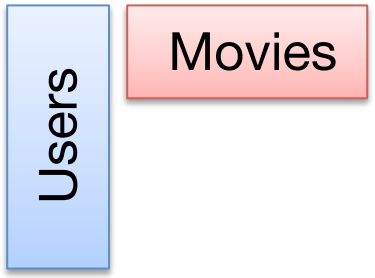
Tensor decomposition

Matrices

Multivariate time-series

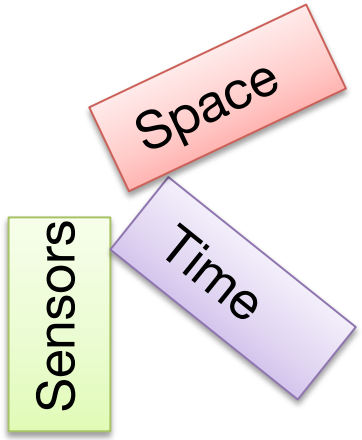


Collaborative filtering

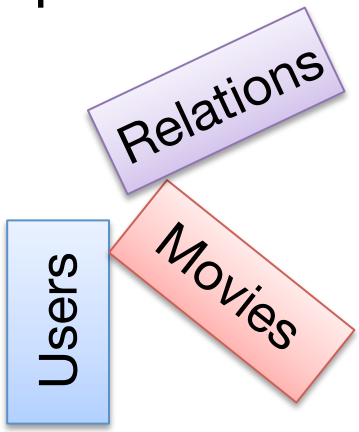


Tensors

Spatio-temoral data

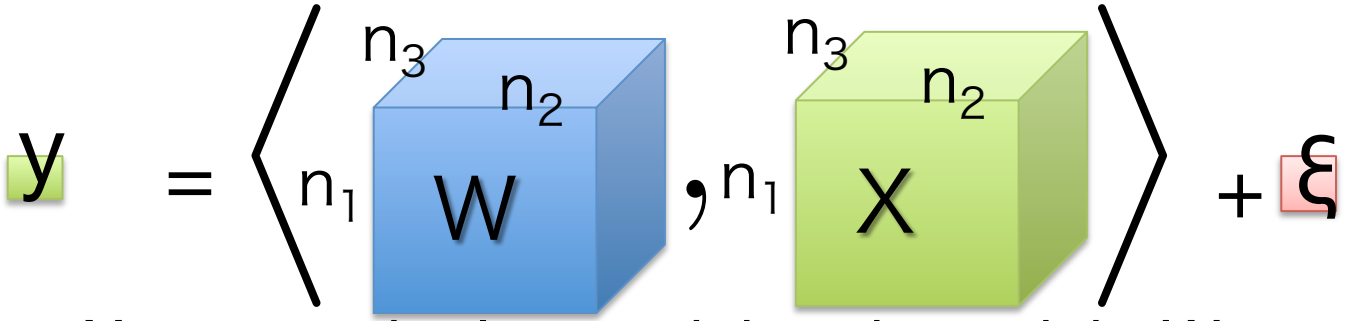


Multiple relations



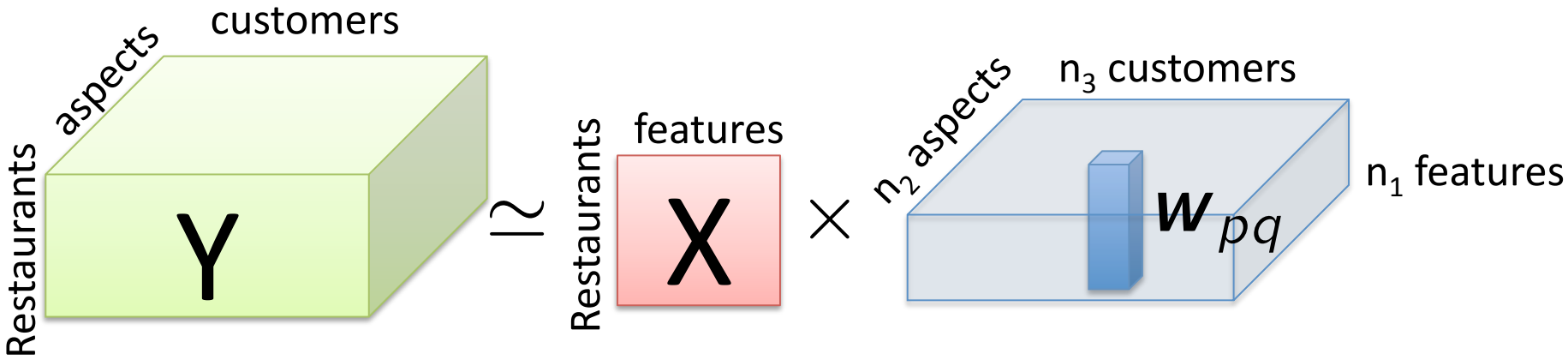
Learning with tensors

- Input: tensor; output: scalar



  - input X may not be low-rank but the weight W may be.

- Multi-task learning with tensors



Learning objectives

Part 1: Matrices

- What is rank? Different views.
- Nuclear norm and factor-based regularization.
- Learning bounds

Part 2: Tensors

- What is rank? Which view generalizes, and how?
- Nuclear norm? Factor-based regularization?

Part 3: Beyond

- Convex relaxation for Tucker model
- Interaction between comptutation and statistics

Lecture format

- I will try to be as interactive as possible.
  - Ask questions when you don't understand.
  - Please interrupt.
- I will use iPython Notebooks to give further details.
  - URL: `https://github.com/ryotat/mlss15/tree/master/python`
  - Python environment?
  - Download Enthought Canopy if you don't

Let's start from matrices

- Rank
- Decomposition
- Regularization
- Theory

$$\boldsymbol{A} \in \mathbb{R}^{m \times n}$$

$$(\text{with } m \leq n, \text{ w.l.o.g.})$$

– and see how these generalize (or does not generalize) to tensors

Rank

- What are the ranks of these matrices?

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$$

Rank 1

$$\begin{pmatrix} 1 & 1 & 0 \\ 2 & 1 & 1 \\ 3 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

Rank 2

Rank

- The rank of a matrix A

= min R such that $\boldsymbol{A} = \sum_{r=1}^{R} \boldsymbol{u}_r \boldsymbol{v}_r{}^{\top}$

= max C such that A_C is linearly independent
(A_C is the submatrix of A indexed by C)

= number of non-zero singular values

Singular values/vectors

- are solutions of

$$\boldsymbol{A}\boldsymbol{v} = \sigma\boldsymbol{u}$$

$$\boldsymbol{A}^\top\boldsymbol{u} = \sigma\boldsymbol{v}$$

σ: singular value; (*u*,*v*): singular vectors

Fact: for (σ_j, u_j, v_j) j=1,…,m, as long as $\sigma_j \neq \sigma_k$

$$\boldsymbol{u}_j{}^\top\boldsymbol{u}_k = 0, \quad \text{and} \quad \boldsymbol{v}_j{}^\top\boldsymbol{v}_k$$

Singular values and eigenvalues

- Eigenvalue equation

$$\begin{bmatrix} 0 & \boldsymbol{A} \\ \boldsymbol{A}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \sigma \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix}$$

Symmetric matrix

$$\begin{bmatrix} 0 & \boldsymbol{A} \\ \boldsymbol{A}^\top & 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \boldsymbol{U} & \boldsymbol{U} \\ \boldsymbol{V} & -\boldsymbol{V} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma} & 0 \\ 0 & -\boldsymbol{\Sigma} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^\top & \boldsymbol{V}^\top \\ \boldsymbol{U}^\top & -\boldsymbol{V}^\top \end{bmatrix}$$

Negative eigenvalues correspond to flipping the sign of *V.*

Singular values and eigenvalues

- Eigenvalue equation

$$\begin{bmatrix} 0 & \boldsymbol{A} \\ \boldsymbol{A}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \sigma \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix}$$

Symmetric matrix

$$\begin{bmatrix} 0 & \boldsymbol{A} \\ \boldsymbol{A}^\top & 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \boldsymbol{U} & \boldsymbol{U} \\ \boldsymbol{V} & -\boldsymbol{V} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma} & 0 \\ 0 & -\boldsymbol{\Sigma} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^\top & \boldsymbol{V}^\top \\ \boldsymbol{U}^\top & -\boldsymbol{V}^\top \end{bmatrix}$$

That is, $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$

Algorithmic view

Fixed-point algorithm

1. Initialize u^0 and v^0 randomly
2. For t=0,1,2,…

$$\boldsymbol{u}^{t+1} = \frac{\boldsymbol{A}\boldsymbol{v}^t}{\|\boldsymbol{A}\boldsymbol{v}^t\|},$$

$$\boldsymbol{v}^{t+1} = \frac{\boldsymbol{A}^\top\boldsymbol{u}^t}{\|\boldsymbol{A}^\top\boldsymbol{u}^t\|}$$

Does this converge? – Yes, essentially power iteration for $\begin{bmatrix} 0 & \boldsymbol{A} \\ \boldsymbol{A}^\top & 0 \end{bmatrix}$

Singular value decomposition

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

$$\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_m, \quad \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_m, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \end{pmatrix}$$

Usual convention: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m \geq 0$.

This gives a decomposition

$$\boldsymbol{A} = \sum_{j=1}^{m} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j{}^\top$$

Is this minimal? Yes – keep only $\sigma_j > 0$.

Experiment

- Open USPS.ipynb

Summary: Rank

1. min R such that $\boldsymbol{A} = \sum_{r=1}^{R} \boldsymbol{u}_r \boldsymbol{v}_r{}^{\top}$

2. max C such that A_C is linearly independent
   (A_C is the submatrix of A indexed by C)

3. number of non-zero singular values

Summary: Rank

- Which definition generalizes to tensors?

1. min R such that $\boldsymbol{A} = \sum_{r=1}^{R} \boldsymbol{u}_r \boldsymbol{v}_r{}^{\top}$

Yes - Rank

2. max C such that A_C is linearly independent
   (A_C is the submatrix of A indexed by C)

Yes – Multilinear rank

3. number of non-zero singular values

???

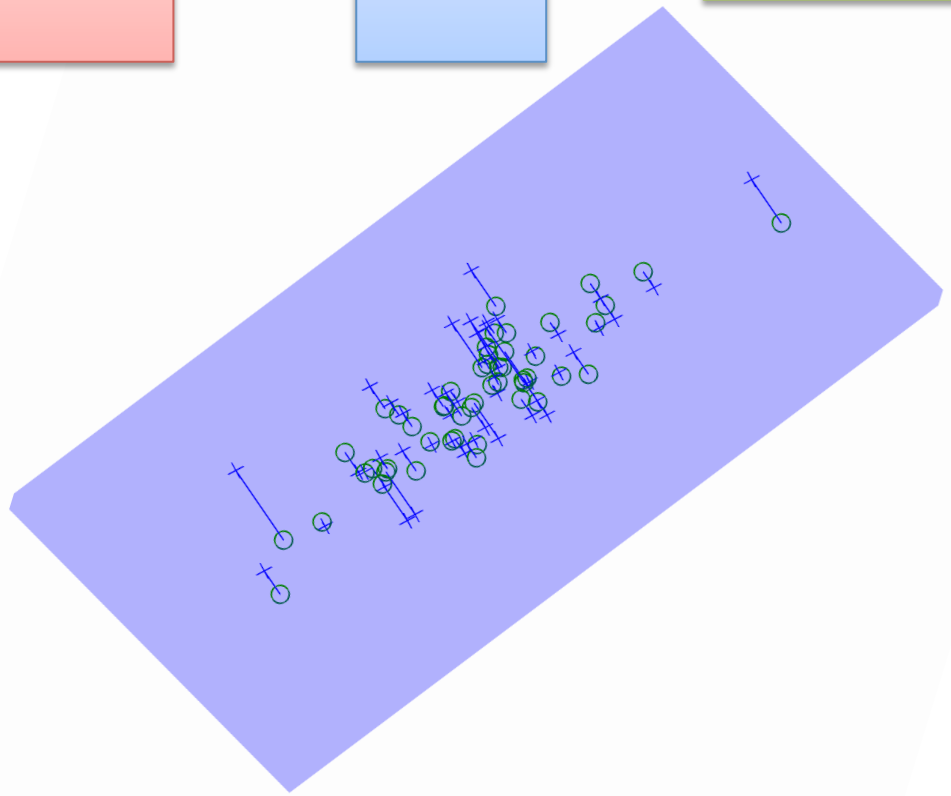
Best rank-*r* approximation

#data points

#dimensions

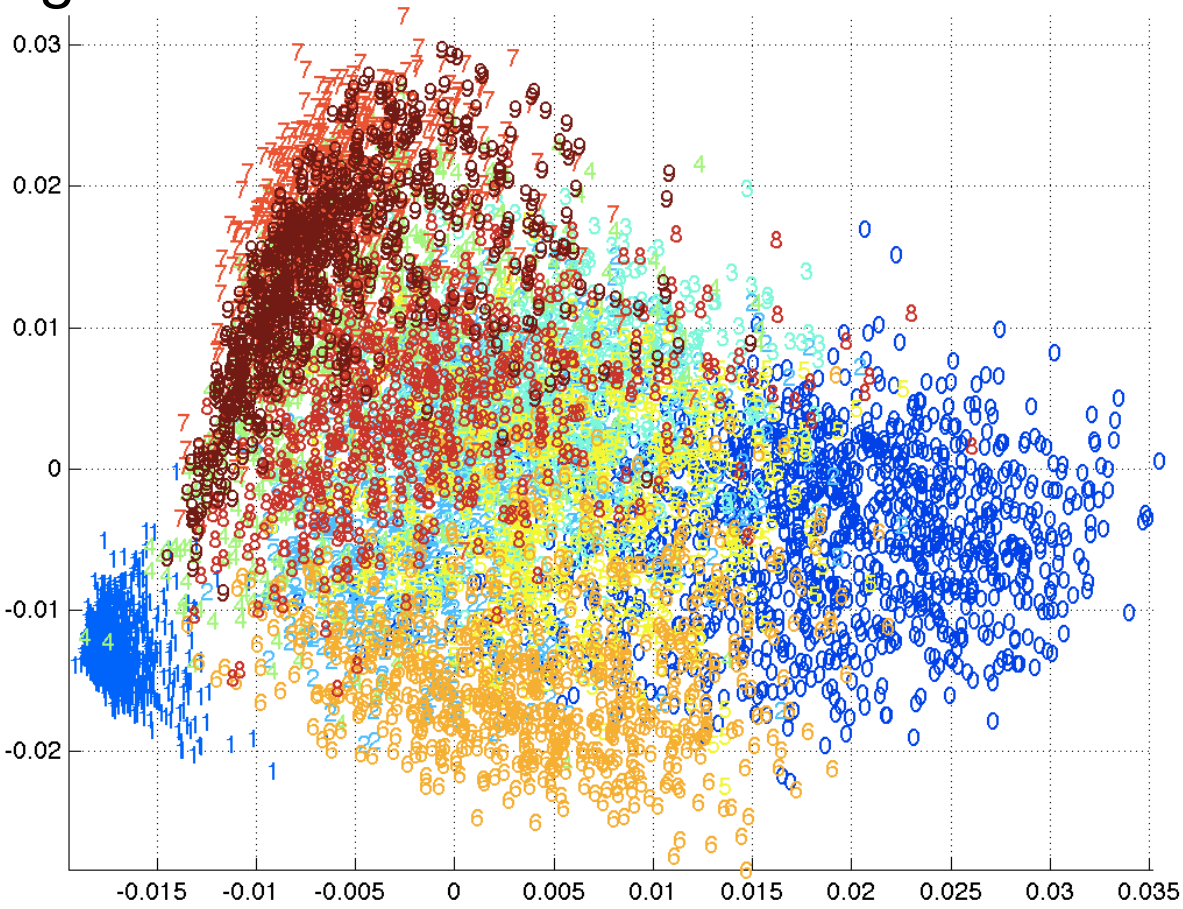
r=#embedding dimensions

$$A \fallingdotseq U \times V^T$$



Best rank-2 approximation

USPS digits dataset



Data from: http://statweb.stanford.edu/~tibs/ElemStatLearn/data.html

Best rank-r approximation

- Given an m x n matrix A, we want to find

$$\hat{\boldsymbol{A}}_r = \underset{\substack{\boldsymbol{X} \in \mathbb{R}^{m \times n}: \\ \operatorname{rank}(\boldsymbol{X}) \leq r}}{\operatorname{argmin}} \|\boldsymbol{A} - \boldsymbol{X}\|_F$$

Best rank-1 approximation

- The simplest case

$$\hat{\boldsymbol{A}}_1 = \underset{\substack{\boldsymbol{X} \in \mathbb{R}^{m \times n}: \\ \mathrm{rank}(\boldsymbol{X}) \leq 1}}{\mathrm{argmin}} \|\boldsymbol{A} - \boldsymbol{X}\|_F$$

Fact: $\hat{\boldsymbol{A}}_1 = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1{}^\top$

σ_1: the largest singular value

Proof

- Let $\boldsymbol{X} = \boldsymbol{u}\boldsymbol{v}^\top$

$$\|\boldsymbol{A} - \boldsymbol{X}\|_F^2 = \|\boldsymbol{A}\|_F^2 - 2\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{v} + \|\boldsymbol{u}\|_2^2 \cdot \|\boldsymbol{v}\|_2^2$$

Setting the derivative to zero, we have

$$\left.\begin{aligned} \boldsymbol{A}\boldsymbol{v} &= \|\boldsymbol{v}\|_2^2 \cdot \boldsymbol{u}, \\ \boldsymbol{A}^\top \boldsymbol{u} &= \|\boldsymbol{u}\|_2^2 \cdot \boldsymbol{v} \end{aligned}\right\}$$

Equations defining singular values/vectors

Solution: $\boldsymbol{u} = \sqrt{\sigma_1}\boldsymbol{u}_1, \quad \boldsymbol{v} = \sqrt{\sigma_1}\boldsymbol{v}_1$

Viewing it as a maximization

- Finding the best rank-one approximation

$$\hat{\boldsymbol{A}}_1 = \underset{\substack{\boldsymbol{X} \in \mathbb{R}^{m \times n}: \\ \operatorname{rank}(\boldsymbol{X}) \leq 1}}{\operatorname{argmin}} \|\boldsymbol{A} - \boldsymbol{X}\|_F$$

is equivalent to finding (*u*,*v*) that maximizes

$$\sigma_1 = \max_{\substack{\boldsymbol{u} \in \mathbb{R}^m, \boldsymbol{v} \in \mathbb{R}^n, \\ \|\boldsymbol{u}\|=1, \|\boldsymbol{v}\|=1}} \boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{v}$$

Note: both problems are non-convex.
Still admit poly-time algorithm (singular value decomp.)

Spectral norm ||A||

- The largest singular value

$$\sigma_1 = \max_{\substack{\boldsymbol{u}\in\mathbb{R}^m, \boldsymbol{v}\in\mathbb{R}^n, \\ \|\boldsymbol{u}\|=1, \|\boldsymbol{v}\|=1}} \boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{v}$$

  is a norm. We denote it by $\|\boldsymbol{A}\| = \sigma_1$

- A norm needs to satisfy
  1) Positive homogeneity $\|\alpha\boldsymbol{A}\| = |\alpha| \cdot \|\boldsymbol{A}\| \quad (\forall \alpha \in \mathbb{R})$
  2) Subadditivity $\|\boldsymbol{A}+\boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$
  3) Separation $\|\boldsymbol{A}\| = 0 \quad \Rightarrow \quad \boldsymbol{A} = \boldsymbol{0}$

Best rank *r* approximation

- For *r*>1, $\hat{\boldsymbol{A}}_r = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r{}^\top$

  (take the first *r* columns of *U* and *V* and corresponding singular values)

PSYCHOMETRIKA—VOL. 1, NO. 3
SEPTEMBER, 1936

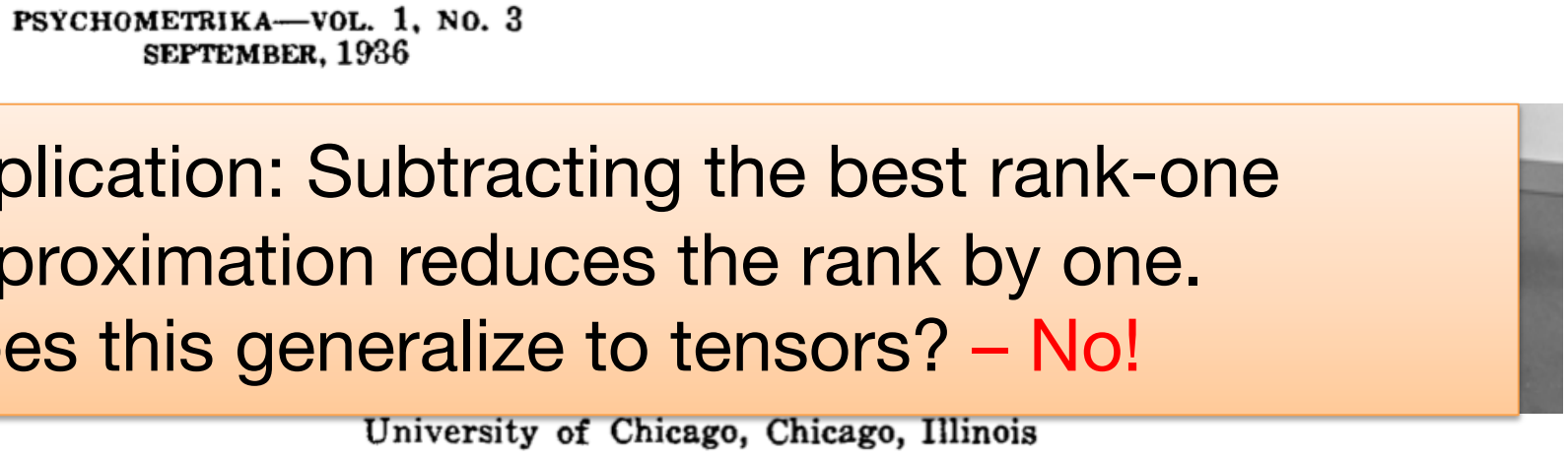
THE APPROXIMATION OF ONE MATRIX BY ANOTHER OF LOWER RANK

CARL ECKART AND GALE YOUNG
University of Chicago, Chicago, Illinois

Best rank *r* approximation

- For $r>1$, $\hat{\boldsymbol{A}}_r = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r{}^\top$

  (take the first *r* columns of *U* and *V* and corresponding singular values)


PSYCHOMETRIKA—VOL. 1, NO. 3
SEPTEMBER, 1936

Implication: Subtracting the best rank-one approximation reduces the rank by one.
Does this generalize to tensors? – No!

University of Chicago, Chicago, Illinois

Low-rank regression

- Input x $\in$ R^d, output y $\in$ R^C

$$\boldsymbol{y} = \boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{\xi}$$

where $$\boldsymbol{W} \in \mathbb{R}^{d \times C},\ \mathrm{rank}(\boldsymbol{W}) \leq r$$

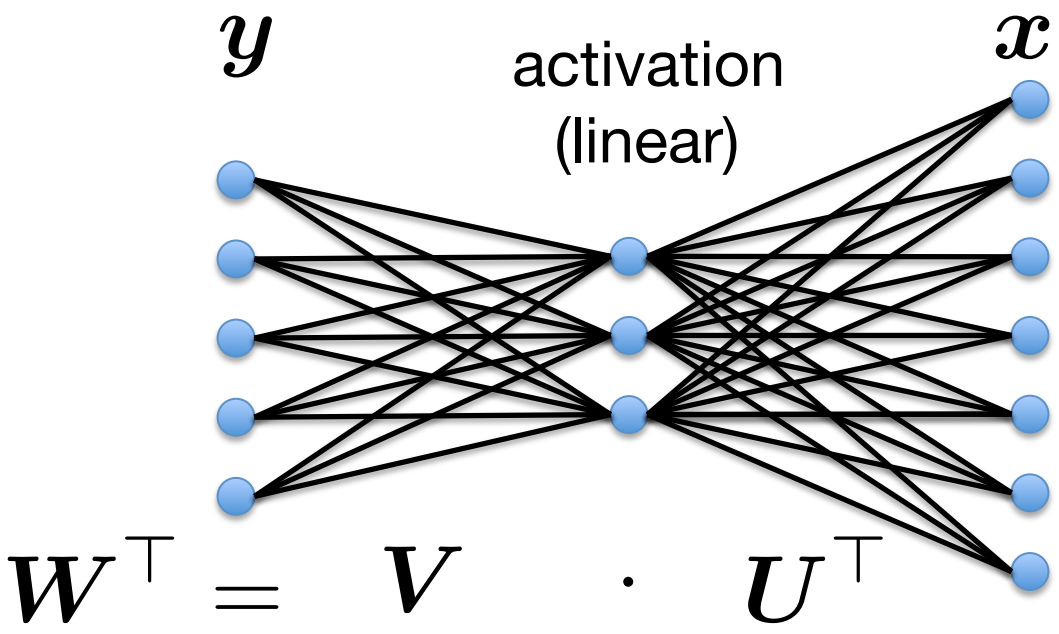
- Given training inputs (x_i,y_i), i=1,…,m

$$\hat{\boldsymbol{W}} = \underset{\substack{\boldsymbol{W} \in \mathbb{R}^{d \times C}: \\ \mathrm{rank}(\boldsymbol{W}) \leq r}}{\mathrm{argmin}} \ \|\boldsymbol{Y} - \boldsymbol{W}^\top \boldsymbol{X}\|_F^2$$

$$\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m], \quad \boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]$$

Connection to linear neural network

- Linear neural network (Baldi & Hornik, 1989)



Matrix factorization = Linear neural network

Analytic solution

$$\boldsymbol{W}^\top = \boldsymbol{P}_V \boldsymbol{\Sigma}_{XY}{}^\top \boldsymbol{\Sigma}_{XX}^{-1}$$

$\boldsymbol{P}_V = \boldsymbol{U}_r \boldsymbol{U}_r{}^\top$,

$\boldsymbol{U}_r$: the top r s.v. of $\boldsymbol{\Sigma}_{XY}{}^\top \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY}$

$\boldsymbol{\Sigma}_{XX} = \sum_{i=1}^m \boldsymbol{x}_i \boldsymbol{x}_i{}^\top, \quad \boldsymbol{\Sigma}_{XY} = \sum_{i=1}^m \boldsymbol{x}_i \boldsymbol{y}_i{}^\top$

Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima

PIERRE BALDI AND KURT HORNIK*

University of California, San Diego

(*Received* 18 *May* 1988; *revised and accepted* 16 *August* 1988)

Matrix completion

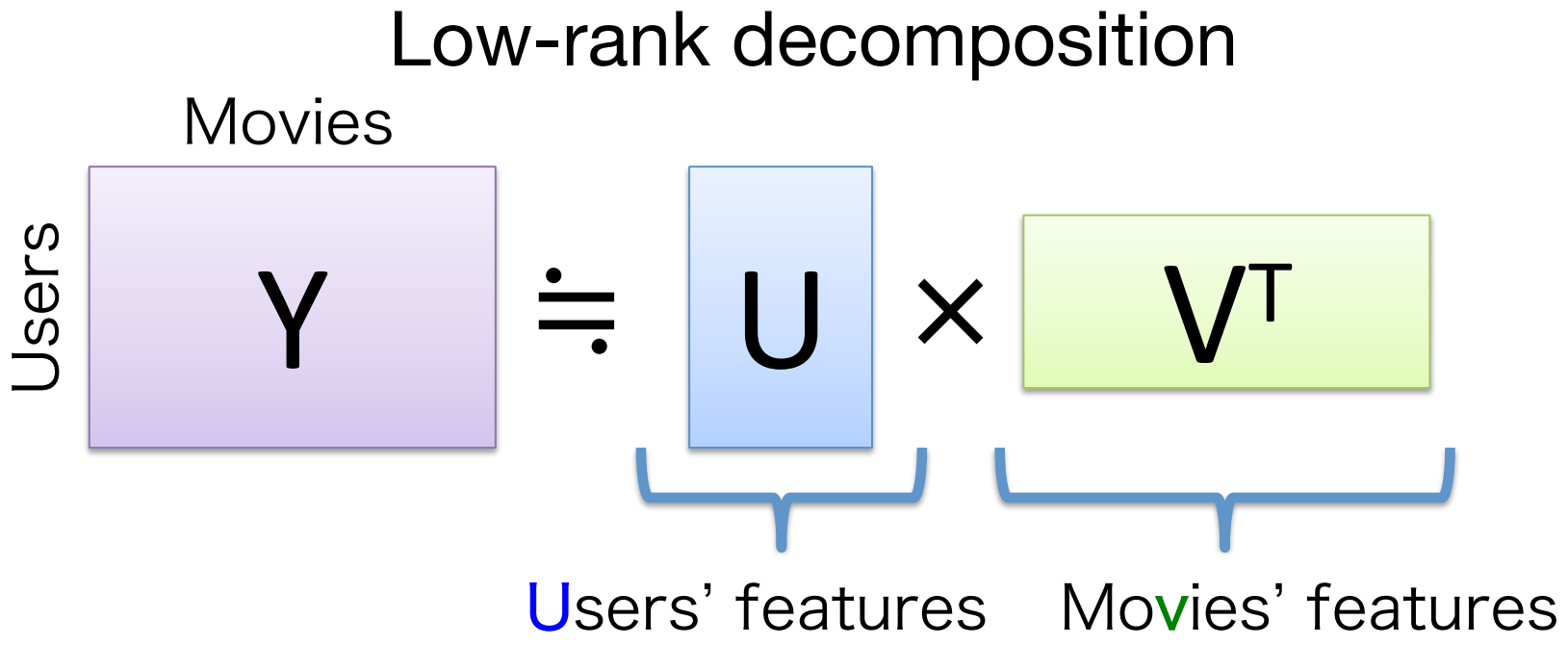
- "Netflix problem" – Fill in the missing entries

$Y =$

	Star wars	Titanic	Blade runner	Lost in translation	
User 1	5	2	4	5	
User 2	1	4	2	4	
User 3	5	1	?	?	
User 4	?	?	?	4	

Matrix completion

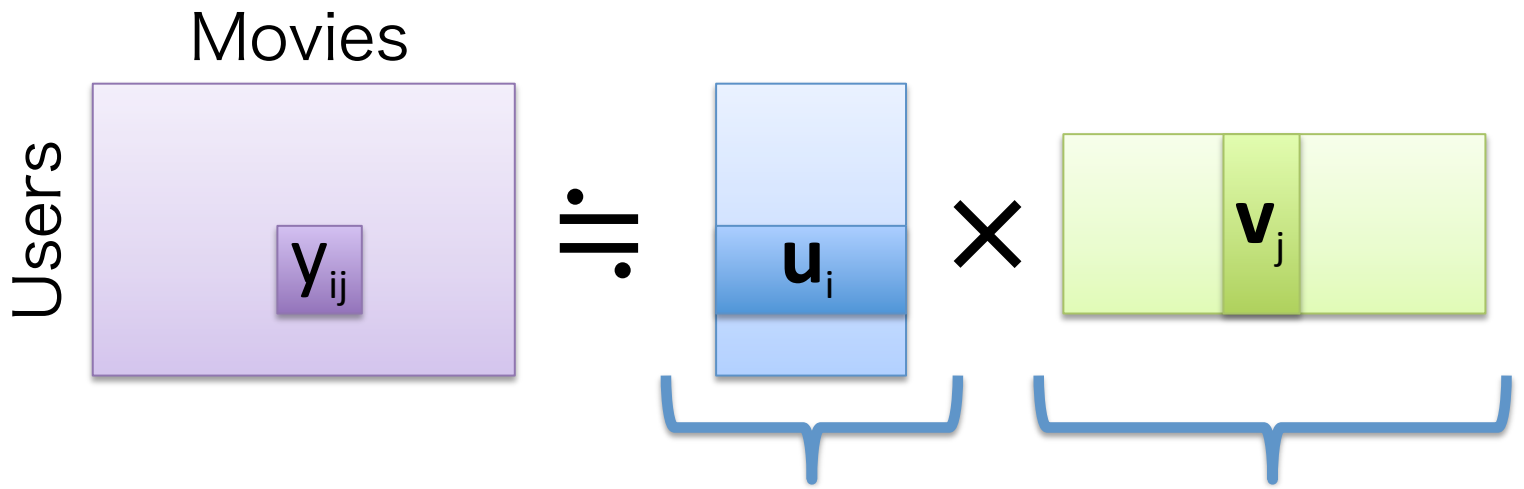
- Impossible without an assumption. (Missing entries can be arbitrary) --- problem is ill-posed
- Most common assumption:



Matrix completion

- Most common assumption:

Low-rank decomposition (rank r)

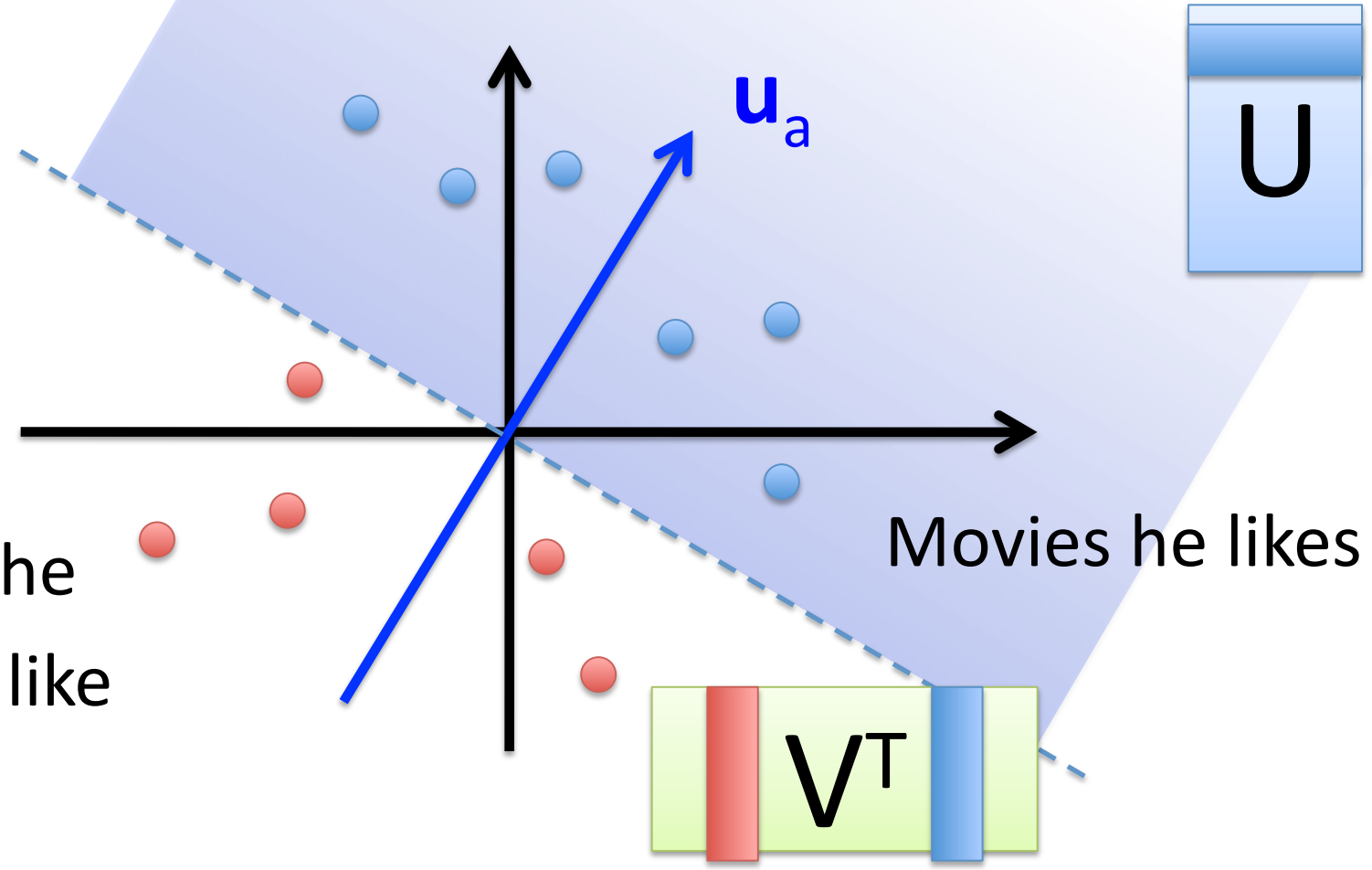


$$y_{ij} = \boldsymbol{u}_i{}^{\top} \boldsymbol{v}_j \quad \begin{pmatrix} \text{dot prodcut} \\ \text{in r-dim space} \end{pmatrix}$$

Geometric Intuition

r-dimensional space

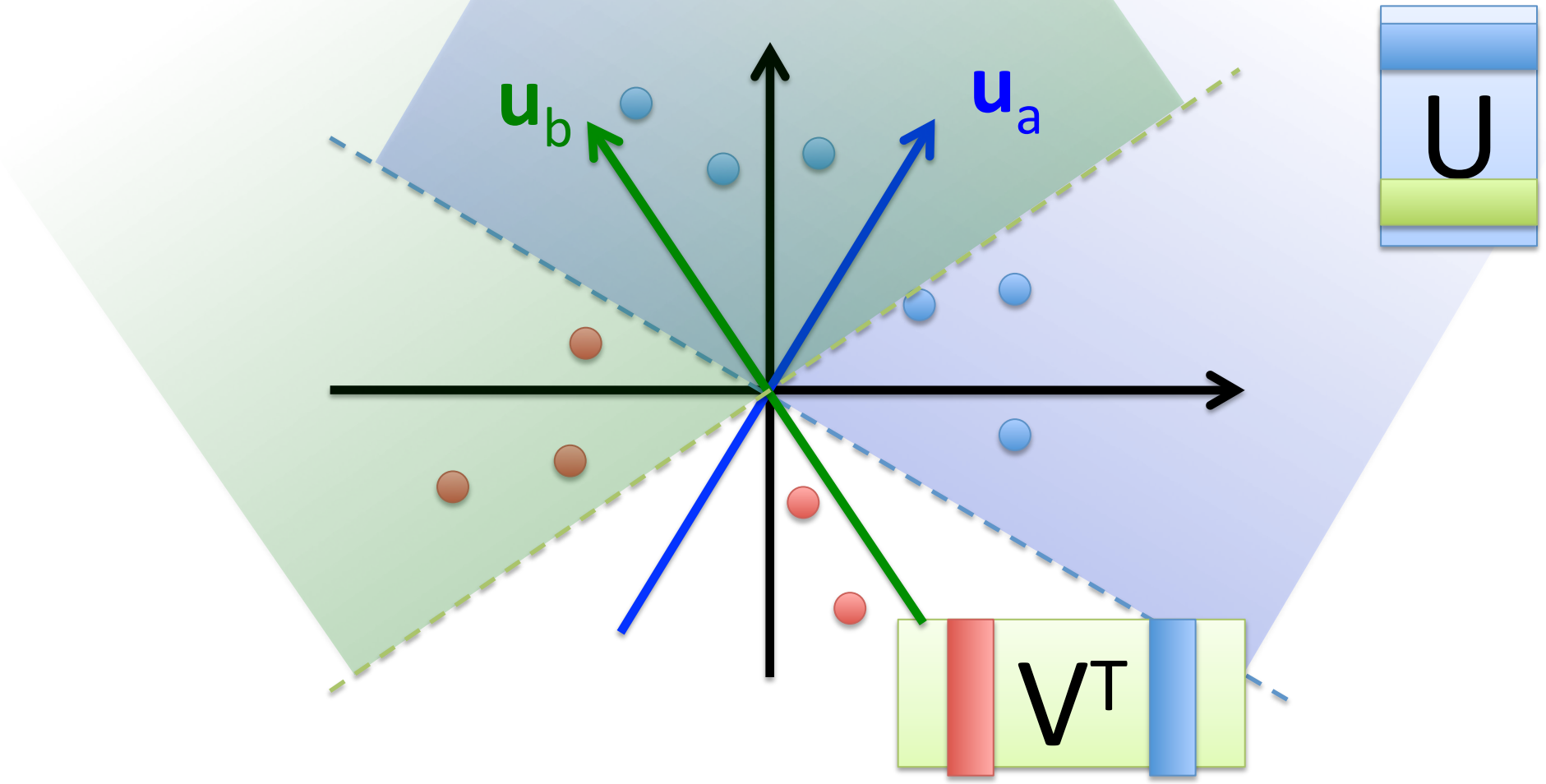
(r: the rank of the decomposition)



Geometric Intuition

r-dimensional space

(r: the rank of the decomposition)



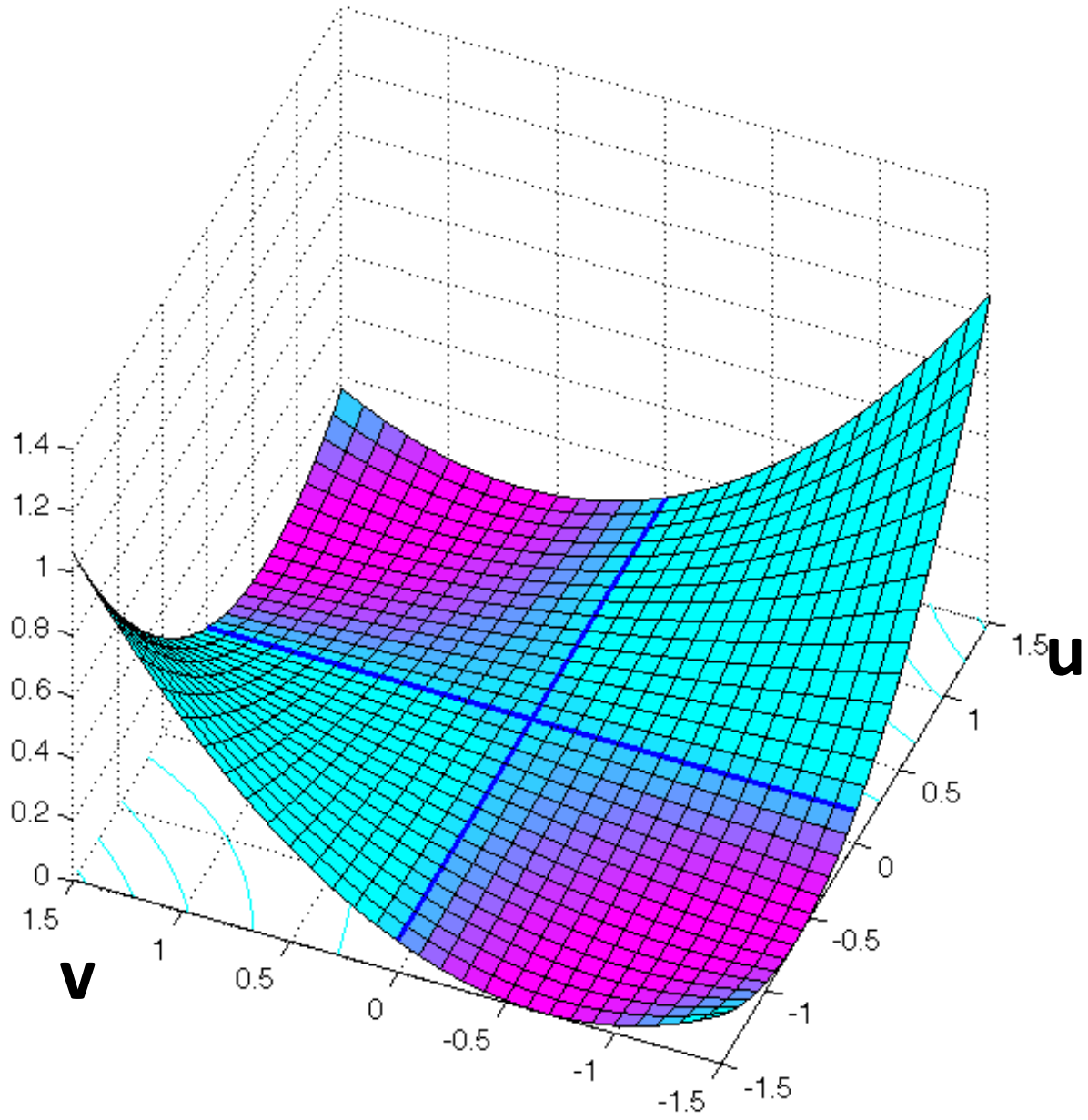
Netflix problem –best rank-*r* approximation from partial observation

$$\underset{\boldsymbol{U}\in\mathbb{R}^{m\times r},\boldsymbol{V}\in\mathbb{R}^{n\times r}}{\text{minimize}} \sum_{(i,j)\in\Omega} \left(Y_{i,j} - \boldsymbol{u}_i{}^\top \boldsymbol{v}_j\right)^2$$

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}_1{}^\top \\ \boldsymbol{u}_2{}^\top \\ \vdots \\ \boldsymbol{u}_m{}^\top \end{bmatrix}, \quad \boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1{}^\top \\ \boldsymbol{v}_2{}^\top \\ \vdots \\ \boldsymbol{v}_n{}^\top \end{bmatrix}$$

Note: u_i and v_j are rows of U and V, respectively.

Non-convexity



Regularization

- Add L2 regularization terms

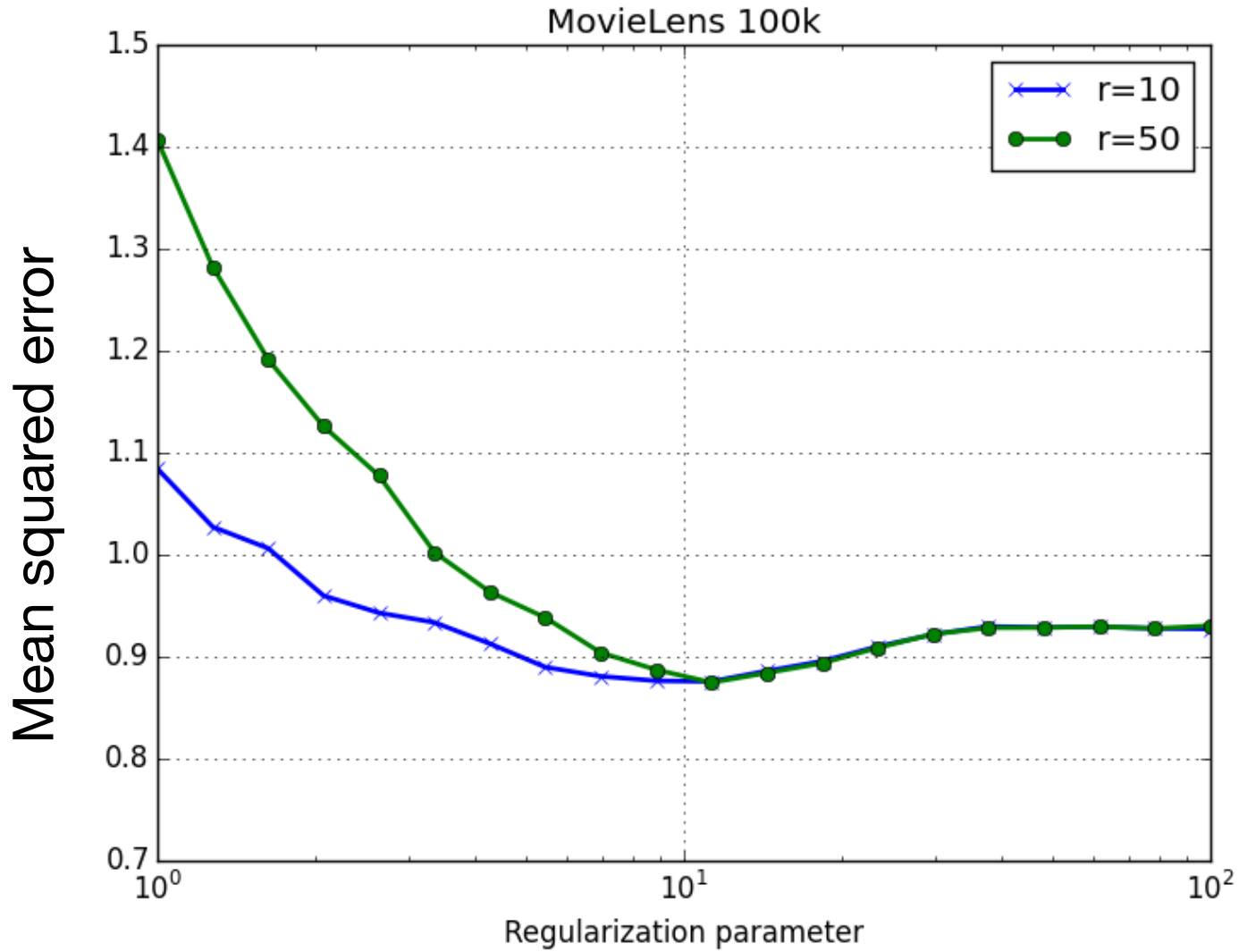
$$\underset{\substack{\boldsymbol{U}\in\mathbb{R}^{m\times r},\\ \boldsymbol{V}\in\mathbb{R}^{n\times r}}}{\text{minimize}} \sum_{(i,j)\in\Omega} \left(Y_{i,j} - \boldsymbol{u}_i{}^{\top}\boldsymbol{v}_j\right)^2 + \frac{\lambda}{2}\left(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2\right)$$

How do we minimize this? –Gradient descent!

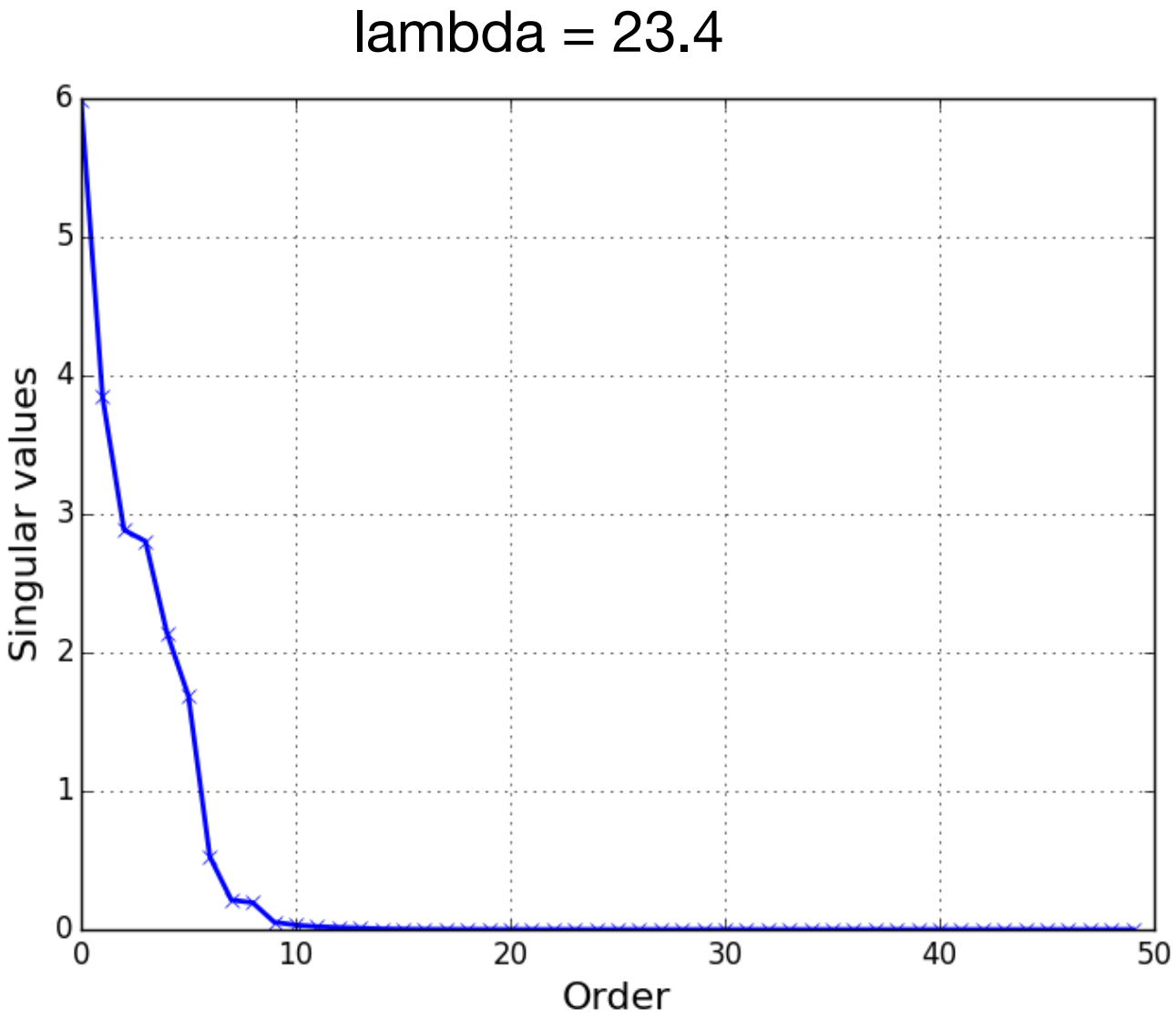
Let's try it with r=10 and r=50

Results for r=10 and r=50

are surprisingly similar



Singular values



What is going on?

- Define $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{V}^\top = (\boldsymbol{u}_i{}^\top \boldsymbol{v}_j)_{i,j}$
- Loss term: only depends on the product *W*.
- Regularizer: let's define

$$g(\boldsymbol{W}) = \min_{\substack{\boldsymbol{U}\in\mathbb{R}^{m\times r}, \boldsymbol{V}\in\mathbb{R}^{n\times r}: \\ \boldsymbol{W}=\boldsymbol{U}\boldsymbol{V}^\top}} \frac{1}{2}\left(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2\right)$$

Note: this is a function of *W* (*U* and *V* are minimized-out)!

Effective regularizer

- Now our objective is

$$\underset{\boldsymbol{W}\in\mathbb{R}^{m\times n}}{\text{minimize}} \sum_{(i,j)\in\Omega} \left(Y_{i,j} - W_{i,j}\right)^2 + \lambda g(\boldsymbol{W})$$

with $$g(\boldsymbol{W}) = \min_{\substack{\boldsymbol{U}\in\mathbb{R}^{m\times r},\boldsymbol{V}\in\mathbb{R}^{n\times r}:\\ \boldsymbol{W}=\boldsymbol{U}\boldsymbol{V}^\top}} \frac{1}{2}\left(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2\right)$$

Is this a convex minimization problem in *W*?

Convexity

Claim: g(w) is a norm (and therefore convex), if *r* is large enough.

Proof:

1. Positive homogeneity:

(for $\boldsymbol{W}' = \alpha \boldsymbol{W}$, take $\boldsymbol{U}' = \text{sign}(\alpha)\sqrt{|\alpha|}\boldsymbol{U}$ and $\boldsymbol{V}' = \sqrt{|\alpha|}\boldsymbol{V}$.)

2. Triangular inequality:

(for $\boldsymbol{W} = \boldsymbol{X} + \boldsymbol{Y}$, take $\boldsymbol{U} = [\boldsymbol{U}_X, \boldsymbol{U}_Y]$ and $\boldsymbol{V} = [\boldsymbol{V}_X, \boldsymbol{V}_Y]$.)

3. Separation:

   g(W) = 0 implies U=V=0, which implies W=0.

Nuclear norm

- The norm $\|\boldsymbol{W}\|_* = \min_{\boldsymbol{W}=\boldsymbol{U}\boldsymbol{V}^\top} \frac{1}{2}\left(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2\right)$

  is called the nuclear norm.

We use the norm notation instead of g(W) because we know that it is a norm.

- Moreover,

$$\|\boldsymbol{W}\|_* = \max_{\substack{\boldsymbol{X}\in\mathbb{R}^{m\times n},\\ \|\boldsymbol{X}\|\leq 1}} \langle \boldsymbol{X}, \boldsymbol{W}\rangle = \sum_{j=1}^{m} \sigma_j$$

Spectral norm

Linear sum of s.v.

Proof

- Easy to see that $\|\boldsymbol{W}\|_* \leq \sum_{j=1}^{m} \sigma_j \leq \max_{\|\boldsymbol{X}\|\leq 1} \langle \boldsymbol{X}, \boldsymbol{W} \rangle$

  (given $\boldsymbol{W} = \tilde{\boldsymbol{U}}\boldsymbol{\Sigma}\tilde{\boldsymbol{V}}^\top$, take $\boldsymbol{U} = \tilde{\boldsymbol{U}}\sqrt{\boldsymbol{\Sigma}}$, $\boldsymbol{V} = \tilde{\boldsymbol{V}}\sqrt{\boldsymbol{\Sigma}}$, and $\boldsymbol{X} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{V}}^\top$.)

- For any $\boldsymbol{W} = \sum_{j=1}^{m} \boldsymbol{u}_j \boldsymbol{v}_j{}^\top$

$$\langle \boldsymbol{X}, \boldsymbol{W} \rangle = \sum_{j=1}^{m} \boldsymbol{u}_j{}^\top \boldsymbol{X} \boldsymbol{v}_j$$

$$\leq \sum_{j=1}^{m} \|\boldsymbol{u}_j\|_2 \cdot \|\boldsymbol{v}_j\|_2 \qquad \text{(because } \|X\| \leq 1\text{)}$$

$$\leq \frac{1}{2} \sum_{j=1}^{m} \left(\|\boldsymbol{u}_j\|_2^2 + \|\boldsymbol{v}_j\|_2^2 \right) \qquad \text{(AM-GM)}$$

Nuclear norm and spectral norm

- Nuclear norm and spectral norm are dual

$$\|\boldsymbol{W}\|_* = \sum_{j=1}^{m} \sigma_j, \qquad \|\boldsymbol{X}\| = \max_j \sigma_j$$

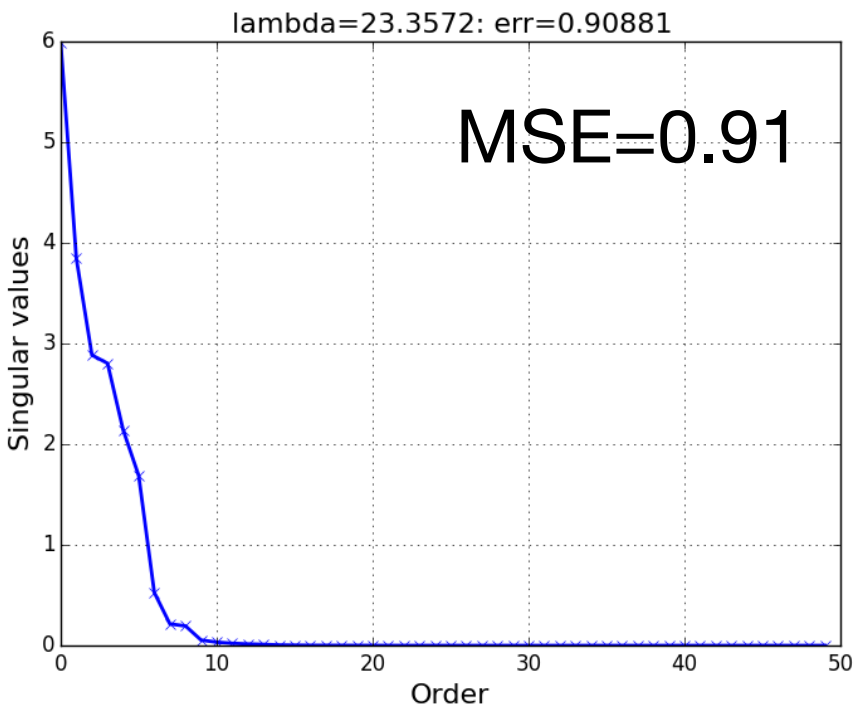
to each other. Just like L_1 and $L\infty$

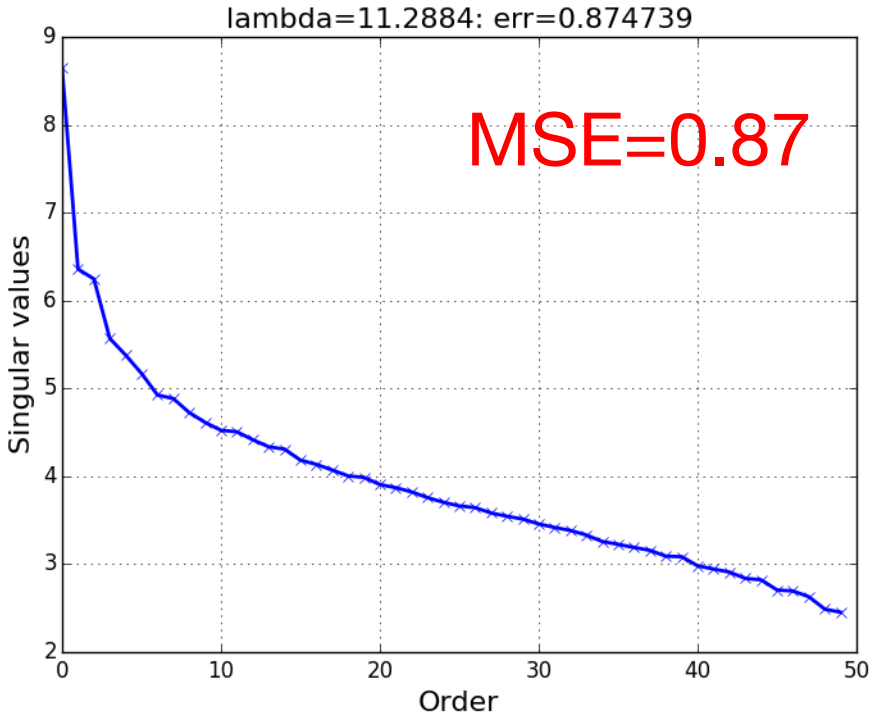
$$\|\boldsymbol{w}\|_1 = \sum_{j=1}^{d} |w_j|, \qquad \|\boldsymbol{x}\|_\infty = \max_j |x_j|$$

When do we use the L1 norm? What can we expect from using the nuclear norm?

Nuclear norm and sparisty

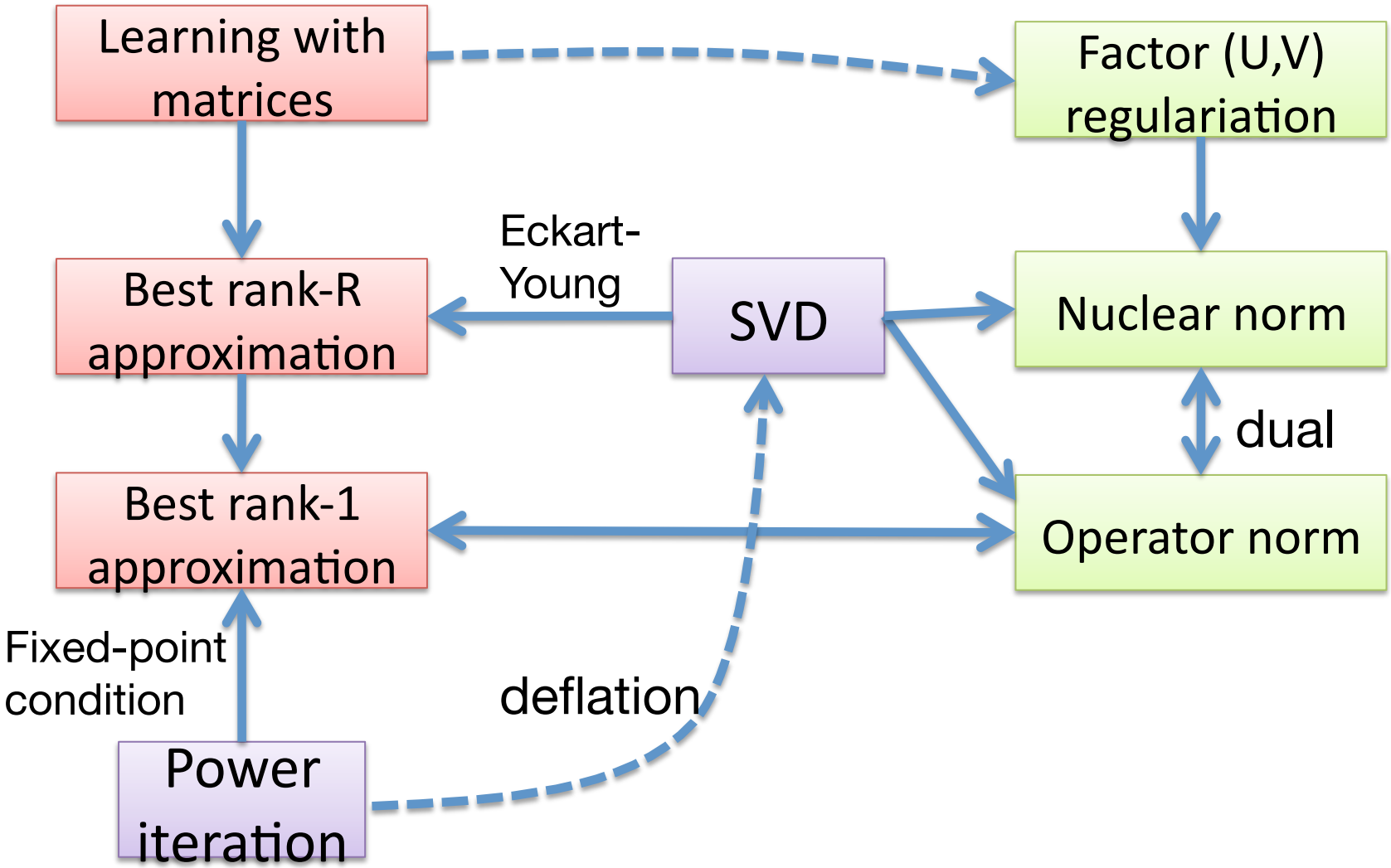
- Nuclear norm promotes *W* to be low rank





but maybe that's not the end of the story…

Summary for matrix decomposition



Little bit of theory

VC dimension

Rademacher complexity

Complexity of low-rank matrices

- Let $f_{\boldsymbol{U},\boldsymbol{V},b}(\boldsymbol{X}) = \left\langle \boldsymbol{X}, \boldsymbol{U}\boldsymbol{V}^\top \right\rangle + b$ and define

$$\mathcal{F}_r = \left\{ f_{\boldsymbol{U},\boldsymbol{V},b} : \boldsymbol{U} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{V} \in \mathbb{R}^{n_2 \times r}, b \in \mathbb{R} \right\}$$

- Examples
  - Low-rank regression: $\boldsymbol{X} = \boldsymbol{x} \cdot \boldsymbol{e}_j{}^\top$
  - Matrix completion: $\boldsymbol{X} = \boldsymbol{e}_i \cdot \boldsymbol{e}_j{}^\top$
  - Matrix classification: general X.

What is the complexity of F_r? How many samples do we need?

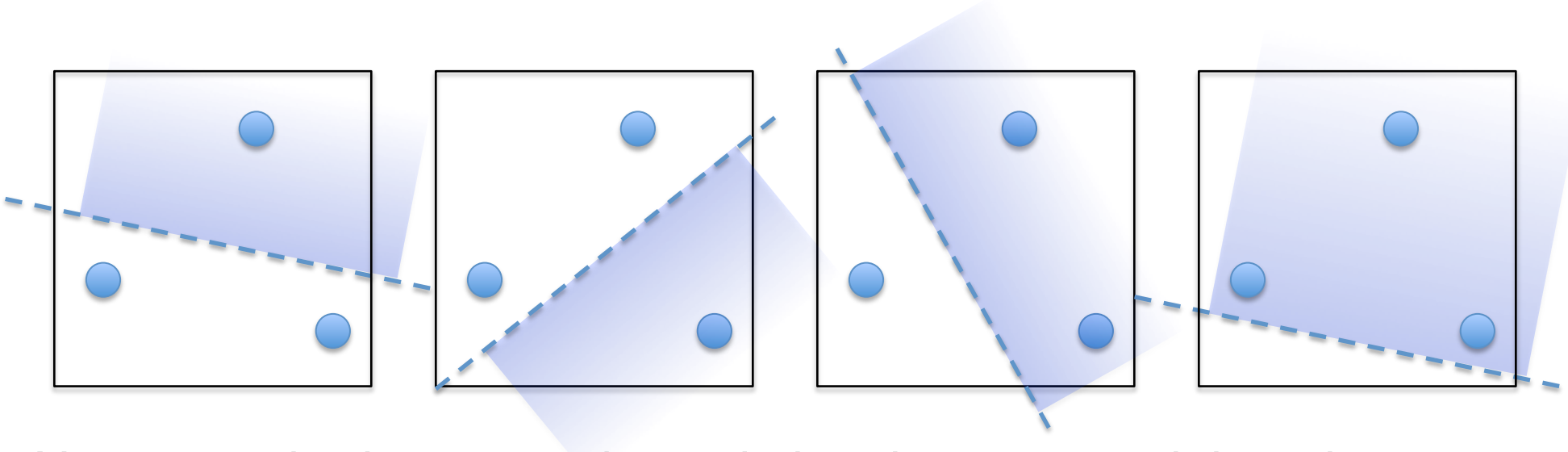
Background

Vapnik

Chervonenkis

- VC dimension of *F*:

$$d(\mathcal{F}) = \max\, m$$
$$\text{s.t. } |\{(\text{sgn}(f(\boldsymbol{X}_1)), \ldots, \text{sgn}(f(\boldsymbol{X}_m)) : f \in \mathcal{F}\}| = 2^m$$



Note: standard asymptotic statistics does not work here because the model is not regular (degenerate Fisher information).

Bound on the VC dimension

$$d(\mathcal{F}_r) \leq (r(n_1 + n_2) + 1) \log(r(n_1 + n_2) + 1)$$

For sufficiently large n_1 and n_2. No more than #parameters!

- This implies that for classification

$$R(f) - R_{\text{emp}}(f) \leq \tilde{O}\left(\sqrt{\frac{(r(n_1 + n_2) + 1)}{m}}\right)$$

for any $f \in \mathcal{F}_r$ with high probability (ignoring the log term)

where $R(f) = \Pr(y \neq f(\boldsymbol{X})), \quad R_{\text{emp}}(f) = \frac{1}{m} \sum_{i=1}^{m} I(y_i \neq f(\boldsymbol{X}_i))$

- Similar bound for real-valued output using pseudodimension.

Proof of the VC dimension

- Warren's theorem

THEOREM 3. *Let $p_1, \ldots, p_m$ be real polynomials in n variables, each of degree at most $d \geqq 1$. If $m \geqq n$, the number of sign sequences* $\operatorname{sgn} \boldsymbol{p}(\boldsymbol{x}) = (\operatorname{sgn} p_1(\boldsymbol{x}), \ldots, \operatorname{sgn} p_m(\boldsymbol{x}))$ *that consist of terms $+1$, -1 does not exceed $(4edm/n)^n$.*

Warren (1968) "Lower bounds for approximation by nonlinear manifolds"

Note that

$$f_{\boldsymbol{U},\boldsymbol{V},b}(\boldsymbol{X}_i) = \left\langle \boldsymbol{X}_i, \boldsymbol{U}\boldsymbol{V}^\top \right\rangle + b$$

is a polynomial of degree at most 2 in $r(n_1+n_2)+1$ variables.

Note: (U,V,b) are variables and X_i are coefficients.

Low-nuclear norm matrices

- Let $f_{\boldsymbol{W}}(\boldsymbol{X}) = \langle \boldsymbol{X}, \boldsymbol{W} \rangle$ and define

$$\mathcal{F}_W = \left\{ f_{\boldsymbol{W}} : \boldsymbol{W} \in \mathbb{R}^{n_1 \times n_2},\ \|\boldsymbol{W}\|_* \leq W \right\}$$

(no bias term for simplicity)

- Examples
  - Low-rank regression: $\boldsymbol{X} = \boldsymbol{x} \cdot \boldsymbol{e}_j{}^\top$
  - Matrix completion: $\boldsymbol{X} = \boldsymbol{e}_i \cdot \boldsymbol{e}_j{}^\top$
  - Matrix classification: general *X*.

Low-nuclear norm matrices

- Let $f_{\boldsymbol{W}}(\boldsymbol{X}) = \langle \boldsymbol{X}, \boldsymbol{W} \rangle$ and define

$$\mathcal{F}_W = \left\{ f_{\boldsymbol{W}} : \boldsymbol{W} \in \mathbb{R}^{n_1 \times n_2},\ \|\boldsymbol{W}\|_* \leq W \right\}$$

(no bias term for simplicity)

- Rademacher complexity

$$\hat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(\boldsymbol{X}_i) \right|$$

(σ_i takes +1 or -1 with probability 1/2)

  - measures how well *F* can fit +1/-1 noise.
  - is scale sensitive.

Rademacher complexity bound

[see Kakade, Shalev-Shwartz, Tewari 12]

If $\|\boldsymbol{X}_i\| \leq X \ (\text{for } i = 1, \ldots, m)$

Spectral norm

$$\hat{\mathcal{R}}_m(\mathcal{F}_W) \leq \sqrt{\frac{2}{m} \log((n_1 + n_2)e)} \cdot X \cdot W.$$

Consequence: for any Lipschitz continuous loss function ℓ ,

$$\mathbb{E}\ell(y, f(\boldsymbol{X})) - \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, f(\boldsymbol{X}_i)) \leq O\left(\sqrt{\frac{X^2 W^2 \log(n_1 + n_2)}{m}}\right)$$

How large are X and W? If $\boldsymbol{X}_i$ (i=1,…,m) are drawn from standard normal distribution, X^2=O(n_1+n_2), and W^2=O(r), if $\boldsymbol{W}$ has a constant Frobenius norm.

Proof of RC bound

- From the duality btwn. the spectral & nuclear norms

$$\frac{1}{m}\sum_{i=1}^{m}\sigma_i \langle \boldsymbol{X}_i, \boldsymbol{W}\rangle \leq \left\|\frac{1}{m}\sum_{i=1}^{m}\sigma_i \boldsymbol{X}_i\right\| \cdot \underbrace{\|\boldsymbol{W}\|_*}_{\leq W}$$

- Thus

$$\hat{\mathcal{R}}_m(\mathcal{F}_W) \leq \mathbb{E}_{\boldsymbol{\sigma}}\left\|\frac{1}{m}\sum_{i=1}^{m}\sigma_i \boldsymbol{X}_i\right\| \cdot W$$

and using random matrix theory,

$$\mathbb{E}_{\boldsymbol{\sigma}}\left\|\frac{1}{m}\sum_{i=1}^{m}\sigma_i \boldsymbol{X}_i\right\| \leq \sqrt{\frac{2}{m}\log((n_1+n_2)e)} \cdot X$$

if $\|\boldsymbol{X}_i\| \leq X$ (for $i = 1, \ldots, m$).

(Using Corollary 4.2 from Tropp (2011) “User-Friendly Tail Bounds for Sums of Random Matrices”)

Tightening by expectation

[Foygel & Srebro, 2011]

- Previous analysis does not assume any distribution for inputs $X_1,\ldots,X_m$.
- Considering the matrix completion setting and taking expectation over the random draw of the observed positions,

$$\mathcal{R}_m(\mathcal{F}_W) := \mathbb{E}\hat{\mathcal{R}}_m(\mathcal{F}_W) = O\left(\sqrt{\frac{(n_1+n_2)\log(n_1+n_2)}{n_1 n_2 m}} \cdot W\right)$$

Expectation over the draw of the observed positions

W can be as large as $(rn_1n_2)^{1/2}$ and sample complexity $O(r(n_1+n_2)/\varepsilon^2)$

Approach from high-dimensional statistics

[Negahban & Wainwright 2011]

- Let the entries of $X_1,\ldots,X_m$ drawn i.i.d. from standard Gaussian distribution and assume

$$y_i = \langle \boldsymbol{X}_i, \underbrace{\boldsymbol{W}^*}_{\text{truth (rank } r\text{)}} \rangle + \xi_i, \quad \xi_i \sim \mathcal{N}(0, \sigma^2), \quad (i = 1, \ldots, m)$$

then the estimator

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W} \in \mathbb{R}^{n_1 \times n_2}}{\operatorname{argmin}} \left(\frac{1}{2m} \sum_{i=1}^{m} (y_i - \langle \boldsymbol{X}_i, \boldsymbol{W} \rangle)^2 + \lambda_m \|\boldsymbol{W}\|_* \right)$$

with $\lambda_m = c \cdot \sigma \sqrt{(n_1 + n_2)/m}$ satisfies

$$\|\hat{\boldsymbol{W}} - \boldsymbol{W}^*\|_F^2 \leq O\left(\frac{\sigma^2 r (n_1 + n_2)}{m} \right)$$

with high probability; r=rank(W*).

Summary so far

- Rank and nuclear norm: both measure complexity of a matrix-based hypothesis class.
- L2 regularization on the factors (U and V) induces nuclear norm regularization on the matrix UV^T.
- Nuclear norm promotes the matrix to be low-rank. However, it can be independently justified even when it doesn't.

Tensors

Rank and multilinear rank

Non-closedness and other weirdness

Tucker decomposition

Graphical representation

Recap: Rank for matrices

1. min R such that $\boldsymbol{A} = \sum_{r=1}^{R} \boldsymbol{u}_r \boldsymbol{v}_r{}^{\top}$

2. max C such that A_C is linearly independent
   (A_C is the submatrix of A indexed by C)

3. number of non-zero singular values

Rank for tensors

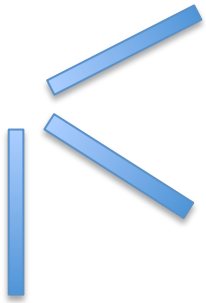
- A *K*th-order tensor

$$\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_K}$$

(we will take *K*=3 below to simplify notation)

- Definition: Rank is the min R such that

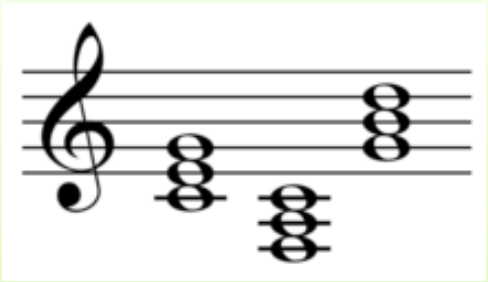
$$\mathcal{A} = \sum_{r=1}^{R} \underbrace{\boldsymbol{u}_r \circ \boldsymbol{v}_r \circ \boldsymbol{w}_r}_{\text{outer product of 3 vectors}}$$



Bibliographical note

- The decomposition

$$\mathcal{A} = \sum_{r=1}^{R} \underbrace{\boldsymbol{u}_r \circ \boldsymbol{v}_r \circ \boldsymbol{w}_r}_{\text{Triad}}$$



Triad

https://en.wikipedia.org/wiki/Triad_(music)

is called CP (Canonical Polyadic) decomposition.

[Hitchcock (1927) “The expression of a tensor or a polyadic as a sum of products”]

- Also known as CANDECOMP [Carroll & Chang, 1970] and PARAFAC [Harshman, 1970].

Rank quiz

- What is the rank of this tensor?

$$\mathcal{A} = \left[\begin{array}{cc|cc} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{array}\right] = \begin{bmatrix}1\\1\end{bmatrix} \circ \begin{bmatrix}1\\1\end{bmatrix} \circ \begin{bmatrix}1\\1\end{bmatrix}$$

Rank 1

1st slice 2nd slice

$$\mathcal{A} = \left[\begin{array}{cc|cc} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 100 \end{array}\right] = \begin{bmatrix}0.0995\\0.995\end{bmatrix} \circ \begin{bmatrix}0.5025\\5.025\end{bmatrix} \circ \begin{bmatrix}1\\10\end{bmatrix} + \begin{bmatrix}0.0995\\-0.995\end{bmatrix} \circ \begin{bmatrix}-0.5025\\5.025\end{bmatrix} \circ \begin{bmatrix}1\\-10\end{bmatrix}$$

Rank 2

$$\mathcal{A} = \left[\begin{array}{cc|cc} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{array}\right]$$

Rank 3

Tensor rank is not closed

- Consider the limit $\alpha \to \infty$ of a rank 2 tensor

$$\mathcal{Y} = \alpha(a_1 + \frac{1}{\alpha}a_2) \circ (b_1 + \frac{1}{\alpha}b_2) \circ (c_1 + \frac{1}{\alpha}c_2) - \alpha a_1 \circ b_1 \circ c_1$$

$$\lim_{\alpha \to \infty} \mathcal{Y} = \underbrace{\boldsymbol{a}_1 \circ \boldsymbol{b}_1 \circ \boldsymbol{c}_2}_{} + \underbrace{\boldsymbol{a}_1 \circ \boldsymbol{b}_2 \circ \boldsymbol{c}_1}_{} + \underbrace{\boldsymbol{a}_2 \circ \boldsymbol{b}_1 \circ \boldsymbol{c}_1}_{}$$

- The limit is rank 3 if $[a_1,a_2]$, $[b_1,b_2]$, $[c_1,c_2]$ are linearly independent.

(Example from Kolda & Bader 2009)

Rank of 2 x 2 x 2 tensors

Table 1
Orbits of $2 \times 2 \times 2$ tensors under the action of invertible multilinear transformation $(\mathbf{S}, \mathbf{T}, \mathbf{U})$ over the real field. The letters D and G stand for "degenerate" (zero volume set in the eight-dimensional space of $2 \times 2 \times 2$ tensors) and "typical" (positive volume set), respectively.

Canonical form	Tensor rank	Multilinear rank	Sign Δ
$D_0: \left[\begin{array}{cc\|cc} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$	0	(0,0,0)	0
$D_1: \left[\begin{array}{cc\|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$	1	(1,1,1)	0
$D_2: \left[\begin{array}{cc\|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right]$	2	(2,2,1)	0
$D_2': \left[\begin{array}{cc\|cc} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right]$	2	(1,2,2)	0
$D_2'': \left[\begin{array}{cc\|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right]$	2	(2,1,2)	0
$G_2: \left[\begin{array}{cc\|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]$	2	(2,2,2)	+
$D_3: \left[\begin{array}{cc\|cc} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{array}\right]$	3	(2,2,2)	0
$G_3: \left[\begin{array}{cc\|cc} -1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{array}\right]$	3	(2,2,2)	−

Stegeman & Comon (2010) "Subtracting a best rank-1 approximation may increase tensor rank"

Bounding the rank

- A trivial bound

$$\operatorname{rank}(\mathcal{A}) \leq n_1 n_2 n_3$$

(consider each entry as a rank-one tensor)

- A slightly better bound

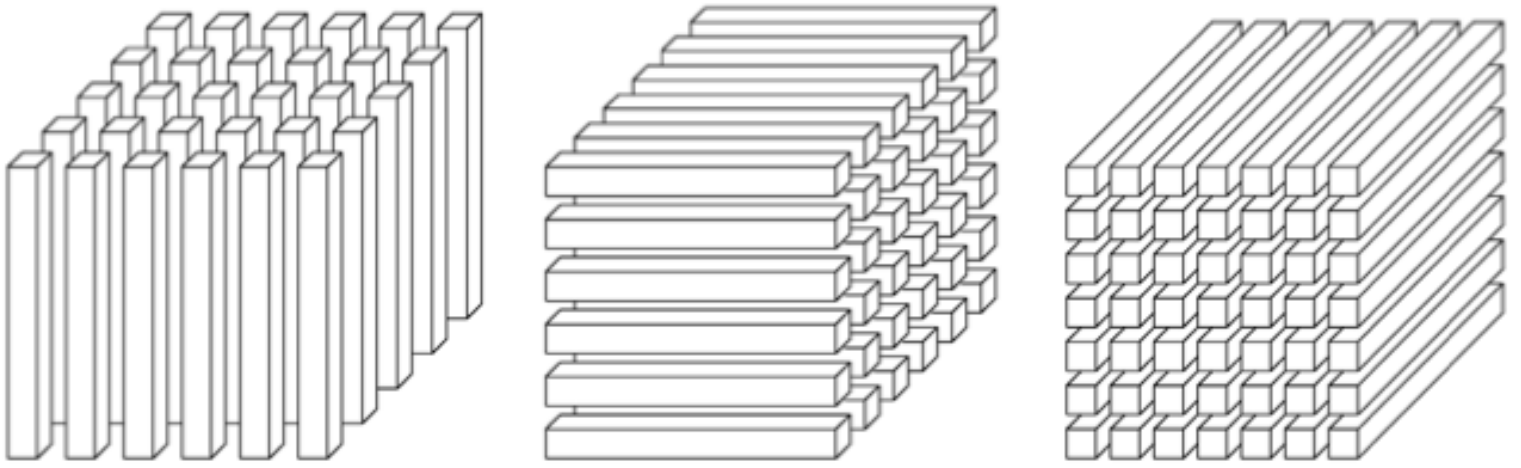
$$\operatorname{rank}(\mathcal{A}) \leq \min(n_1 n_2, n_2 n_3, n_1 n_3)$$

(consider each fiber as a rank-one tensor)

Rank = number of linearly independent columns?

- What is a column vector of a tensor? –fiber

A mode-*k* fiber is a column vector obtained by fixing all but the *k*th index.



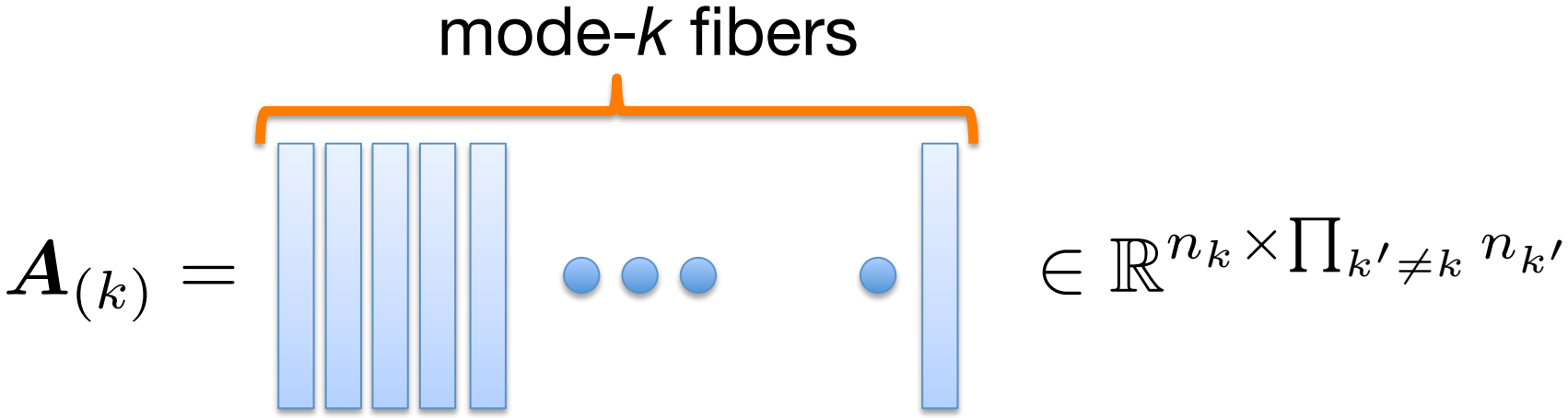
(a) Mode-1 (column) fibers: $\mathbf{x}_{:jk}$ (b) Mode-2 (row) fibers: $\mathbf{x}_{i:k}$ (c) Mode-3 (tube) fibers: $\mathbf{x}_{ij:}$

Fig. 2.1 *Fibers of a 3rd-order tensor.*

(Kolda & Bader 2009)

Unfolding and mode-*k* rank

- mode-*k* unfolding



- mode-*k* rank $\quad \mathrm{rank}_k(\mathcal{A}) = \mathrm{rank}(\boldsymbol{A}_{(k)})$

- Multilinear rank $\quad (\mathrm{rank}_1(\mathcal{A}), \ldots, \mathrm{rank}_K(\mathcal{A}))$

Questions

- How do these two notions of rank relate to each other? When are they equal?
- What kind of decomposition does low-multilinear rank imply?

Equivalence when rank=1

Detecting a rank-one tensor is easy!

- Suppose A is rank one

$$\mathcal{A} = \boldsymbol{u} \circ \boldsymbol{v} \circ \boldsymbol{w} = \left(u_i v_j w_k\right)_{i,j,k}$$

(again assuming *K*=3 for simplicity)

Rank 1

Mode-1 unfolding

$$\boldsymbol{A}_{(1)} = \begin{bmatrix} \boldsymbol{u}v_1 w_1 & \boldsymbol{u}v_2 w_1 & \cdots & \boldsymbol{u}v_{n_2} w_{n_3} \end{bmatrix}$$

$$= \boldsymbol{u}(\boldsymbol{v} \otimes \boldsymbol{w})^\top \quad (\otimes\text{: Kronecker product})$$

Similarly $\boldsymbol{A}_{(2)}$ and $\boldsymbol{A}_{(3)}$ are also rank 1.

Thus multilinear rank=(1,1,1). The other way is also true.

General relation

$$\mathrm{rank}_k(\mathcal{A}) \leq \mathrm{rank}(\mathcal{A}) \leq \prod_{k=1}^{K} \mathrm{rank}_k(\mathcal{A})$$

- Left hand side:

$$\boldsymbol{A}_{(1)} = \sum_{r=1}^{R} \boldsymbol{u}_r (\boldsymbol{v}_r \otimes \boldsymbol{w}_r)^\top$$

- Right hand side: given by the so-called Tucker decomposition

Preliminaries

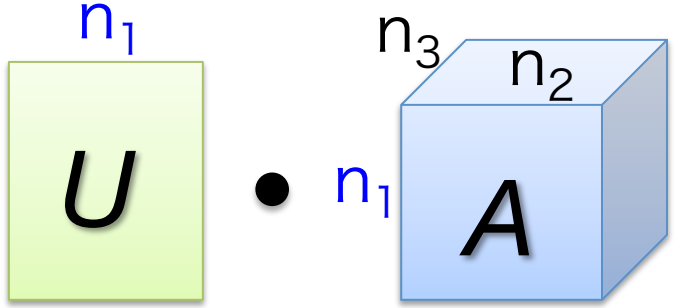
- Define the mode-*k* product

$$\mathcal{A} \times_k \boldsymbol{U} := \left(\sum_{j=1}^{n_k} A_{i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K} U_{i_k j} \right)_{i_1,\ldots,i_K}$$

for $\mathcal{A} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ and $\boldsymbol{U} \in \mathbb{R}^{n'_k \times n_k}$

For example,

$$\mathcal{A} \times_1 \boldsymbol{U} =$$



In other words,

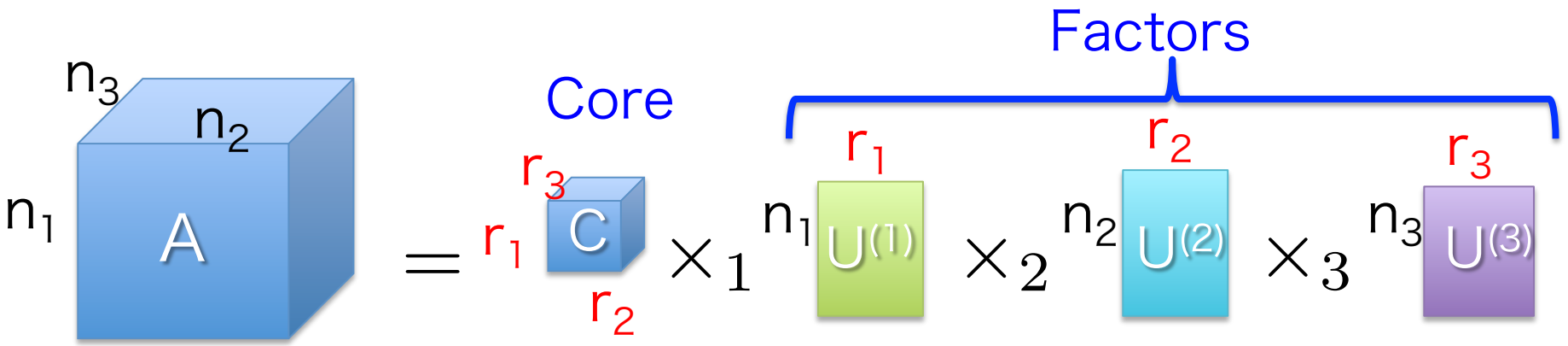
$$(\mathcal{A} \times_k \boldsymbol{U})_{(k)} = \boldsymbol{U} \boldsymbol{A}_{(k)}$$

U applies to each mode-k fiber independently

Tucker decomposition

[Tucker (1966) “Some mathematical notes on three-mode factor analysis”]

- Tensor A has multilinear rank $(r_1,\ldots,r_K)$ implies



$$\left(A_{ijk} = \sum_{a=1}^{r_1} \sum_{b=1}^{r_2} \sum_{c=1}^{r_3} C_{abc} U^{(1)}_{ia} U^{(2)}_{jb} U^{(3)}_{kc} \right)$$

- Also known as higher-order SVD [De Lathauwer+00]

Intuition

- Recall that $(\mathcal{A} \times_k \boldsymbol{U})_{(k)} = \boldsymbol{U}\boldsymbol{A}_{(k)}$
- Take $\boldsymbol{U}^{(k)}$, the (full) left singular vectors of $\boldsymbol{A}_{(k)}$

$$(\mathcal{A} \times_k \boldsymbol{U}^{(k)\top})_{(k)} = \mathsf{r}_k \; \boldsymbol{U}^{(k)\mathsf{T}} \bullet \mathsf{n}_k \; \boldsymbol{A} \qquad \mathsf{n}_k \quad \prod_{k' \neq k} n_{k'}$$

- Repeat

$$\mathcal{A} \times_1 \boldsymbol{U}^{(1)\top} \cdots \times_K \boldsymbol{U}^{(K)\top} = \mathsf{r}_1 \; \mathsf{C} \; \mathsf{r}_2 \; \mathsf{r}_3$$

- Reconstruct

$$\mathcal{A} = \mathcal{C} \times_1 \boldsymbol{U}^{(1)} \times_2 \boldsymbol{U}^{(2)} \cdots \times_K \boldsymbol{U}^{(K)}$$

Properties

- Tensor A has multilinear rank (r_1,…,r_K) implies

$$\mathcal{A} = \sum_{a=1}^{r_1} \sum_{b=1}^{r_2} \sum_{c=1}^{r_3} C_{abc} \boldsymbol{u}_a{}^{(1)} \circ \boldsymbol{u}_b{}^{(2)} \circ \boldsymbol{u}_c{}^{(3)}$$

  Thus

$$\operatorname{rank}(\mathcal{A}) \leq \prod_{k=1}^{K} \operatorname{rank}_k(\mathcal{A})$$

- In fact,

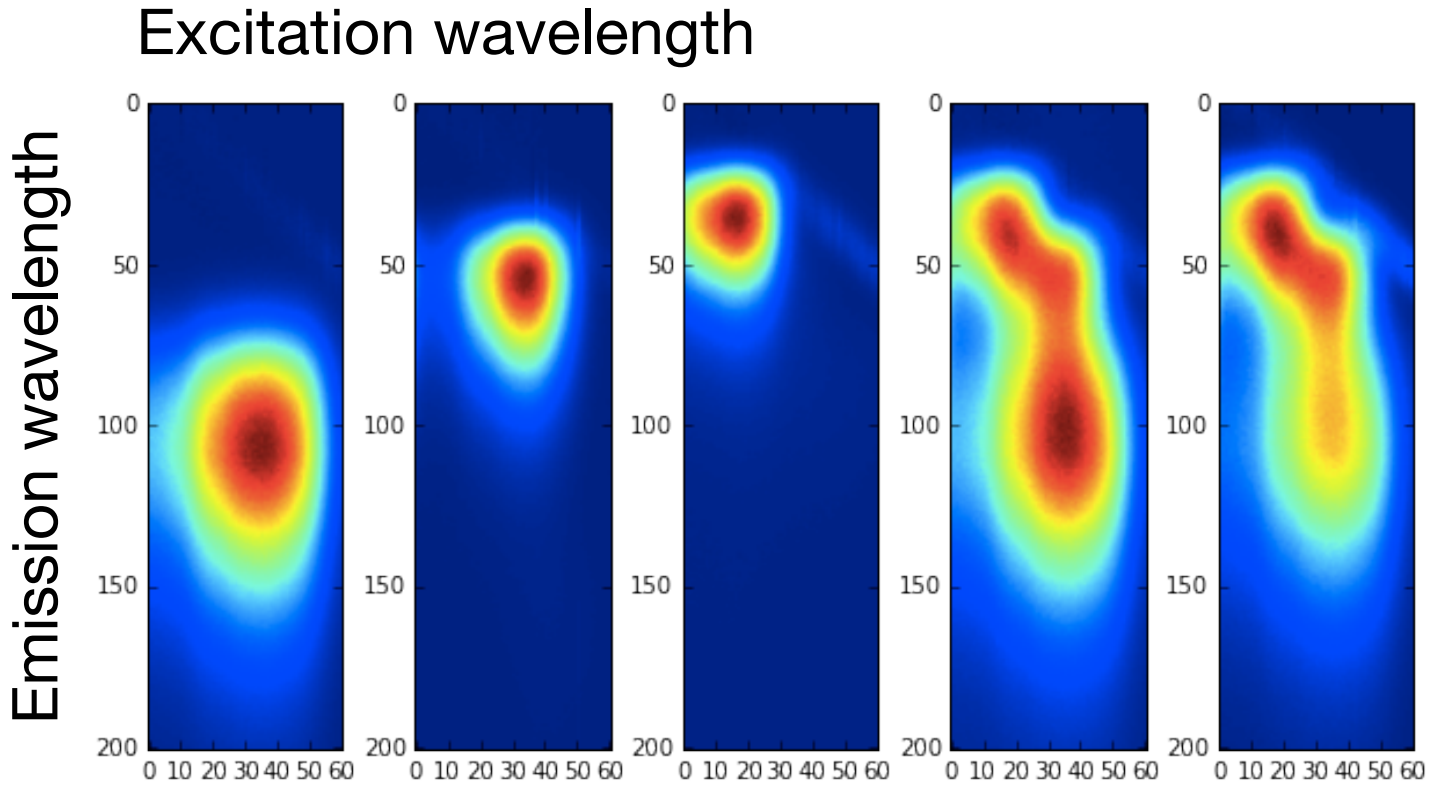
$$\operatorname{rank}(\mathcal{A}) = \operatorname{rank}(\mathcal{C})$$

  if the columns of $\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(K)}$ are independent.

Amino acids fluorescence data

[Bro (1997) “PARAFAC: Tutorial and applications”]

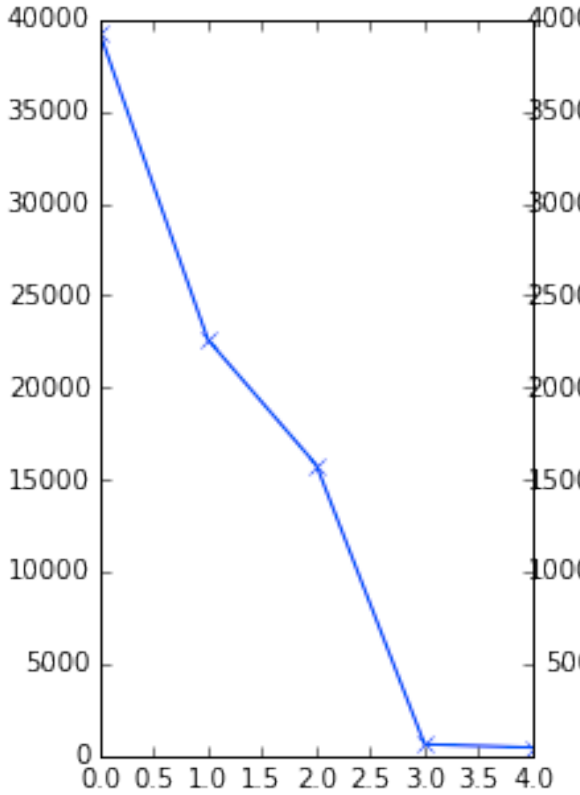
- Five samples containing different amounts of three types of amino acids (Tyr, Trp, and Phe) measured by fluorescence spectrometer



Mode-wise SVD

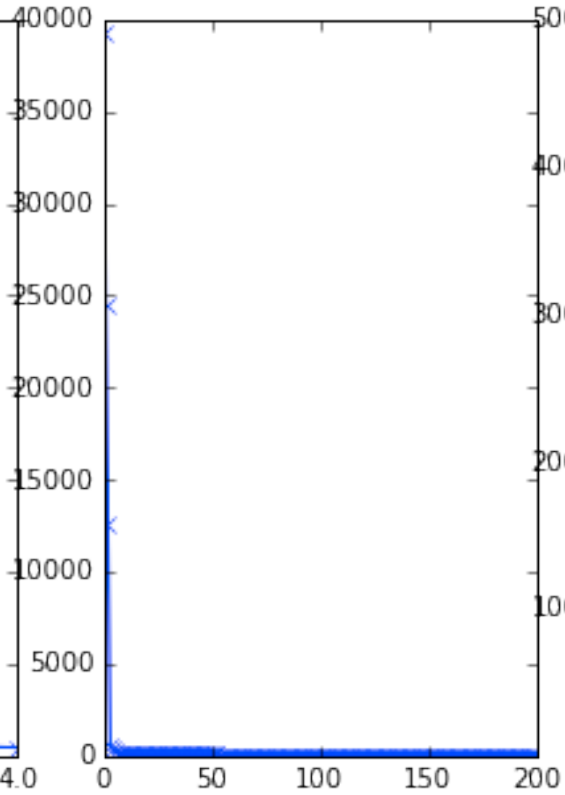
Sample dim.

$SVD(A_{(1)})$



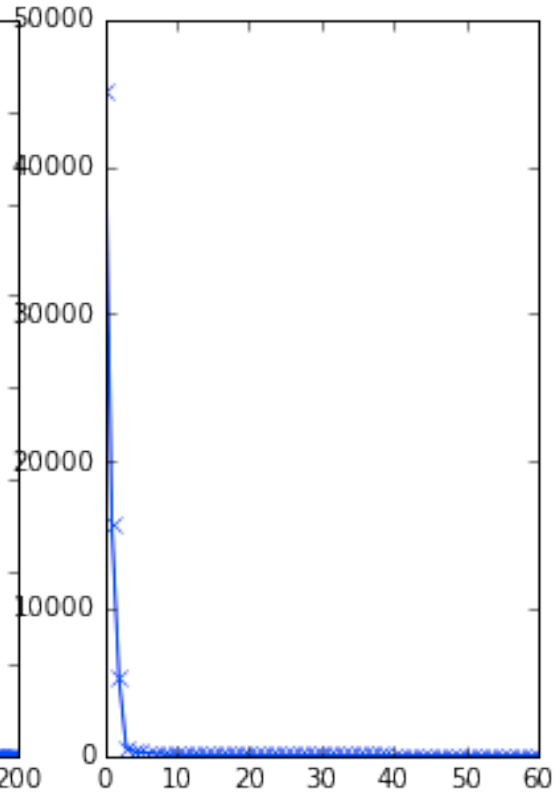
Emission dim.

$SVD(A_{(2)})$



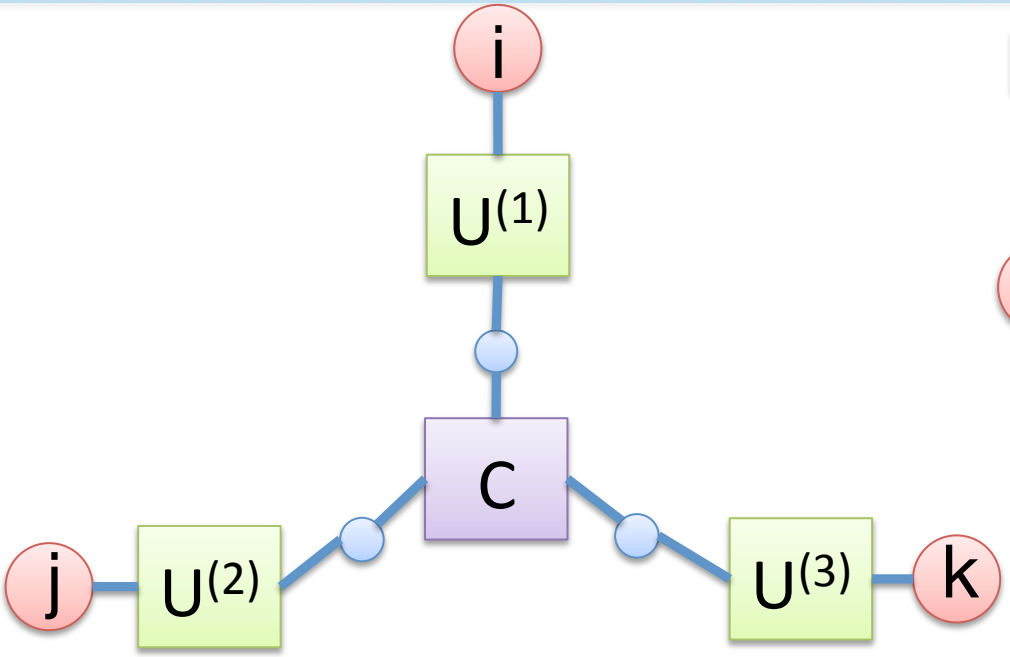
Excitation dim.

$SVD(A_{(3)})$

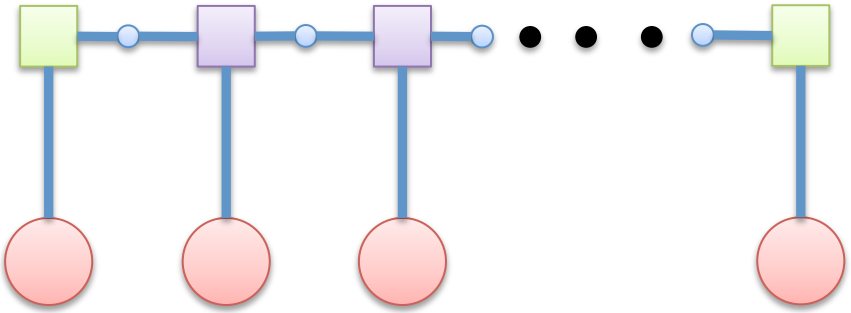


Multilinear rank $\doteqdot (3, 3, 3) \Rightarrow 3 \leq \text{rank}(A) \leq 9$

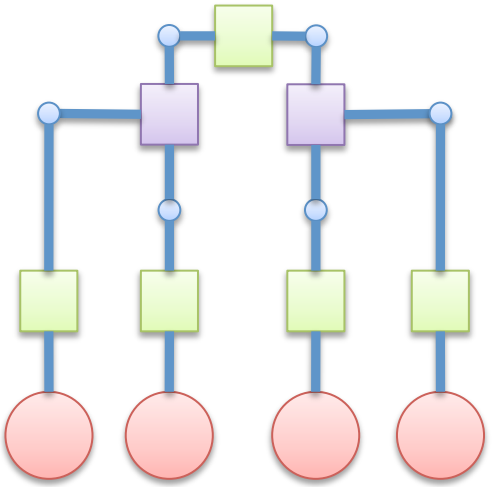
Graphical representation



Tucker decomposition



Tensor train decomposition
[Oseledets, 2011]



Hierarchical Tucker decomposition
[Grasedyck, 2010]

Short summary

- Tensor rank is a much more delicate concept than matrix rank
  - difficulty of computing (NP-hard [Håstad, 90])
  - non-closedness
  - may increase by subtracting the best rank-1 approximation [Stegeman & Comon 2010]
- Rank can be bounded by the multilinear rank, which can be obtained easily by mode-wise SVD.

Tensor decomposition

Best rank-k approximation

Singular values and eigenvalues

Power method

Random projection

Best rank-*R* approximation

- Minimization problem

$$\underset{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}}{\text{minimize}} \left\| \mathcal{A} - \sum_{r=1}^{R} \boldsymbol{u}_r \circ \boldsymbol{v}_r \circ \boldsymbol{w}_r \right\|_F$$

  (Frobenius norm defined as $\|\mathcal{A}\|_F = \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle}$)

- Hardness
  - Known to be NP-hard even for R=1 (3-SAT)
- Questions: for R=1
  - Does it lead to equations defining SVD?
  - Does it relate to the operator norm?

Connection to spectral norm

- Finding

$$(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{w}}) = \underset{\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}}{\operatorname{argmin}} \|\mathcal{A} - \boldsymbol{u} \circ \boldsymbol{v} \circ \boldsymbol{w}\|_F$$

is equivalent to computing

$$\|\mathcal{A}\| = \max_{\substack{\|\tilde{\boldsymbol{u}}\|\leq 1,\\ \|\tilde{\boldsymbol{v}}\|\leq 1,\\ \|\tilde{\boldsymbol{w}}\|\leq 1}} \langle \mathcal{A}, \tilde{\boldsymbol{u}} \circ \tilde{\boldsymbol{v}} \circ \tilde{\boldsymbol{w}} \rangle$$

In fact, $\hat{\boldsymbol{u}} = \sigma\tilde{\boldsymbol{u}}$, $\hat{\boldsymbol{v}} = \tilde{\boldsymbol{v}}$, $\hat{\boldsymbol{w}} = \tilde{\boldsymbol{w}}$ with $\sigma = \|\mathcal{A}\|$ is a best rank-one approximation.

Connection to singular values

- Stationary condition implies

$$\mathcal{A} \times_2 \tilde{\boldsymbol{v}}^\top \times_3 \tilde{\boldsymbol{w}}^\top = \sigma \tilde{\boldsymbol{u}}$$

$$\mathcal{A} \times_1 \tilde{\boldsymbol{u}}^\top \times_3 \tilde{\boldsymbol{w}}^\top = \sigma \tilde{\boldsymbol{v}}$$

$$\mathcal{A} \times_1 \tilde{\boldsymbol{u}}^\top \times_2 \tilde{\boldsymbol{v}}^\top = \sigma \tilde{\boldsymbol{w}}$$

  generalizes equations defining s.v. for matrices.

- We can define any triplet of unit vectors satisfying the above equations singular vectors [Lim, 2005]
- Difference:
  - LHS is quadratic, whereas RHS is linear (motivation for defining L_p singular vectors with $p=K$.)
  - Generally, not orthogonal.

Symmetrization

- Define a super-symmetric n' x n' x n' tensor

(n'=n_1+n_2+n_3)

$$\bar{\mathcal{A}} = \left[\begin{array}{ccc|ccc|ccc} 0 & 0 & 0 & 0 & 0 & \mathcal{A}^{(1,3,2)} & 0 & \mathcal{A}^{(1,2,3)} & 0 \\ 0 & 0 & \mathcal{A}^{(2,3,1)} & 0 & 0 & 0 & \mathcal{A}^{(2,1,3)} & 0 & 0 \\ 0 & \mathcal{A}^{(3,2,1)} & 0 & \mathcal{A}^{(3,1,2)} & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

- Eigenvectors of $\bar{\mathcal{A}}$, solutions of

$$\bar{\mathcal{A}} \times_2 \bar{\boldsymbol{u}}^\top \times_3 \bar{\boldsymbol{u}}^\top = \lambda \bar{\boldsymbol{u}}$$

include vectors of the form $\bar{\boldsymbol{u}} = \begin{bmatrix} \tilde{\boldsymbol{u}} \\ \tilde{\boldsymbol{v}} \\ \tilde{\boldsymbol{w}} \end{bmatrix}$

E.V.P. for symmetric tensors

- Power iteration (detailed below) will quickly converge to one of the solutions –but which solution?
- The solutions are not orthogonal to each other. Cannot use the usual deflation approach (it may actually increase the rank…)
- The orthogonally decomposable case

$$\mathcal{A} = \sum_{r=1}^{R} \lambda_r \boldsymbol{u}_r \circ \boldsymbol{u}_r \circ \boldsymbol{u}_r, \quad \left(\boldsymbol{u}_r \perp \boldsymbol{u}_{r'}, \forall r \neq r'\right)$$

has attracted considerable attention [Anandkumar, Ge, Hsu, Kakade, Telgarsky, 2012]

Power method for symmetric tensors

Fixed-point algorithm

1. Initialize u^0 randomly
2. For t=0,1,2,…

$$\boldsymbol{u}^{t+1} = \frac{\mathcal{A} \times_2 (\boldsymbol{u}^t)^\top \times_3 (\boldsymbol{u}^t)^\top}{\|\mathcal{A} \times_2 (\boldsymbol{u}^t)^\top \times_3 (\boldsymbol{u}^t)^\top\|_2}$$

Why does it work?

- For symmetric orthogonally decomposable tensors,

$$\mathcal{A} = \sum_{r=1}^{R} \lambda_r \boldsymbol{u}_r \circ \boldsymbol{u}_r \circ \boldsymbol{u}_r, \quad (\boldsymbol{u}_r \perp \boldsymbol{u}_{r'}, \forall r \neq r')$$

$$\boldsymbol{u}^{t+1} \propto \mathcal{A} \times_2 (\boldsymbol{u}^t)^\top \times_3 (\boldsymbol{u}^t)^\top$$

$$= \sum_{r=1}^{R} \lambda_r (\langle \boldsymbol{u}_r, \boldsymbol{u}^t \rangle)^2 \boldsymbol{u}_r$$

$$\langle \boldsymbol{u}_r, \boldsymbol{u}^{t+1} \rangle \propto \lambda_r (\langle \boldsymbol{u}_r, \boldsymbol{u}^t \rangle)^2 \qquad \text{(Used orthogonality)}$$

Notes on power iteration for orthogonally decomposable tensors

$$\left\langle \boldsymbol{u}_r, \boldsymbol{u}^{t+1} \right\rangle \propto \lambda_r (\left\langle \boldsymbol{u}_r, \boldsymbol{u}^{t} \right\rangle)^2$$

- The inner product $\left\langle \boldsymbol{u}_r, \boldsymbol{u}^{t} \right\rangle$ will converge to zero exponentially fast except for the one with the largest absolute value.
- The rate is *K*-1: power iteration converges faster for higher order tensors.
- Odd order *K*: can take $\lambda_r \geq 0$. RHS≥0. Even order *K*: RHS can be negative (oscillation).
- Deflation works (because the factors are orthogonal).
- Robustness to noise (see Anandkumar et al. 12)

Power method for asymmetric tensors

1. Initialize u^0, v^0, w^0 randomly
2. For t=0,1,2,...

$$\boldsymbol{u}^{t+1} = \frac{\mathcal{A} \times_2 (\boldsymbol{v}^t)^\top \times_3 (\boldsymbol{w}^t)^\top}{\|\mathcal{A} \times_2 (\boldsymbol{v}^t)^\top \times_3 (\boldsymbol{w}^t)^\top\|_2},$$

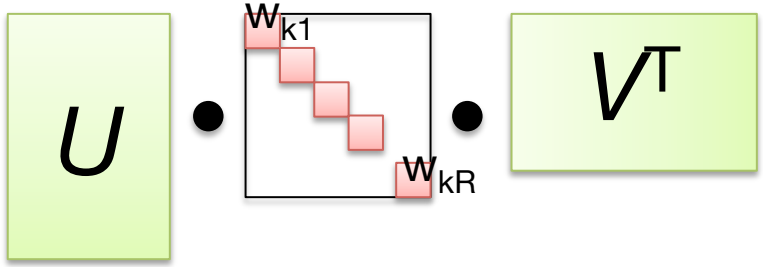
$$\boldsymbol{v}^{t+1} = \frac{\mathcal{A} \times_1 (\boldsymbol{u}^t)^\top \times_3 (\boldsymbol{w}^t)^\top}{\|\mathcal{A} \times_1 (\boldsymbol{u}^t)^\top \times_3 (\boldsymbol{w}^t)^\top\|_2},$$

$$\boldsymbol{w}^{t+1} = \frac{\mathcal{A} \times_1 (\boldsymbol{u}^t)^\top \times_2 (\boldsymbol{v}^t)^\top}{\|\mathcal{A} \times_1 (\boldsymbol{u}^t)^\top \times_2 (\boldsymbol{v}^t)^\top\|_2}$$

Tensor and simultaneous diagonalization

- If $R<\min(n_1,n_2)$, slices are low-rank

$$\boldsymbol{A}_{:,:,k} = \sum_{r=1}^{R} w_{kr}\boldsymbol{u}_r\boldsymbol{v}_r{}^{\top} =$$

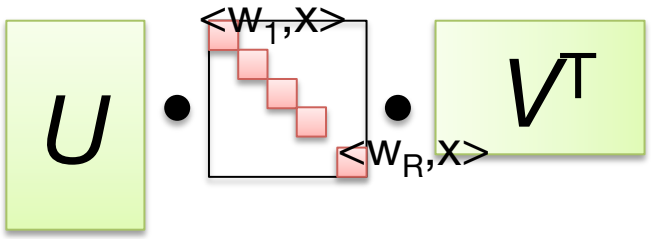


- Factors are common to all slices.
- If we can simultaneously diagonalize all slices, we are done.

Random projection

- Consider the random projection

$$\boldsymbol{A} \times_3 \boldsymbol{x}^\top = \sum_{r=1}^{R} \langle \boldsymbol{w}_r, \boldsymbol{x} \rangle \, \boldsymbol{u}_r \boldsymbol{v}_r{}^\top =$$



x is a random vector. The projection is again a low-rank matrix if $R<\min(n_1,n_2)$.

  - Random slice is a special case of random projection

- How many random projections/slices do we need?

  - Two! (if the factors are linearly independent)

[Harshman (1972) Determination and Proof of Minimum Uniqueness Conditions for PARAFAC1]

Random projection algorithm

[Harshman 1972; Goyal, Vempala, Xiao, 2014; Moitra et al.]

1. Let A_1 and A_2 be two independent random projections of rank-R tensor A. (see Random Projection.ipynb)
2. Compute truncated rank-R SVD $\boldsymbol{A}_1 = \boldsymbol{U}_1 \boldsymbol{S}_1 \boldsymbol{V}_1{}^\top$
3. Compute eigen-decomposition

$$\boldsymbol{S}_1^{-1} \boldsymbol{U}_1{}^\top \boldsymbol{A}_2 \boldsymbol{V}_1 = \boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{P}^{-1}$$

4. Recover

$$\boldsymbol{U} = \boldsymbol{U}_1 \boldsymbol{S}_1 \boldsymbol{P}$$

$$\boldsymbol{V} = \boldsymbol{V}_1 \boldsymbol{P}^{-\top}$$

5. Note that

$$\boldsymbol{A}_1 = \boldsymbol{U} \boldsymbol{V}^\top, \quad \boldsymbol{A}_2 = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{V}^\top$$

Learning with tensors

Regularization

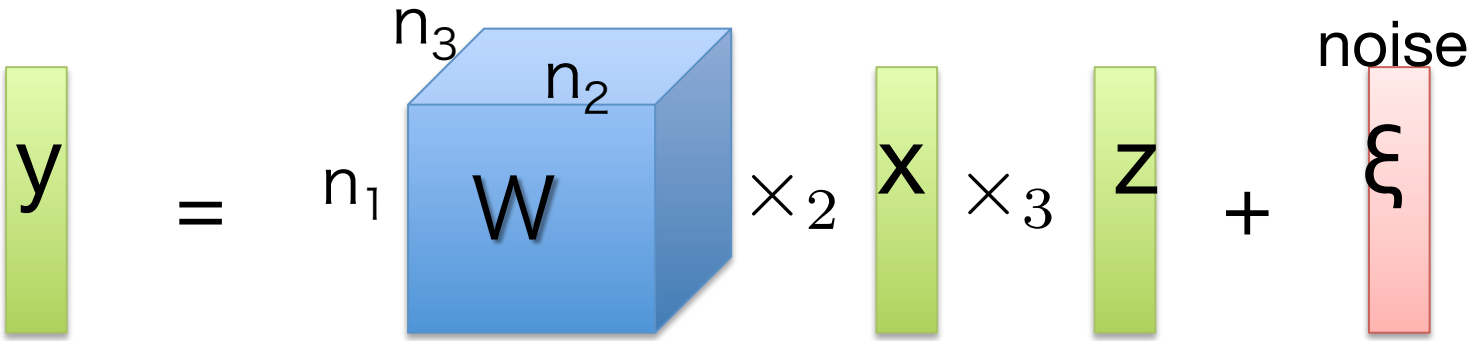
Nuclear norm

Nuclear norm as an atomic norm

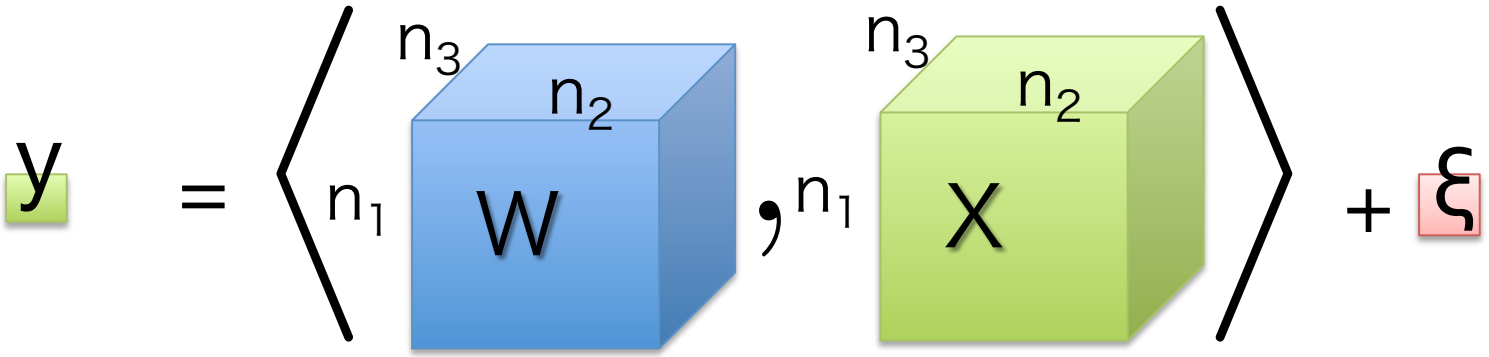
Hardness

Learning with tensors

- Examples
  - Input: pair of vectors (x, z); output: vector y.



  - Input: tensor; output: scalar



Regularization

- In learning setting, we are interested in minimizing

$$\underset{\mathcal{W}\in\mathbb{R}^{n_1\times\cdots\times n_K}}{\text{minimize}} \frac{1}{m}\sum_{i=1}^{m}\ell\left(y_i,\langle\mathcal{X}_i,\mathcal{W}\rangle\right)+R(\mathcal{W})$$

- Examples
  - Low-rank regression: $\mathcal{X}=\boldsymbol{e}_{\text{out}}\circ\boldsymbol{x}\circ\boldsymbol{z}$
  - Tensor completion: $\mathcal{X}=\boldsymbol{e}_i\circ\boldsymbol{e}_j\circ\boldsymbol{e}_k$
  - Tensor classification: (general tensor X)

Regularization

- In learning setting, we are interested in minimizing

$$\underset{\mathcal{W}\in\mathbb{R}^{n_1\times\cdots\times n_K}}{\text{minimize}} \frac{1}{m}\sum_{i=1}^{m}\ell\left(y_i,\langle\mathcal{X}_i,\mathcal{W}\rangle\right)+R(\mathcal{W})$$

- Questions:
  - What regularization term R(W) should we use?
  - How is it related to rank?
  - How does regularizing the factors ($U^{(1)},\ldots,U^{(K)}$) relate to a regullarization for W?

Regularizing the factors

- Recall that the nuclear norm

$$\|\boldsymbol{W}\|_* = \min_{\substack{\boldsymbol{U},\boldsymbol{V}: \\ \boldsymbol{W}=\boldsymbol{U}\boldsymbol{V}^\top}} \frac{1}{2}\left(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2\right)$$

  minimum over all factorizations $W=UV^T$.

- A naïve idea

$$R_{2,2}(\mathcal{W}) = \min_{\substack{\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\boldsymbol{U}^{(3)}: \\ \mathcal{W}=\sum_{r=1}^{R} \boldsymbol{u}_r^{(1)} \circ \boldsymbol{u}_r^{(2)} \circ \boldsymbol{u}_r^{(3)}}} \frac{1}{3}\left(\|\boldsymbol{U}^{(1)}\|_F^2 + \|\boldsymbol{U}^{(2)}\|_F^2 + \|\boldsymbol{U}^{(3)}\|_F^2\right)$$

Is this convex? Is this a norm?

Non-convexity

- Given any decomposition

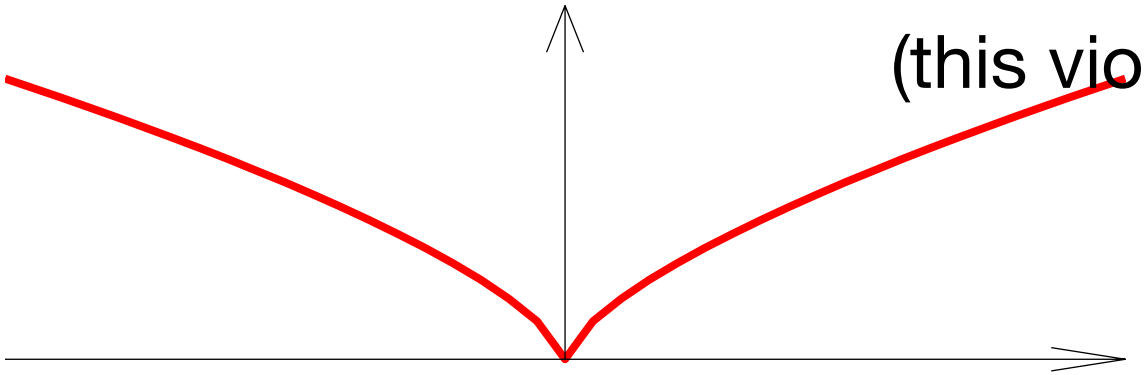
$$R_{2,2}(\mathcal{W}) \leq \left(\|\boldsymbol{U}^{(1)}\|_F \cdot \|\boldsymbol{U}^{(2)}\|_F \cdot \|\boldsymbol{U}^{(3)}\|_F \right)^{\frac{2}{3}}$$

(AM-GM inequality)

- Thus taking $(\hat{\boldsymbol{U}}^{(1)}, \hat{\boldsymbol{U}}^{(2)}, \hat{\boldsymbol{U}}^{(3)}) = \underset{\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}}{\operatorname{argmin}} \left(\cdots \right)$

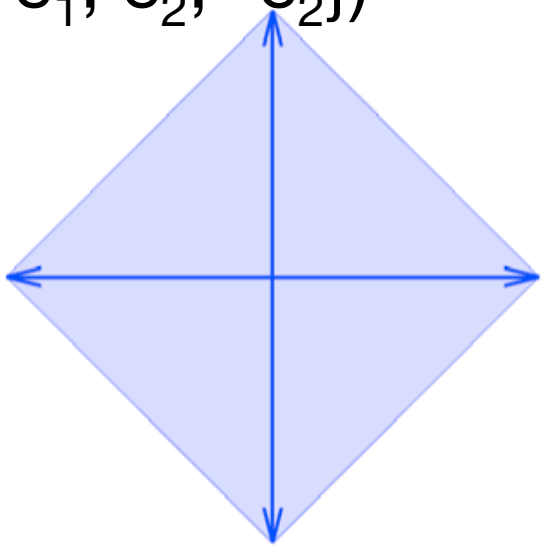
$$R_{2,2}(\alpha\mathcal{W}) \leq \left(|\alpha| \|\hat{\boldsymbol{U}}^{(1)}\|_F \cdot \|\hat{\boldsymbol{U}}^{(2)}\|_F \cdot \|\hat{\boldsymbol{U}}^{(3)}\|_F \right)^{\frac{2}{3}} = |\alpha|^{\frac{2}{3}} R_{2,2}(\mathcal{W})$$

(this violates convexity)



Atomic norm

- Atomic norm [Chandrasekaran+2012] is a norm whose unit ball is the convex hull of an atomic set.
- Example:
  - L_1 norm unit ball in 2D = conv($\{e_1, -e_1, e_2, -e_2\}$)
  - Group lasso
  - Nuclear norm
  - etc.



(Tensor) nuclear norm

- Atomic norm with respect to the atomic set

$$\mathcal{A}_{\text{rank1}} = \left\{ \underbrace{\boldsymbol{u}^{(1)} \circ \cdots \circ \boldsymbol{u}^{(K)}}_{\text{rank-one tensor}} : \|\boldsymbol{u}^{(1)}\|_2 = \cdots = \|\boldsymbol{u}^{(K)}\|_2 = 1 \right\}$$

- Can be written as a minimum

$$\|\mathcal{W}\|_{\text{nuc}} = \min_{\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \ldots, \boldsymbol{U}^{(K)}} \sum_{r=1}^{R} \|\boldsymbol{u}_r^{(1)}\|_2 \cdot \|\boldsymbol{u}_r^{(2)}\|_2 \cdots \|\boldsymbol{u}_r^{(K)}\|_2$$

$$\text{s.t.} \quad \mathcal{W} = \sum_{r=1}^{R} \boldsymbol{u}_r^{(1)} \circ \boldsymbol{u}_r^{(2)} \circ \cdots \circ \boldsymbol{u}_r^{(K)}$$

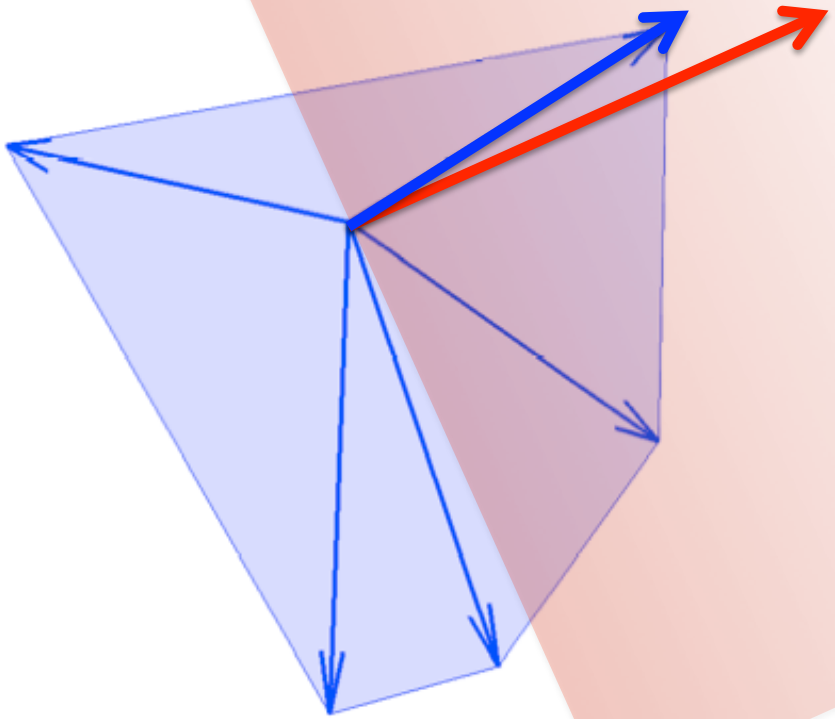
- In other words,

$$\mathcal{W} \in \text{conv}(\mathcal{A}_{\text{rank1}}) \Leftrightarrow \exists \text{decomposition} \sum_{r=1}^{R} \|\boldsymbol{u}_r^{(1)}\|_2 \cdots \|\boldsymbol{u}_r^{(K)}\|_2 \leq 1$$

Dual of nuclear norm is the spectral norm

- Dual of nuclear norm

$$\|\mathcal{A}\|_{\text{nuc}^*} = \max_{\mathcal{W}} \langle \mathcal{A}, \mathcal{W} \rangle , \quad \text{s.t.} \quad \mathcal{W} \in \text{conv}(\mathcal{A}_{\text{rank1}})$$



$$\|\mathcal{A}\|_{\text{nuc}^*} = \|\mathcal{A}\|$$

Summary

Fact:

$$\max_{\substack{\mathcal{X}: \\ \|\mathcal{X}\| \le 1}} \langle \mathcal{X}, \mathcal{W} \rangle = \min_{\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}} \sum_{r=1}^{R} \|\boldsymbol{u}_r^{(1)}\|_2 \cdot \|\boldsymbol{u}_r^{(2)}\|_2 \cdot \|\boldsymbol{u}_r^{(3)}\|_2$$

$$\text{s.t.} \quad \mathcal{W} = \sum_{r=1}^{R} \boldsymbol{u}_r^{(1)} \circ \boldsymbol{u}_r^{(2)} \circ \boldsymbol{u}_r^{(3)}$$

Dual of (tensor) spectral norm = (tensor) nuclear norm = Atomic norm wrt $\mathcal{A}_{\text{rank1}}$

How do we compute this norm?

- Unfortunately it is NP hard to compute both the nuclear norm and the spectral norm for tensors.
- Wait, they are norms (convex functions) aren't they easy to deal with?
- Equivalently written as convex problem with $|\mathcal{A}_{\text{rank1}}|$ many constraints. Just finding the most violated constraint is NP hard.

Generative model?

- By AM-GM inequality

$$\|\mathcal{W}\|_{\text{nuc}} = \min_{\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\ldots,\boldsymbol{U}^{(K)}} \sum_{r=1}^{R} \frac{1}{K}\left(\|\boldsymbol{u}_r^{(1)}\|_2^K + \cdots + \|\boldsymbol{u}_r^{(K)}\|_2^K\right)$$

$$\text{s.t.} \quad \mathcal{W} = \sum_{r=1}^{R} \boldsymbol{u}_r^{(1)} \circ \boldsymbol{u}_r^{(2)} \circ \cdots \circ \boldsymbol{u}_r^{(K)}$$

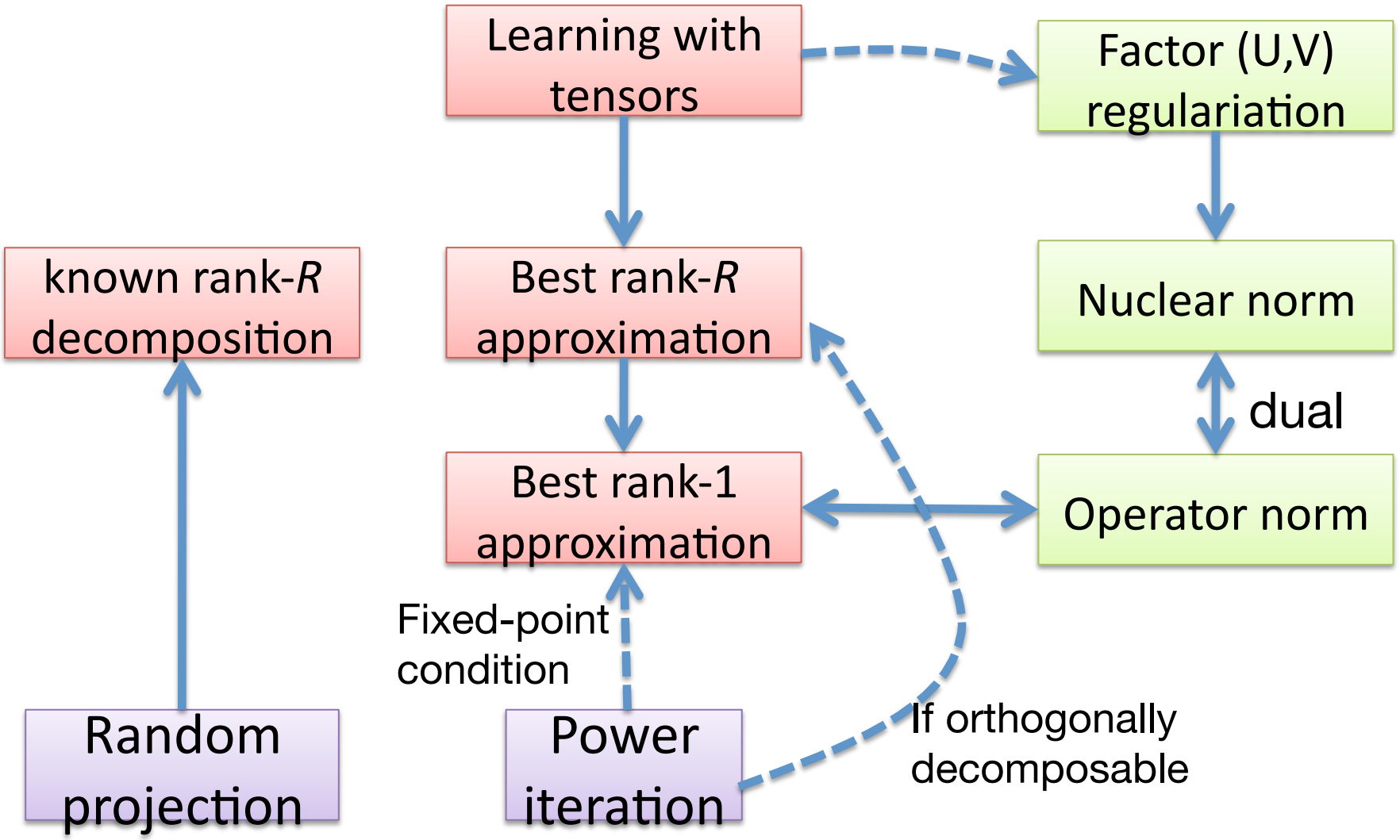
$$P(\boldsymbol{U}^{(1)},\ldots,\boldsymbol{U}^{(K)}) \propto \prod_{k=1}^{K} \exp\left(-\frac{1}{K}\sum_{r=1}^{R}\|\boldsymbol{u}_r^{(k)}\|_2^K\right)$$

- If we consider instead

$$\mathcal{A}_{\text{rank1},K} = \left\{\boldsymbol{u}^{(1)} \circ \cdots \circ \boldsymbol{u}^{(K)} : \|\boldsymbol{u}^{(1)}\|_K = \cdots = \|\boldsymbol{u}^{(K)}\|_K = 1\right\}$$

$$\|\mathcal{W}\|_{\text{nuc}} = \min_{\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\ldots,\boldsymbol{U}^{(K)}} \sum_{r=1}^{R} \frac{1}{K}\left(\|\boldsymbol{u}_r^{(1)}\|_K^K + \cdots + \|\boldsymbol{u}_r^{(K)}\|_K^K\right)$$

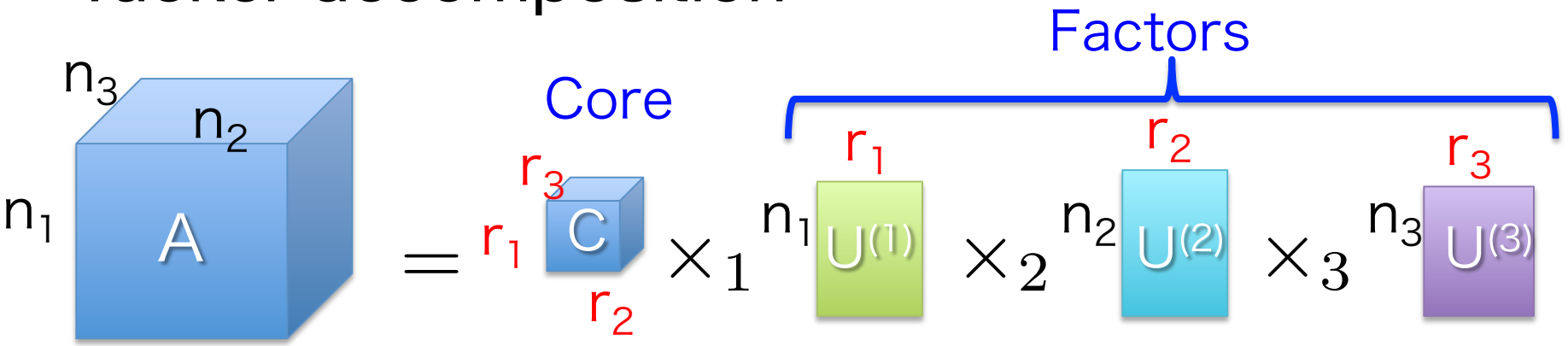
Summary for tensor decomposition



Convex relaxation of Tucker

Overlapped regularization

- If a tensor W is low-rank in the sense of Tucker decomposition



- All unfoldings $W_{(1)}$, $W_{(2)}$, $W_{(3)}$ are simultaneously low-rank.
- How can we encourage W to be simultaneously low-rank?

Approach 1: As a matrix

- Pick a mode *k*, and hope the tensor to be low-rank in mode *k*.

$$\underset{\mathcal{W}\in\mathbb{R}^{n_1\times\cdots\times n_K}}{\text{minimize}} \quad \frac{1}{2m}\|\boldsymbol{y}-\mathfrak{X}(\mathcal{W})\|_2^2+\lambda_m\|\boldsymbol{W}_{(k)}\|_*$$

Observation operator $\mathfrak{X}:\mathbb{R}^{n_1\times\cdots\times n_K}\to\mathbb{R}^m$

$$\mathfrak{X}(\mathcal{W})=\left(\langle\mathcal{X}_1,\mathcal{W}\rangle,\ldots,\langle\mathcal{X}_m,\mathcal{W}\rangle\right)^\top$$

Pro: Basically a matrix problem

⇒Easy to carry out, theory available

Con: Have to be lucky to pick the right mode

Approach 2: Overlapped trace norm

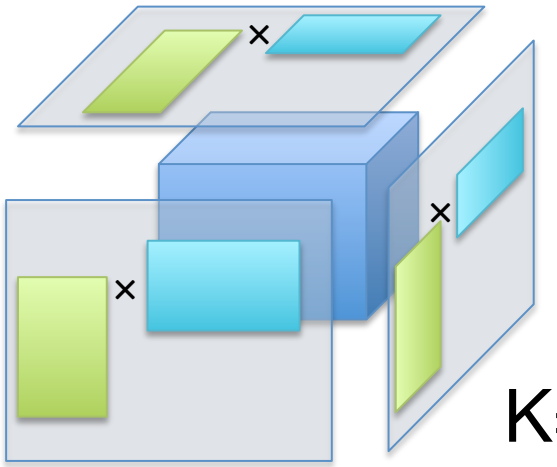
[Liu+09, Signoretto+10, T+10, Gandy+11]

- Regularize by the sum of mode-wise nuclear norms

$$\underset{\mathcal{W}\in\mathbb{R}^{n_1\times\cdots\times n_K}}{\text{minimize}} \quad \frac{1}{2m}\|\boldsymbol{y}-\mathfrak{X}(\mathcal{W})\|_2^2+\lambda_m\sum_{k=1}^{K}\|\boldsymbol{W}_{(k)}\|_*$$

Nuclear norm of unfoldings

- Intuition:
  - the same tensor is regularized to be simultaneously low-rank w.r.t. all modes



K=3

Overlapped nuclear norm

[T+10; Signoretto+10; Gandy+11; Liu+09]

- Regularize the sum of nuclear norms

$$\|\mathcal{W}\|_{\underline{S_1/1}} := \sum_{k=1}^{K} \underbrace{\|\boldsymbol{W}_{(k)}\|_*}_{\text{Nuclear norm of unfoldings}}$$

- Is this a norm?

  Yes –convex because sum of convex functions and positive homogeneous.

- Is this easy to compute?

  Yes –*K*-unfoldings and SVDs.

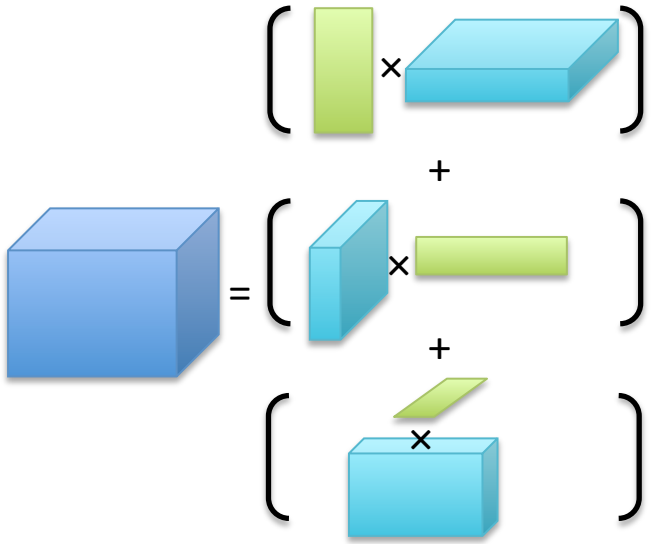
Approach 3: Mixture of low-rank tensors

[T, Hayashi, Kashima10]

- Each mixture component $W^{(k)}$ is regularized to be low-rank in the corresponding mode.

$$\underset{\mathcal{W}\in\mathbb{R}^{n_1\times\cdots\times n_K}}{\text{minimize}} \quad \frac{1}{2m}\left\|\boldsymbol{y}-\mathfrak{X}\left(\sum_{k=1}^{K}\mathcal{W}^{(k)}\right)\right\|_2^2+\lambda_m\sum_{k=1}^{K}\|\boldsymbol{W}_{(k)}^{(k)}\|_*$$

- Each $W^{(k)}$ takes care of each mode
- Sparse solution (group lasso like effect)
- Note: the sum may not be low rank



K=3

Latent nuclear norm

- Effective regularization term

$$\|\mathcal{W}\|_{\overline{S_1/1}} := \min_{\mathcal{W}^{(1)},\ldots,\mathcal{W}^{(K)}} \sum_{k=1}^{K} \left\| \boldsymbol{W}^{(k)}_{(k)} \right\|_* \quad \text{s.t.} \quad \mathcal{W} = \sum_{k=1}^{K} \mathcal{W}^{(k)}$$

- Is this a norm?

  Yes –it is the dual of the overlapped S∞/∞-norm:

$$\|\mathcal{X}\|_{\underline{S_\infty/\infty}} := \max_k \left\| \boldsymbol{X}_{(k)} \right\|$$

spectral norm

- Is it easy to compute?

  Yes –convex minimization problem involving matrix nuclear norm

Property of the norms

- Nuclear norm

$$\|\boldsymbol{W}\|_* \leq \sqrt{r}\|\boldsymbol{W}\|_F \quad \text{if } \operatorname{rank}(\boldsymbol{W}) \leq r,$$

- Overlapped trace norm

$$\|\mathcal{W}\|_{\underline{S_1/1}} \leq \sum_{k=1}^{K} \sqrt{r_k}\|\mathcal{W}\|_F$$

if $\mathcal{W}$ has multilinear rank $(r_1, \ldots, r_K)$

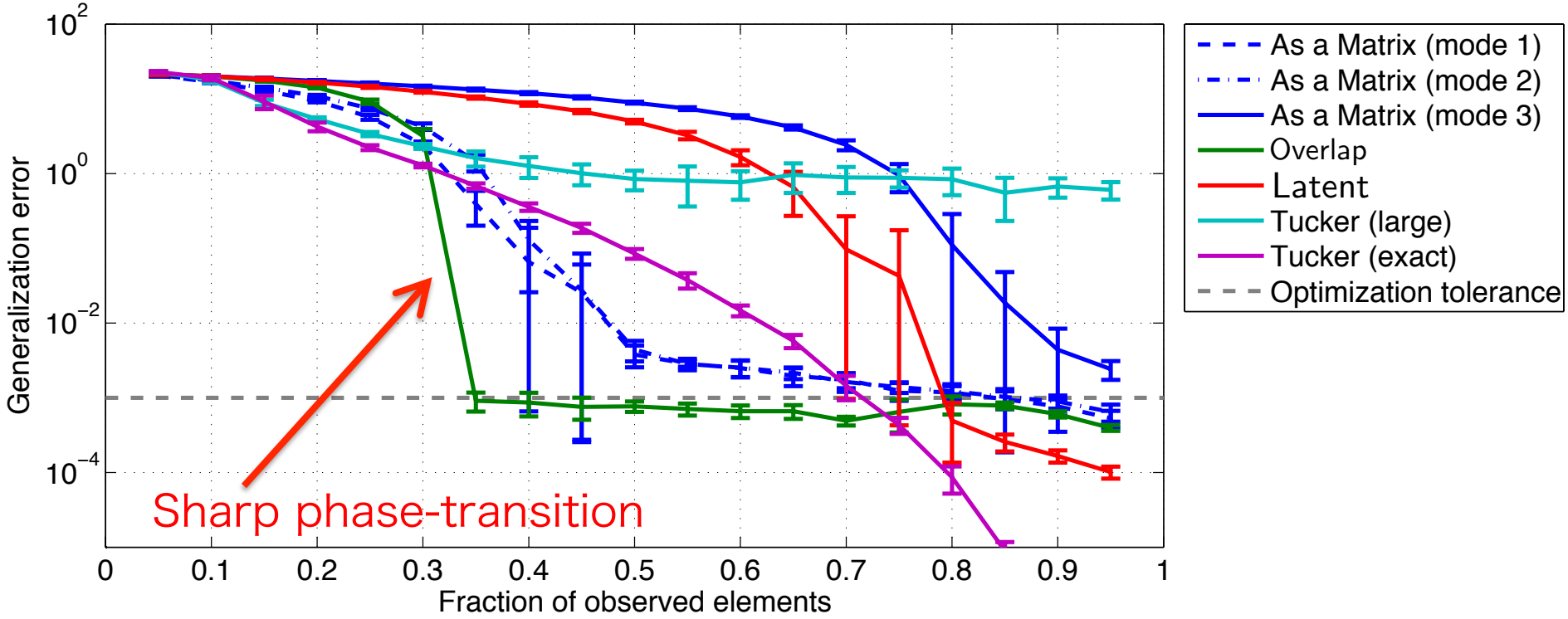
- Latent trace norm

$$\|\mathcal{W}\|_{\overline{S_1/1}} \leq \min_k \sqrt{r_k}\|\mathcal{W}\|_F$$

if $\mathcal{W}$ has multilinear rank $(r_1, \ldots, r_K)$

Tensor completion result

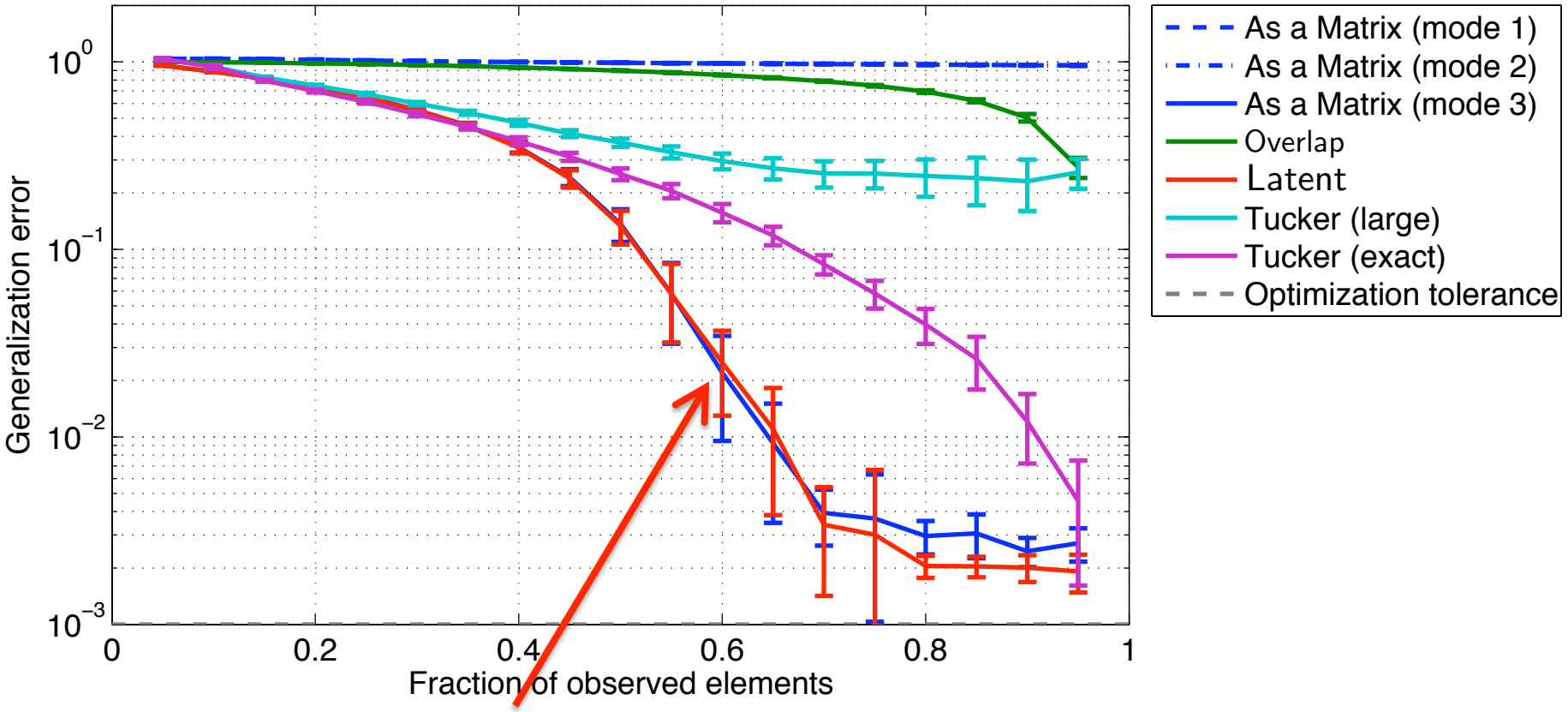
- True tensor: 50x50x20, rank 7x8x9. No noise (λ=0).
- Random train/test split.



Tucker implemented in n-way toolbox. [Andersson & Bro, 00]

Almost full-rank case

True tensor: Size 50x50x20, rank 50x50x5. No noise (λ=0).



Latent approach can automatically figure out which mode is low-rank

Overlapped norm: key properties

- How it relates to the multilinear rank

$$\|\mathcal{W}\|_{\underline{S_1/1}} \leq \underbrace{\sum_{k=1}^{K} \sqrt{r_k}}_{\text{average of the ranks}} \|\mathcal{W}\|_F$$

- How the dual norm behaves

$$\mathbb{E}\|\mathcal{X}\|_{(\underline{S_1/1})^*} \leq O\left(\sigma \frac{1}{K^2} \sum_{k=1}^{K} \sqrt{N/n_k}\right), \qquad (N = \prod_k n_k)$$

Random Gaussian tensor N(0, σ^2)

Overlapped norm: denoising guarantee

[T,Suzuki,Hayashi,Kashima,11]

- Let Y be a noisy observation of W*

$$\mathcal{Y} = \mathcal{W}^* + \mathcal{E} \qquad (\mathcal{E} \sim \mathcal{N}(0, \sigma^2))$$

- Then the estimator

$$\hat{\mathcal{W}} = \operatorname*{argmin}_{\mathcal{W}} \left(\frac{1}{2} \left\| \mathcal{Y} - \mathcal{W} \right\|_F^2 + \lambda \left\| \mathcal{W} \right\|_{\underline{S_1/1}} \right)$$

with $\lambda = c_0 \sigma (\sum_{k=1}^{K} \sqrt{N/n_k})/K$ gives

$$\frac{1}{N} \left\| \hat{\mathcal{W}} - \mathcal{W}^* \right\|_F^2 \leq c_1 \sigma^2 \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{r_k} \right)^2 \left(\frac{1}{K} \sum_{k=1}^{K} 1/\sqrt{n_k} \right)^2$$

with prob. at least 1-exp(-poly(n))

Latent norm: key properties

- How it relates to the multilinear rank

$$\|\mathcal{W}\|_{\overline{S_1/1}} \leq (\min_k \sqrt{r_k}) \|\mathcal{W}\|_F$$

minimum rank

- How the dual norm behaves

$$\mathbb{E}\|\mathcal{X}\|_{(\overline{S_1/1})^*} \leq \tilde{O}\left(\sigma \max_k \sqrt{N/n_k}\right), \quad (N = \prod_k n_k)$$

Random Gaussian tensor N(0, σ^2)

Denoising guarantees

- Overlapped nuclear norm: sensitive to the average rank

$$\frac{1}{N}\left\|\left|\hat{\mathcal{W}}-\mathcal{W}^{*}\right|\right\|_{F}^{2} \leq c_{1}\sigma^{2}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{r_{k}}\right)^{2}\left(\frac{1}{K}\sum_{k=1}^{K}1/\sqrt{n_{k}}\right)^{2}$$

- Latent nuclear norm: sensitive to the minimum rank

$$\frac{1}{N}\left\|\left|\hat{\mathcal{W}}-\mathcal{W}^{*}\right|\right\|_{F}^{2} \leq c_{2}\sigma^{2}\min_{k} r_{k}/\min_{k} n_{k}$$

- Matters only if $r_k \neq r_k'$ or $n_k \neq n_k'$ (typical many tensor problems)

Could we obtain a bound of $\min_k(r_k/n_k)$?

Scaled latent nuclear norm

[Wimalawarne, Sugiyama, T, 2014]

- Normalize by $\sqrt{n_k}$

$$\|\mathcal{W}\|_{\text{scaled}} = \inf_{\mathcal{W}^{(1)},\ldots,\mathcal{W}^{(K)}} \sum_{k=1}^{K} \frac{1}{\sqrt{n_k}} \|\boldsymbol{W}_{(k)}^{(k)}\|_{S_1} \quad \text{s.t.} \quad \sum_{k=1}^{K} \mathcal{W}^{(k)} = \mathcal{W}$$

  – relation to the multilinear rank

$$\|\mathcal{W}\|_{\text{scaled}} \leq \left(\min_k \sqrt{\frac{r_k}{n_k}} \right) \|\mathcal{W}\|_F$$

  – behavior of the dual norm

$$\mathbb{E}\|\mathcal{X}\|_{\text{scaled}^*} \leq \tilde{O}\left(\sigma\sqrt{N}\right), \quad (N = \textstyle\prod_k n_k)$$

Denoising guarantees

- Overlapped nuclear norm: sensitive to the average rank

$$\frac{1}{N}\left\|\left|\hat{\mathcal{W}}-\mathcal{W}^*\right|\right\|_F^2 \leq c_1\sigma^2\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{r_k}\right)^2\left(\frac{1}{K}\sum_{k=1}^{K}1/\sqrt{n_k}\right)^2$$

- Latent nuclear norm: sensitive to the minimum rank

$$\frac{1}{N}\left\|\left|\hat{\mathcal{W}}-\mathcal{W}^*\right|\right\|_F^2 \leq c_2\sigma^2\min_k r_k/\min_k n_k$$

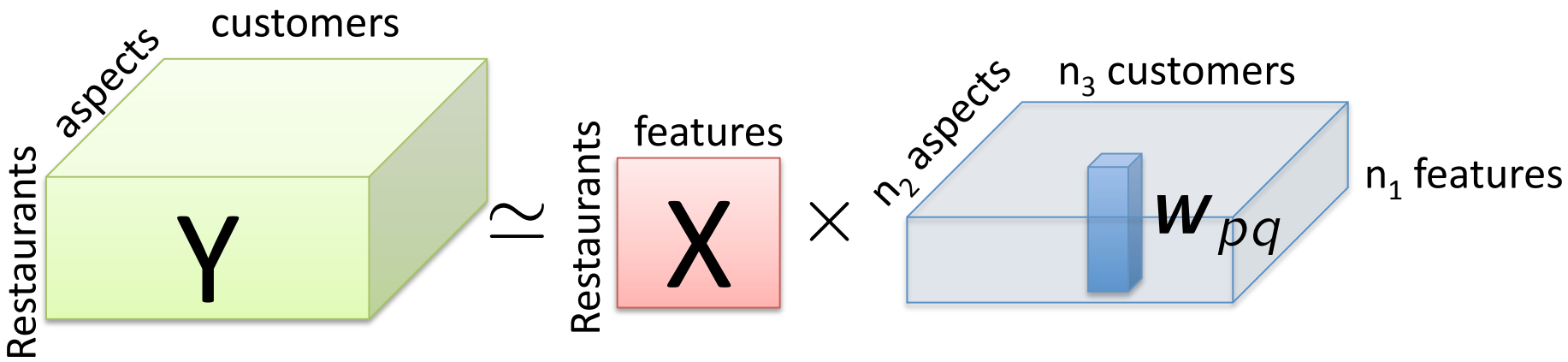
- Scaled latent norm: sensitive to the minimum rank-to-dimension ratio

$$\frac{1}{N}\left\|\left|\hat{\mathcal{W}}-\mathcal{W}^*\right|\right\|_F^2 \leq c_3\sigma^2\min_k \frac{r_k}{n_k}$$

Multi-task learning with tensors

[Romera-Paredes+13; Signoretto+13]

- Example: multi-aspect restaurant recommendation



- Issue: heterogeneous dimensions and ranks
  - Number of aspects is usually much smaller than the number of customers.
  - Rank (amount of sharing of information) may be different for aspect, customers, and features.

Theory for multi-linear MTL

- n_2 x n_3 tasks. Observed tasks $S \subseteq [n_2] \times [n_3]$
- Training samples
  - many tasks may have no samples (zero-shot learning) $(\boldsymbol{x}_{ipq}, y_{ipq})_{i=1}^{m_{pq}}, \quad (p,q) \in S$
- Training

$$\hat{\mathcal{W}} = \underset{\mathcal{W} \in \mathbb{R}^{n_1 \times n_2 \times n_3}}{\operatorname{argmin}} \hat{L}(\mathcal{W}), \text{ s.t. } \|\|\mathcal{W}\|\|_\star \leq C$$

where $\hat{L}(\mathcal{W}) = \frac{1}{|S|} \sum_{(p,q) \in S} \frac{1}{m_{pq}} \sum_{i=1}^{m_{pq}} \ell(\langle \boldsymbol{x}_{ipq}, \boldsymbol{w}_{pq} \rangle, y_{ipq})$

Analysis

[Wimalawarne, Sugiyama, T, 14]

- Lemma: Let W* be any tensor with rank (r1,r2,r3) and element-wise bounded by B. For $C = B\phi(\boldsymbol{r})\sqrt{N}$

$$L(\hat{\mathcal{W}}) - L(\mathcal{W}^*) \leq O\left(\Lambda B \phi(\boldsymbol{r}) \cdot \frac{\sqrt{N}}{|S|} \cdot \mathbb{E}\left\| \sum_{(p,q)\in S} \frac{1}{m_{pq}} \sum_{i=1}^{m_{pq}} \sigma_{ipq} \mathcal{X}_{ipq} \right\|_{\star^*}\right)$$

How the norm relates to the rank

$$\left\|\mathcal{W}\right\|_\star \leq \phi(\boldsymbol{r}) \left\|\mathcal{W}\right\|_F$$

How the dual norm behaves

σ_{ipq}: Rademacher RV

Λ : Lipschitz constant of the loss ℓ

κ : Measure of feature correlation $\mathbb{E}\left[\boldsymbol{x}_{ipq}\boldsymbol{x}_{ipq}{}^\top\right] = C_{pq} \preceq \kappa/n_1 \boldsymbol{I}_{n_1}$

$L(\mathcal{W})$: Expected loss $\quad L(\mathcal{W}) = \frac{1}{n_2 n_3} \sum_{(p,q)\in[n_2]\times[n_3]} \mathbb{E}\ell\left(\left\langle \boldsymbol{x}_{pq}, \boldsymbol{w}_{pq}\right\rangle, y_{pq}\right)$

Theorem (overlap norm)

[Wimalawarne, Sugiyama, T, 14]

- Let $\mathcal{W}^*$ be any tensor with rank (r_1,r_2,r_3) and element-wise bounded by B. For $C = B\sqrt{\|\boldsymbol{r}\|_{1/2} n_1 n_2 n_3}$

$$L(\hat{\mathcal{W}}) - L(\mathcal{W}^*) \leq O\left(\Lambda B \sqrt{\frac{\kappa}{m|S|}\|\boldsymbol{r}\|_{1/2} \min_k \left(D_k \log D_k\right)}\right)$$

Λ : Lipschitz constant of the loss ℓ

κ : Measure of feature correlation $\mathbb{E}\left[\boldsymbol{x}_{ipq}\boldsymbol{x}_{ipq}{}^{\top}\right] = C_{pq} \preceq \kappa/n_1 \boldsymbol{I}_{n_1}$

$$\|\boldsymbol{r}\|_{1/2} = \left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{r_k}\right)^2,$$

$$D_1 := n_1 + n_2 n_3, \quad D_2 := n_2 + n_1 n_3, \quad D_3 := n_3 + n_1 n_2$$

Theorem (latent norm)

[Wimalawarne, Sugiyama, T, 14]

- Let W* be any tensor with rank (r_1,r_2,r_3) and element-wise bounded by *B*. For $C = B\sqrt{\min_k r_k n_1 n_2 n_3}$

$$L(\hat{\mathcal{W}}) - L(\mathcal{W}^*) \leq O\left(\Lambda B \sqrt{\frac{\kappa}{m|S|} \min_k r_k \max_k (D_k \log D_k)}\right)$$

Λ : Lipschitz constant of the loss ℓ

κ : Measure of feature correlation $\mathbb{E}\left[\boldsymbol{x}_{ipq}\boldsymbol{x}_{ipq}{}^{\top}\right] = C_{pq} \preceq \kappa/n_1 \boldsymbol{I}_{n_1}$

$D_1 := n_1 + n_2 n_3, \quad D_2 := n_2 + n_1 n_3, \quad D_3 := n_3 + n_1 n_2$

Theorem (scaled latent norm)

[Wimalawarne, Sugiyama, T, 14]

- Let W* be any tensor with rank (r_1,r_2,r_3) and element-wise bounded by *B*. For $C = B\sqrt{\min_k (r_k/n_k) n_1 n_2 n_3}$

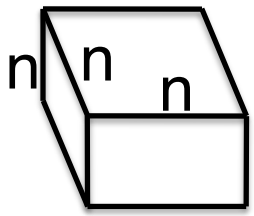
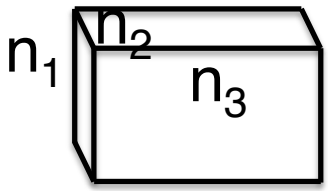
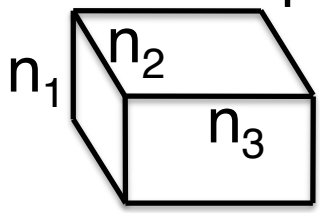
$$L(\hat{\mathcal{W}}) - L(\mathcal{W}^*) \leq O\left(\Lambda B \sqrt{\frac{\kappa}{m|S|} \min_k \left(\frac{r_k}{n_k}\right) n_1 n_2 n_3 \log(\max_k D_k)}\right)$$

Λ : Lipschitz constant of the loss ℓ

κ : Measure of feature correlation $\mathbb{E}\left[\boldsymbol{x}_{ipq}\boldsymbol{x}_{ipq}{}^{\top}\right] = C_{pq} \preceq \kappa/n_1 \boldsymbol{I}_{n_1}$

$D_1 := n_1 + n_2 n_3, \quad D_2 := n_2 + n_1 n_3, \quad D_3 := n_3 + n_1 n_2$

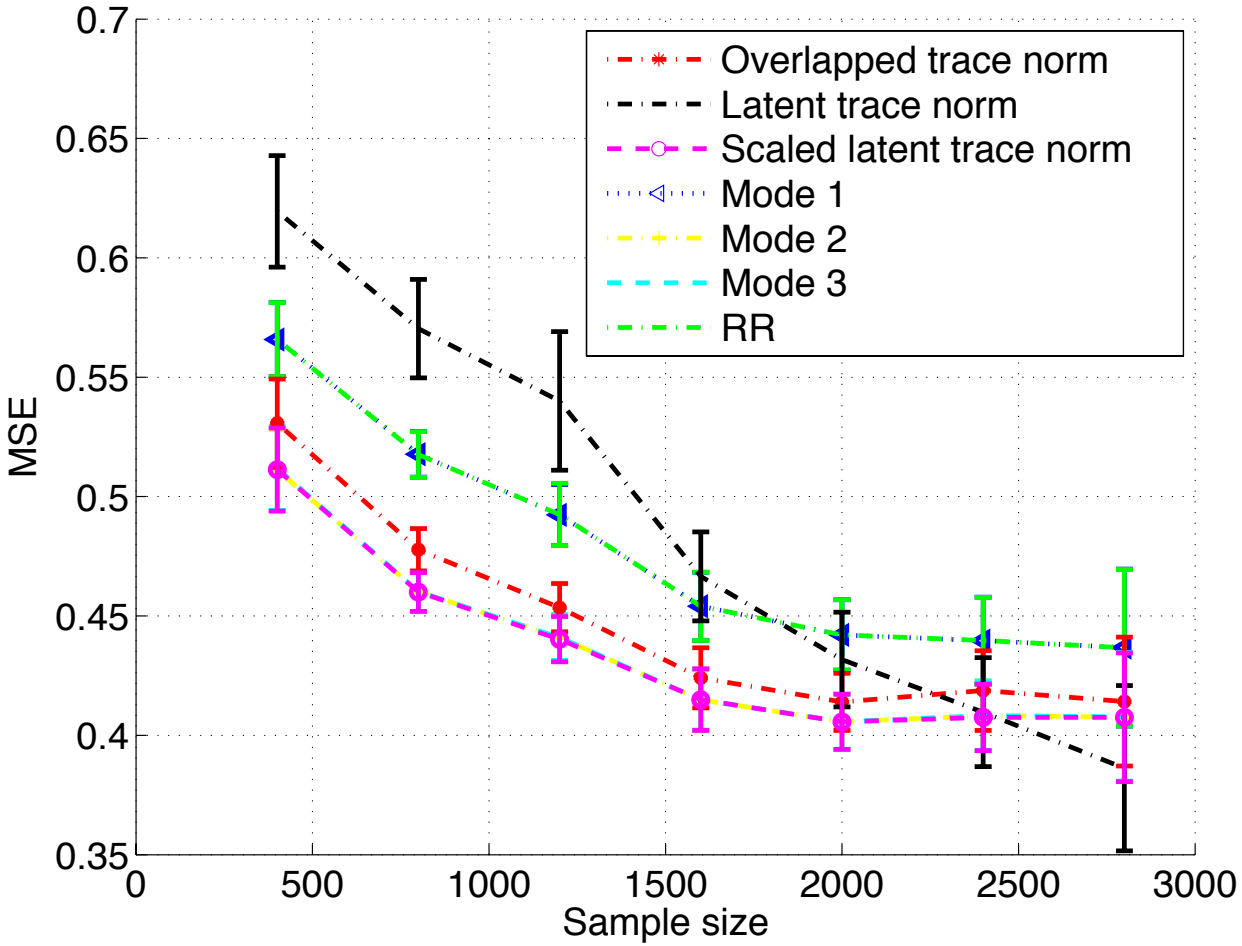
Sample complexities
(ignoring log terms and per $1/\varepsilon^2$)

	Overlap	Latent	Scaled latent
Multi-task learning (homogeneous) 	$\kappa \lVert \boldsymbol{r} \rVert_{1/2} n^2$	$\kappa (\min_k r_k) n^2$	$\kappa (\min_k r_k) n^2$
Multi-task learning (heterogeneous) 	$\kappa \lVert \boldsymbol{r} \rVert_{1/2} n_2 n_3$	$\kappa n_1 n_2 n_3$	$\kappa \min(r_1 n_2 n_3, n_1 n_2 r_3)$
Tensor completion 	$\lVert \boldsymbol{r} \rVert_{1/2} \min_k D_k$	$\min_k r_k \max_k D_k$	$\min_k \frac{r_k}{n_k} N$

Restaurant recommendation data

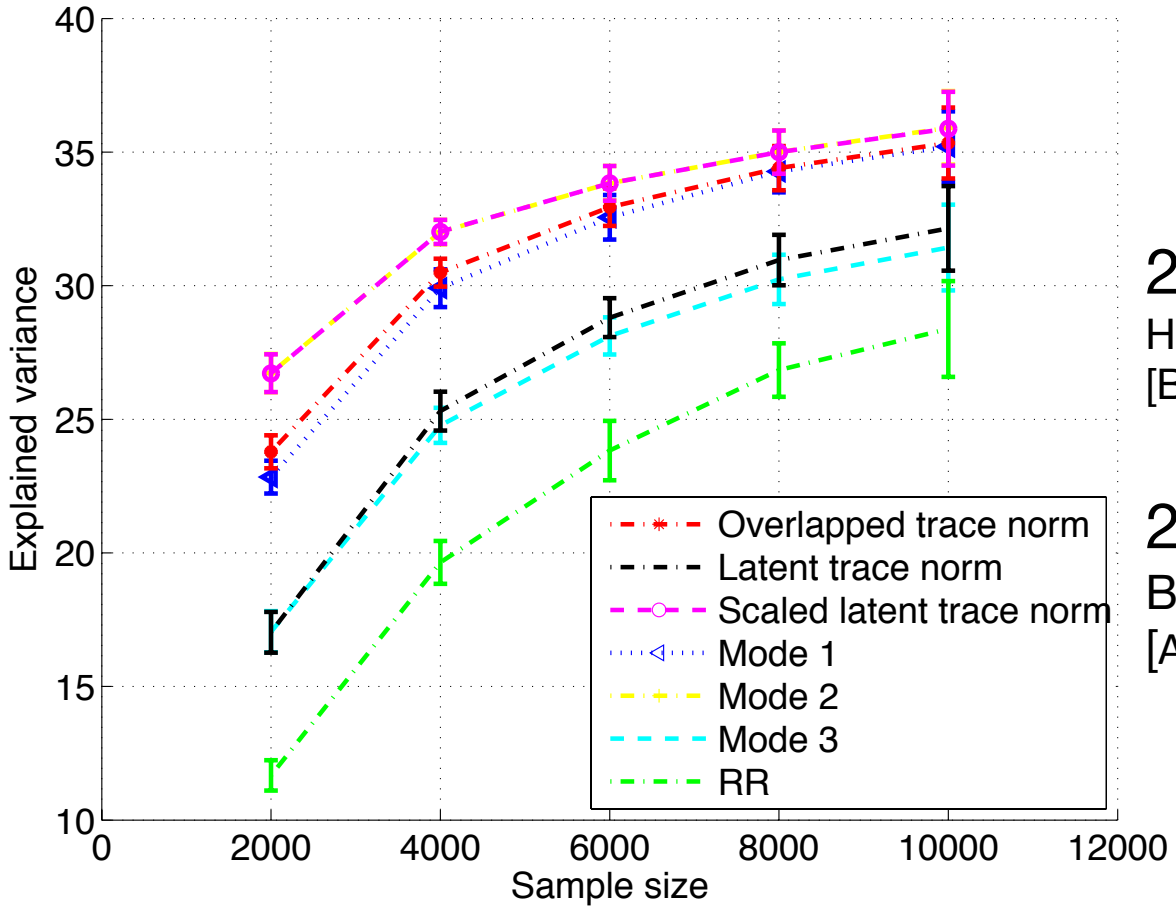
[Vargas-Govea+2011; Romera-Paredes+2013]

- 45 features x 3 aspects x 138 customers



ILEA School data [ILEA; Bakker & Heskes 2003]

- 24 features x 139 schools x 3 years



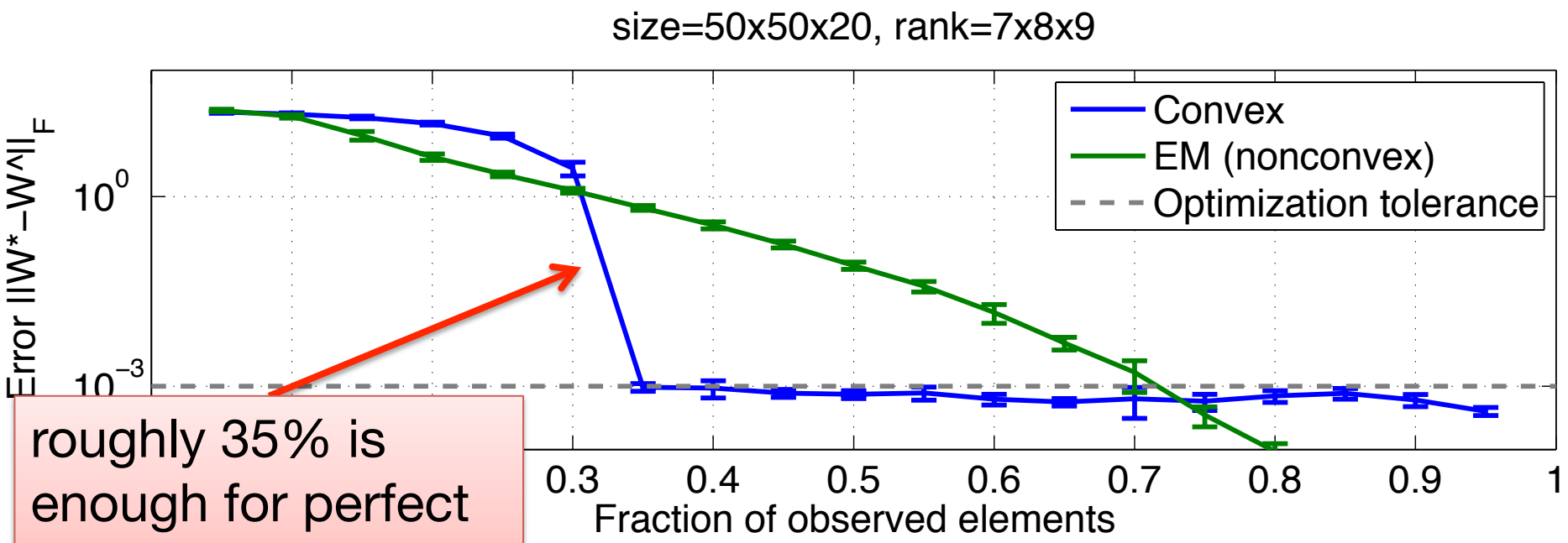
29.5%
Hierarchical Bayes
[Bakker & Heskes 03]

26.7%
Bilinear MTL
[Argyriou 08]

Tensor completion

$$\min_{\mathcal{W}} \quad \|\mathcal{W}\|_{S_1/1}$$

$$\text{s.t.} \quad W_{i,j,k} = Y_{i,j,k}$$



roughly 35% is enough for perfect recovery!

EM algorithm implemented in n-way toolbox. [Andersson & Bro, 00]

Problem setting

Observation $\mathcal{W}^*$: true K-way tensor with rank (r_1,...,r_K)

$$y_i = \langle \mathcal{X}_i, \mathcal{W}^* \rangle + \epsilon_i \quad (i = 1, \ldots, m)$$

Gaussian noise $N(0, \sigma^2)$

Optimization Likelihood Regularization

$$\hat{\mathcal{W}} = \underset{\mathcal{W} \in \mathbb{R}^{n_1 \times \cdots \times n_K}}{\operatorname{argmin}} \left(\frac{1}{2m} \|\boldsymbol{y} - \mathfrak{X}(\mathcal{W})\|_2^2 + \lambda_m \|\mathcal{W}\|_{S_1/1} \right)$$

Reg. constant

Observation operator $\mathfrak{X} : \mathbb{R}^{n_1 \times \cdots \times n_K} \to \mathbb{R}^m$

$$\mathfrak{X}(\mathcal{W}) = (\langle \mathcal{X}_1, \mathcal{W} \rangle, \ldots, \langle \mathcal{X}_m, \mathcal{W} \rangle)^\top$$

Theorem (random Gaussian design)

There are constants c_1, c_2 such that

$$\left\| \hat{\mathcal{W}} - \mathcal{W}^* \right\|_F^2 \leq c_2 \sigma^2 \frac{r n^{K-1}}{M}$$

holds with high probability if

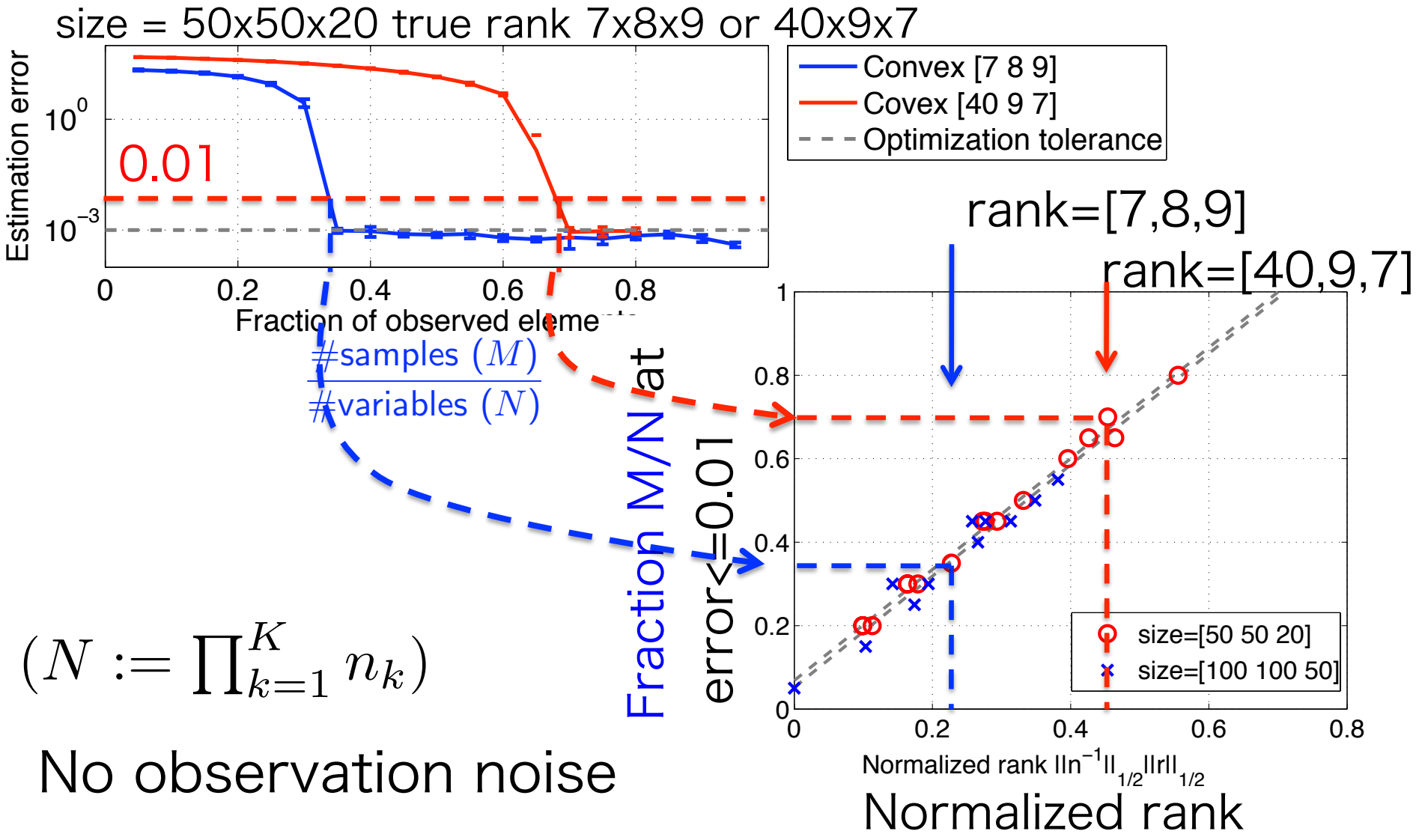
$$\frac{\#\mathrm{samples}(m)}{\#\mathrm{variables}(\prod_k n_k)} \geq c_1 \frac{r}{n} \quad \text{and} \quad \lambda_m = 8\sigma\sqrt{n^{K-1}/m}$$

Condition for the sample size m:
- Depends on the normalized rank r/n
- Independent of the noise σ

Condition for the reg. parameter λ_m:
- Independent of the rank r
- Deepends on the noise σ

For simplicity $n_1,\ldots,n_K=n$, $r_1,\ldots,r_K=r$

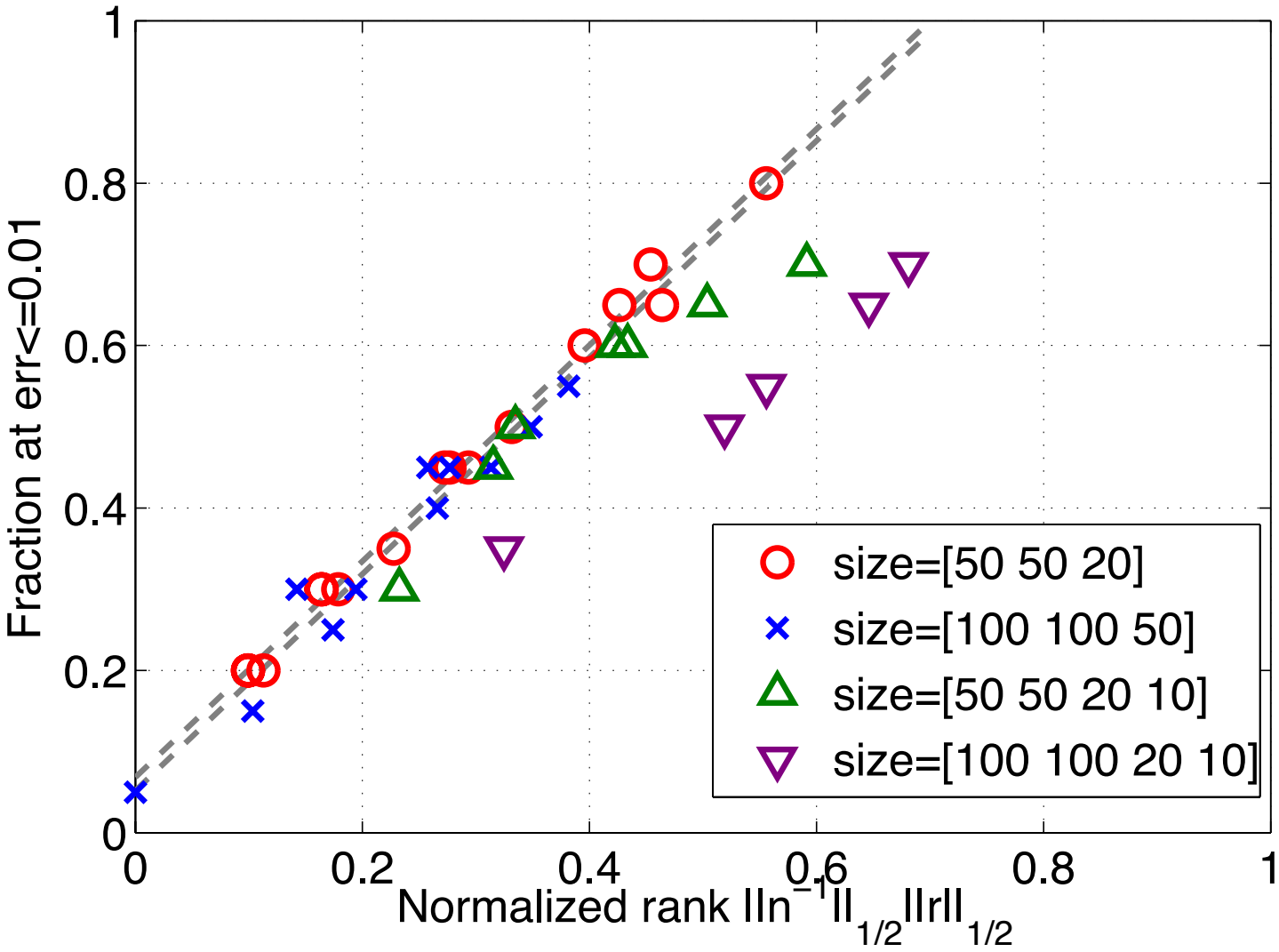
Tensor completion



$(N := \prod_{k=1}^{K} n_k)$

No observation noise

Theory vs. Experiments (4th order)



Limitation: too many samples required!

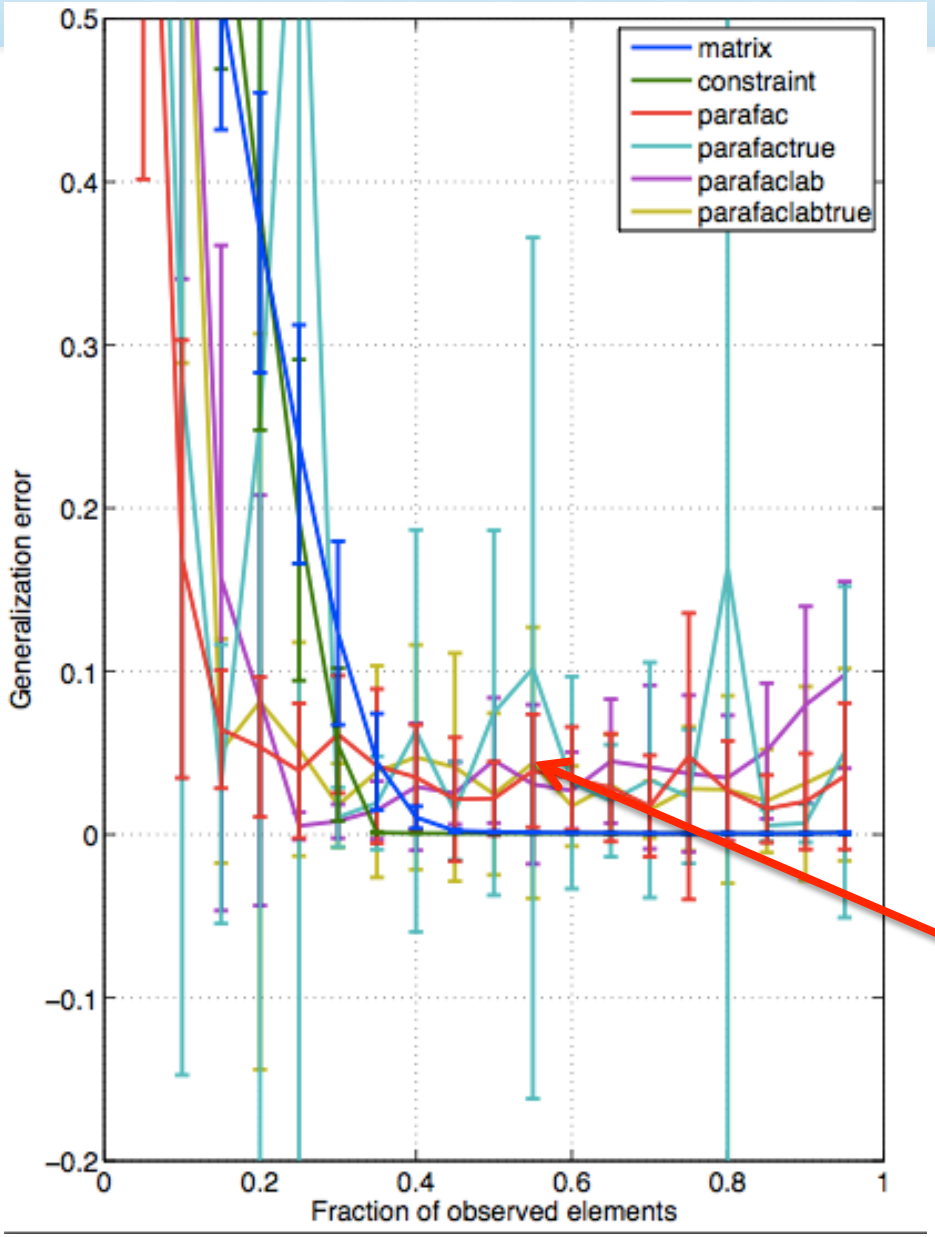
- Sample complexity = number of samples required to obtain error ε

$$\left\| \hat{\mathcal{W}} - \mathcal{W}^* \right\|_F^2 \leq c_2 \sigma^2 \frac{r n^{K-1}}{M} \quad \Leftrightarrow \quad M = O(r n^{K-1} / \epsilon)$$

- For *K*=2: O(rn) = number of parameters rank r matrix has.
- For *K*>2: too many. In fact, for the 50 x 50 x 20 rank (7,8,9) case, about 1500 (≒50*7+50*8+20*9+7*8*9) samples (3% << 35%) should be right.

Tensorlab

Sorber, Van Barel and De Lathauwer (2014).

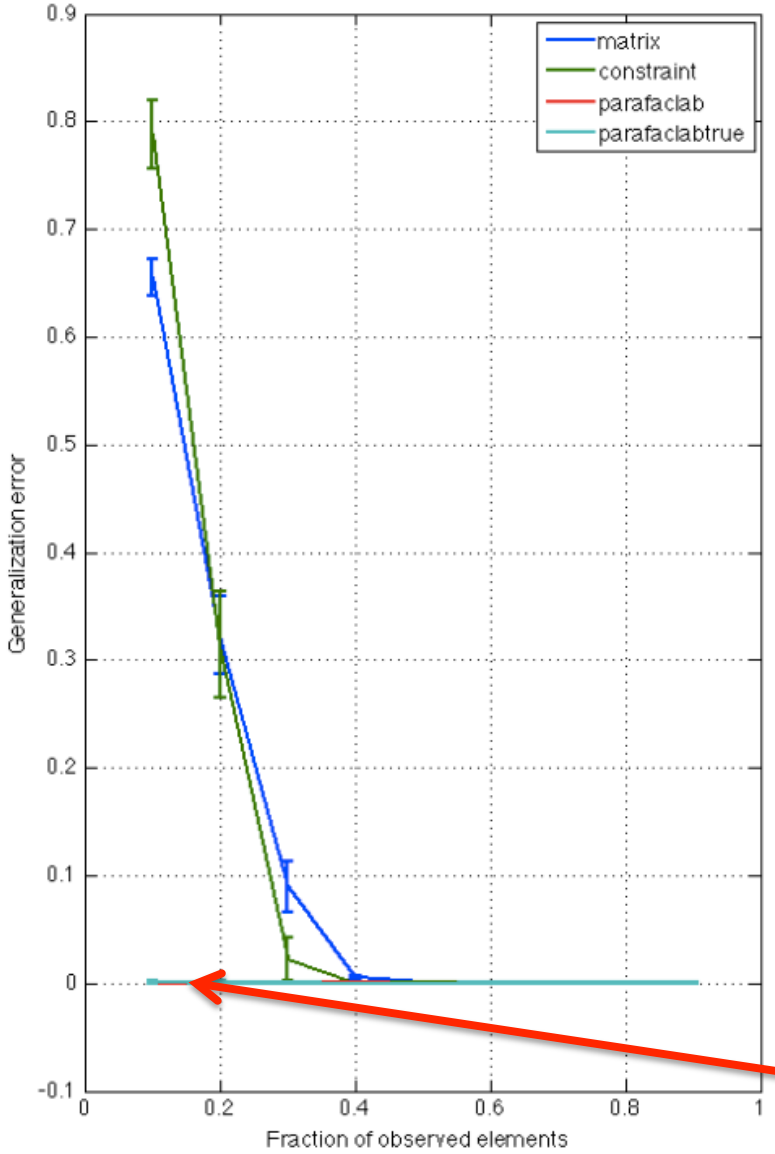


- RT “Hi, I was trying cpd with missing entries and found that the error does not go to zero as the number of observed entries increases…”

Tensorlab

Tensorlab

Sorber, Van Barel and De Lathauwer (2014).



- Laurent Sorber's reply "Thank you for your detailed example … I have been experimenting further with your code, and below is a method that gets good performance with Tensorlab …"

Tensorlab

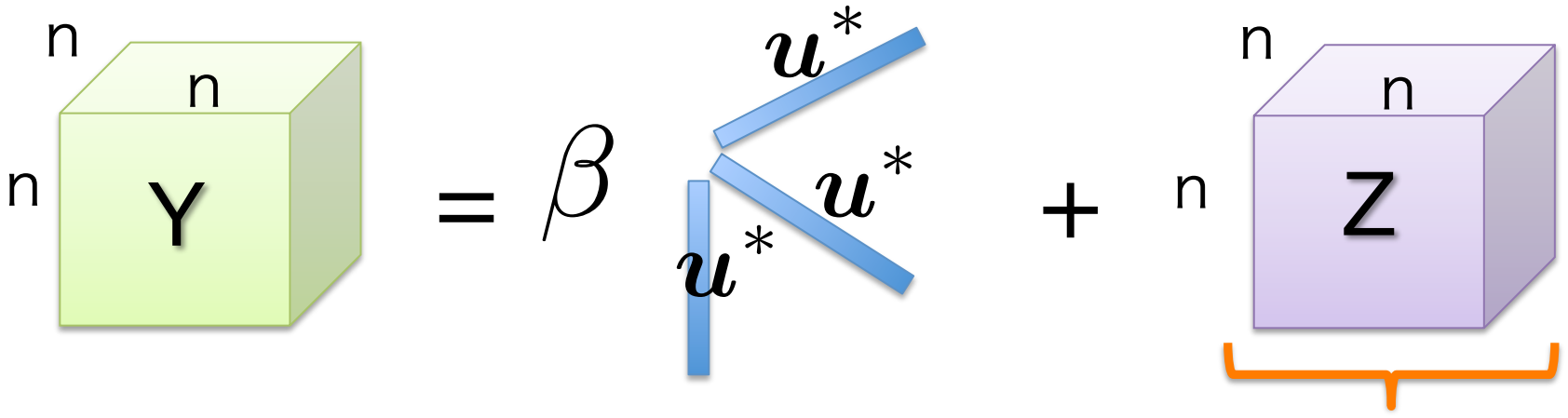
Lesson learned

- When you are doing something non-convex, details matter a lot!
  - Random initialization turns out to be better than the default intelligent initialization.
  - Make sense to try several short restarts (4 x 5).
  - Default max iteration too small.
- Although global guarantee may be absent, nonlinear optimization could work very well in practice.

Computation-statistics trade-off

Simplest case [Montanari & Richard 14]

- Noise corrupted symmetric rank-one tensor



β: signal-to-noise ratio

- How large does β need to be to recover u* from *Y*?

Optimal: maximum likelihood

[Richard & Montanari 14]

- Consider the maximum likelihood estimator

$$\hat{\boldsymbol{u}}_{\mathrm{ML}} = \operatorname*{argmax}_{\boldsymbol{u}\in\mathbb{R}^n} \langle \mathcal{Y}, \boldsymbol{u}\circ\boldsymbol{u}\circ\boldsymbol{u}\rangle$$

- If $\beta \geq \mathbb{E}\left|\!\left|\!\left| \mathcal{Z} \right|\!\right|\!\right|$, with high probability

Operator norm

$$1 - |\langle \hat{\boldsymbol{u}}_{\mathrm{ML}}, \boldsymbol{u}^*\rangle| = O\left(\frac{\mathbb{E}\left|\!\left|\!\left| \mathcal{Z} \right|\!\right|\!\right|}{\beta}\right)$$

- Note that $\mathbb{E}\left|\!\left|\!\left| \mathcal{Z} \right|\!\right|\!\right| = O\left(\sqrt{nK\log(K)}\right)$ [T&Suzuki 14]
- This is optimal (no estimator can do better)
- But computationally interactable (best rank-1 approx.)!

Power method

[Anandkumar+12; Richard & Montanari 14]

- Is a computationally tractable algorithm
- With $\beta \geq c \cdot \sqrt{n^K K \log(K)}$ randomly initialized power iteration converges to a solution with

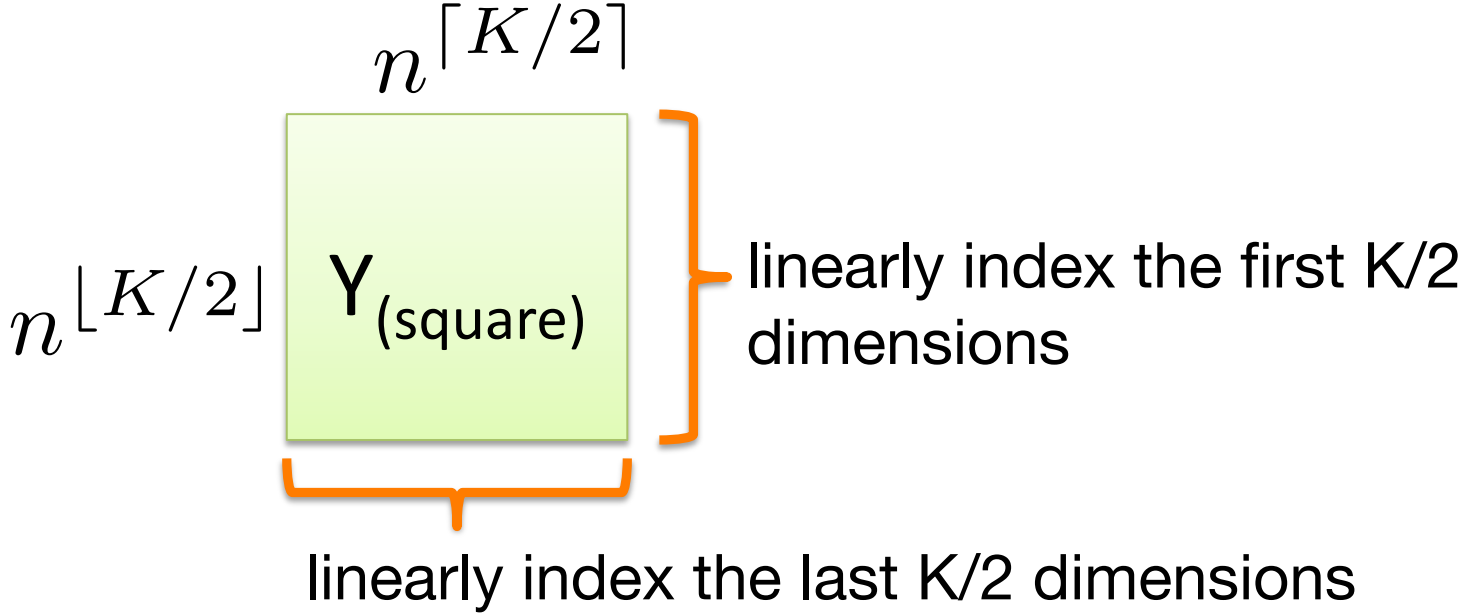
$$1 - |\langle \hat{\boldsymbol{u}}, \boldsymbol{u}^* \rangle| = O\left(\frac{\mathbb{E}|||\mathcal{Z}|||}{\beta}\right)$$

  with high probability.
- Note the sharp increase in the required β (trading statistical performance for tractability)

Recursive unfolding

[Richard & Montanari 14]

1. Matricize *Y* so that it looks as square as possible



2. Compute the leading singular vector v. Note that v is $n^{K/2}$ dimensional.
3. Reshape v into an n x $n^{K/2-1}$ matrix and return the top-left singular vector.

Performance of (recursive) unfolding

- Richard & Montanari: recursive unfolding attains

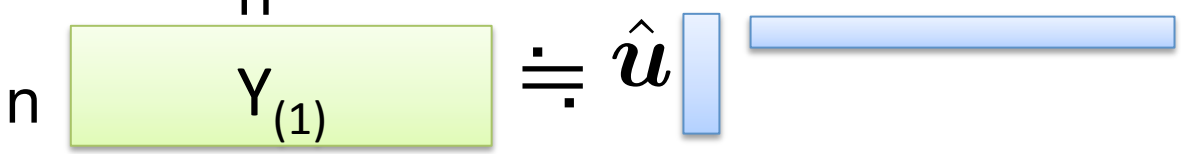
$$1 - |\langle \hat{\boldsymbol{u}}, \boldsymbol{u}^* \rangle| = O\left(\frac{K n^{\lceil K/2 \rceil}}{\beta^2}\right)$$

under $\beta \geq c \cdot \sqrt{n^{\lceil K/2 \rceil} K}$ with high probability.

  - loose for odd-order tensors –conjectured $n^{K/4}$ suffice

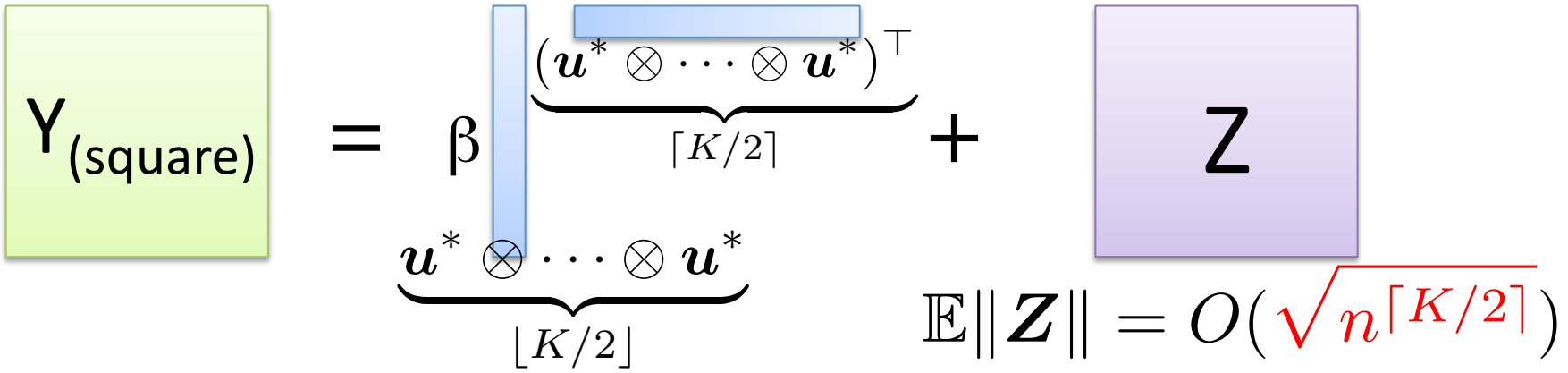
- A tighter analysis [Zheng & T, 2015]: taking the top-left singular vector of the ordinary (rectangular) unfolding is sufficient to obtain a threshold

$$\beta \geq c' \cdot n^{K/4}$$



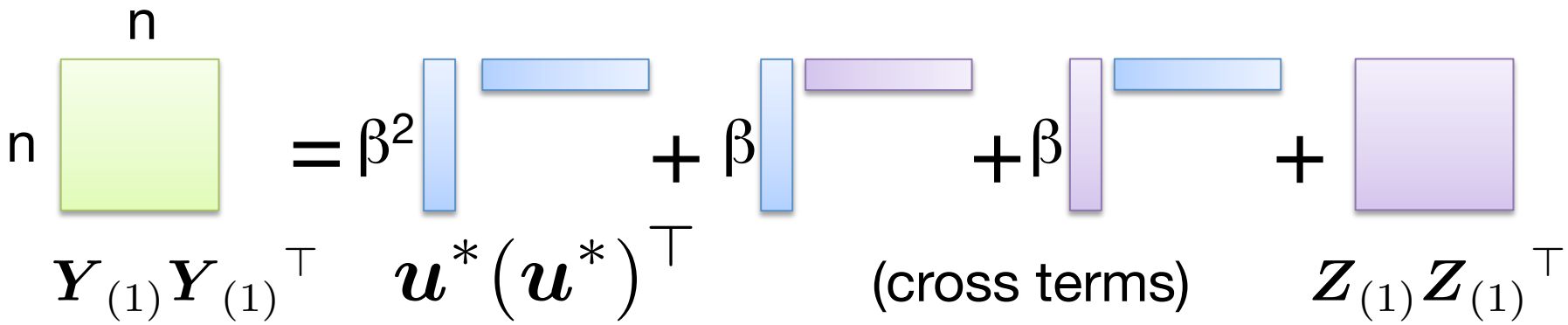
Why simpler method gets better bound

- R&M analyzes the perturbation of



$$\mathbb{E}\|\boldsymbol{Z}\| = O(\sqrt{n^{\lceil K/2 \rceil}})$$

- Z&T analyzes the perturbation of

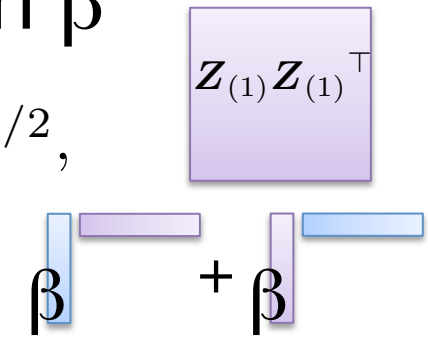


(better controlled perturbation!)

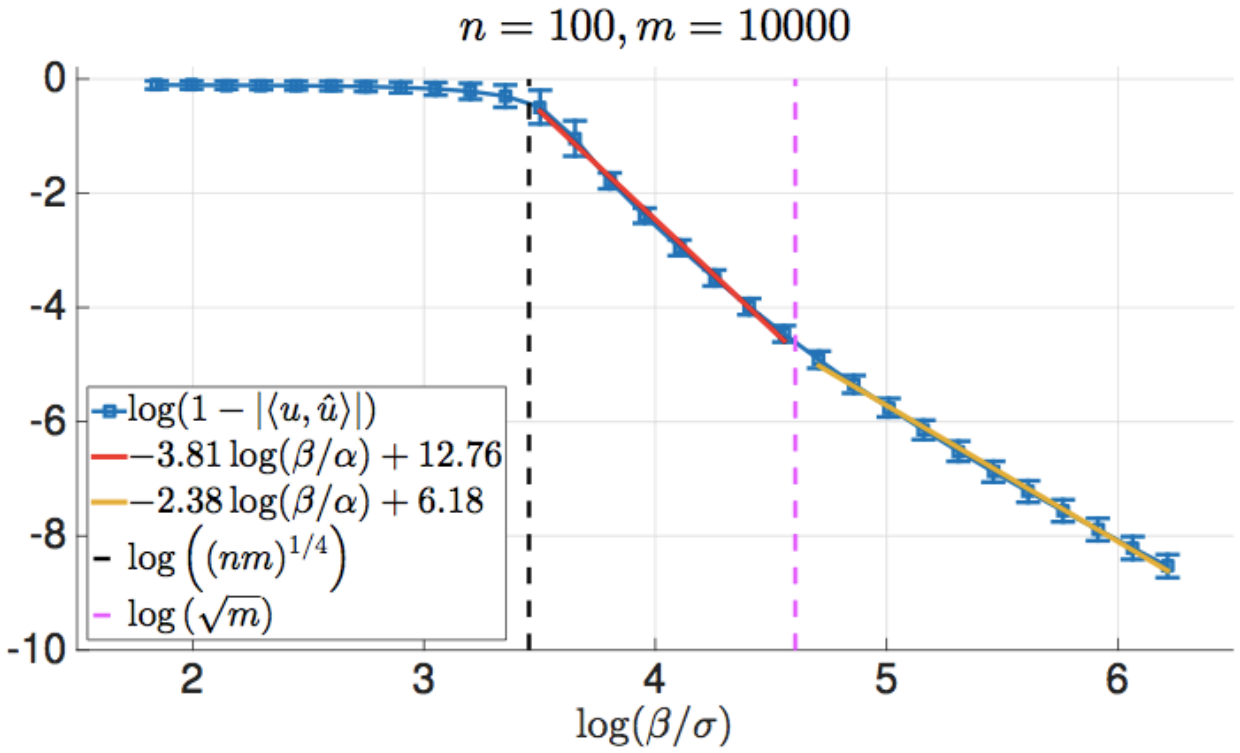
Two-phase behavior

- Because the cross term depends on β

$$1 - |\langle \hat{\boldsymbol{u}}, \boldsymbol{u}^* \rangle| = \begin{cases} O\left(\frac{n^K}{\beta^4}\right), & \text{if } c \cdot n^{K/4} \leq \beta \leq n^{(K-1)/2}, \\ O\left(\frac{n}{\beta^2}\right), & \text{if } n^{(K-1)/2} \leq \beta \end{cases}$$



(dominant term)



Summary of rank-one tensor recovery

- Ideal ML estimator: $\beta \geq c \cdot \sqrt{nK \log(K)}$
- Randomly initialized PI: $\beta \geq c' \cdot \sqrt{n^K K \log(K)}$
  - Note: in practice, it seems to be much more robust!
- (Recursive) unfolding: $\beta \geq c'' \cdot n^{K/4}$
- Questions:
  - Is there a fundamental limit how small β can be and still computationally tractable?
  - How about more general problems?

Barak & Moitra (2015) "Tensor Prediction, Rademacher Complexity and Random 3-XOR"

- Studied the atomic norm w.r.t

$$\mathcal{A}_{\text{rank1},\infty} = \left\{ \boldsymbol{u} \circ \boldsymbol{u} \circ \boldsymbol{u} : \|\boldsymbol{u}\|_\infty \leq \frac{C}{\sqrt{n}} \right\}$$

atoms are "balanced"

  for symmetric *n* x *n* x *n* tensors.

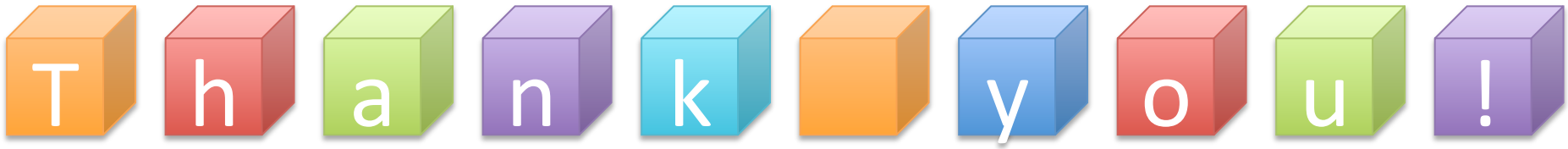
- For tensor completion, the atomic norm has almost optimal $O(n/\varepsilon^2)$ samples complextiy but intractable.
- Sixth order sum-of-squares relaxation yields $O(n^{K/2}/\varepsilon^2)$ sample complexity and polytime –$O(n^6)$ size SDP.
- Suggested no efficient algorithm can beat $O(n^{K/2}/\varepsilon^2)$.

Jain & Oh (2014) “Provable Tensor Factorization with Missing Data”

- Showed that alternating minimization algorithm with careful initialization achieves $O(r^5 n^{3/2}(\log(n))^4 \log(r\|W^*\|_F/\varepsilon))$ sample complexity for K=3.
  - Slightly high scaling with the rank (probably not optimized)
  - The same $O(n^{K/2})$ scaling
  - Note that if we consider $\sigma^2=1/\beta^2=1/m$ (inverse sample size), the same $m \geq c''n^{K/2}$ scaling for (recursive) unfolding.

Conclusion

- Tensors lie in the intersection of a lot of interesting areas –theory, data analysis, machine learning, and numerical linear algebra.
- Many tensor operations are known to be hard but often gradient descent/power iteration works in practice
  - Interaction between computation and statistics: can we trade statistical performance for tractability?
  - Need to understand hardness in contexts.
- Learning with tensors is an area already explored empirically (relational data, neural networks) but not much theoretically


Thank you!