Towards better computation-statistics trade-off in tensor decomposition

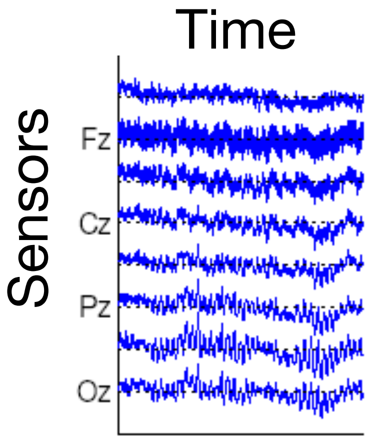
Ryota Tomioka
TTI Chicago

Joint work with: T. Suzuki, K. Hayashi, & H. Kashima

Matrices and Tensors in machine learning

Matrices

Multivariate time-series



Collaborative filtering

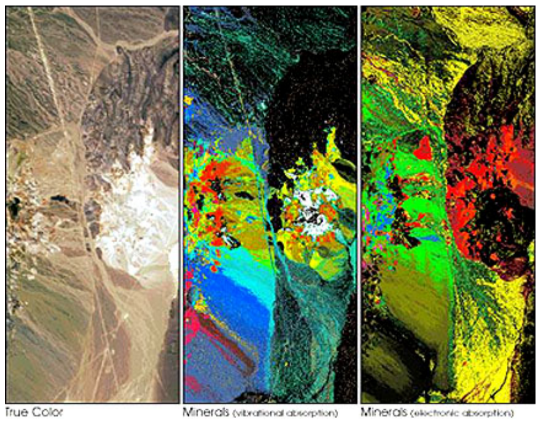
Movies

Users

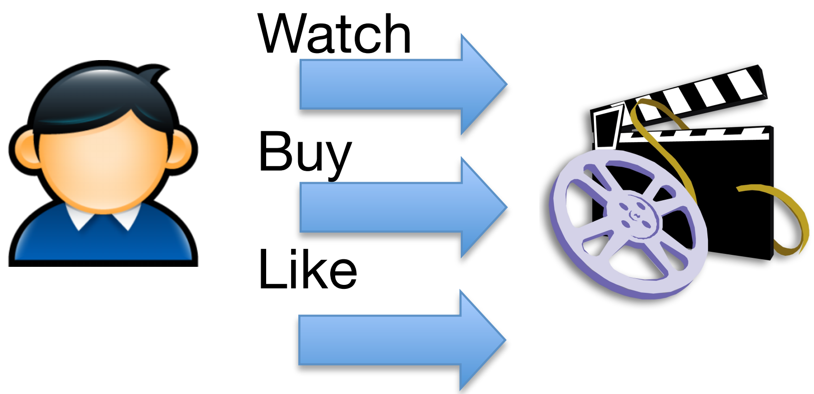
	Star Wars	Titanic	Blade Runner
User 1	5	2	4
User 2	1	4	2
User 3	5	?	?

Tensors

Spatio-temoral data



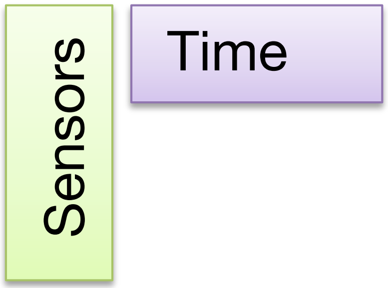
Multiple relations



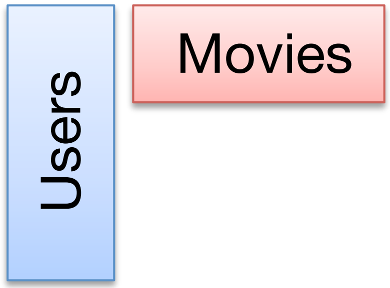
Matrices and Tensors in machine learning

Matrices

Multivariate time-series

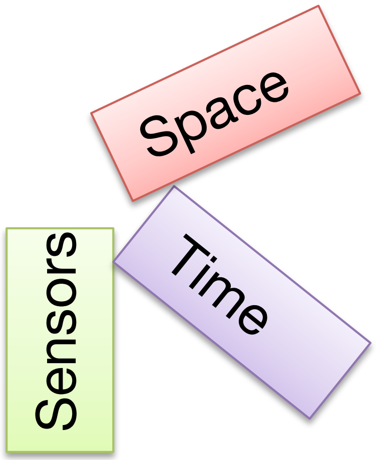


Collaborative filtering

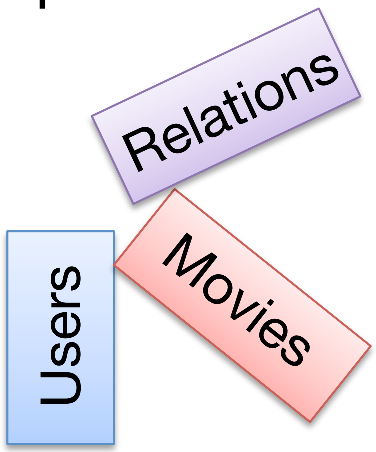


Tensors

Spatio-temoral data



Multiple relations



From matrices to tensors

- Trace norm: convex relaxation of matrix rank

$$\|\boldsymbol{W}\|_{S_1} = \sum_{j=1}^{r} \sigma_j(\boldsymbol{W})$$

Induces low-rank-ness (spectral sparsity)

  - It works like L1 regularization on the singular values
  - Performance guarantees [Srebro & Schraibman 2005; Candes & Recht 2009; Candes & Tao 2010; Negahban & Wainwright 2011]

Similar relaxation possible for tensor rank?

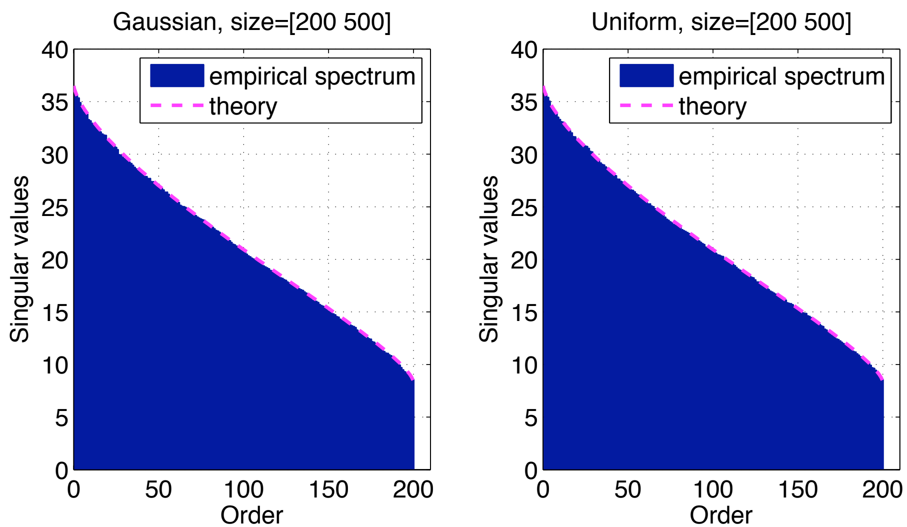
From matrices to tensors

- Spectral norm of random Gaussian matrix

$$\mathbb{E}\|\boldsymbol{X}\|_{S_\infty} \leq \sigma\left(\sqrt{m}+\sqrt{n}\right)$$

- Marchenko-Pastur distribution

[Marchenko & Pastur 1967]



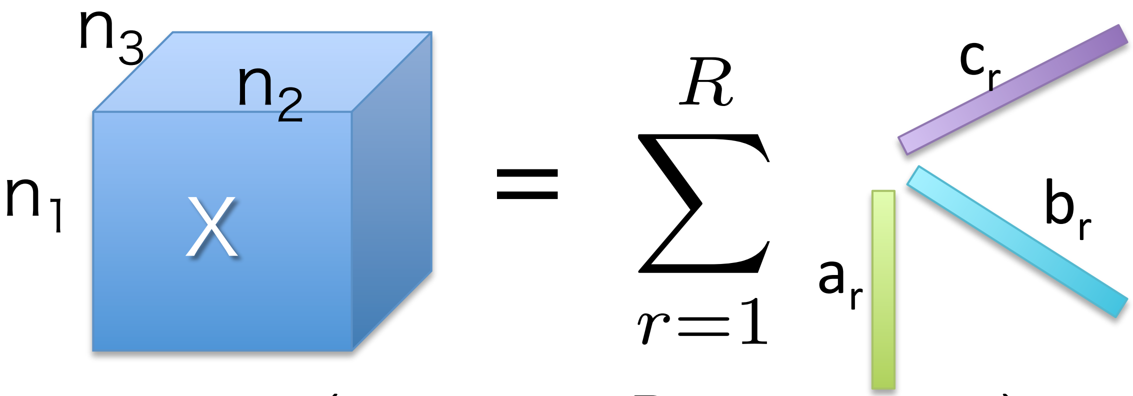
Random *tensor* theory?

Outline

- Tensor ranks and decompositions
- Overlapped trace norm (moderate computation)
  - Limitations: requires $O(rn^{K-1})$ samples
- Balanced trace norm (heavy computation) [Mu et al. 2013]
  - requires $O(r^{K/2}n^{K/2})$ samples
- Tensor trace norm (probably intractable)
  - requires only $O(rn)$ samples

Tensor rank

- Minimum number R such that



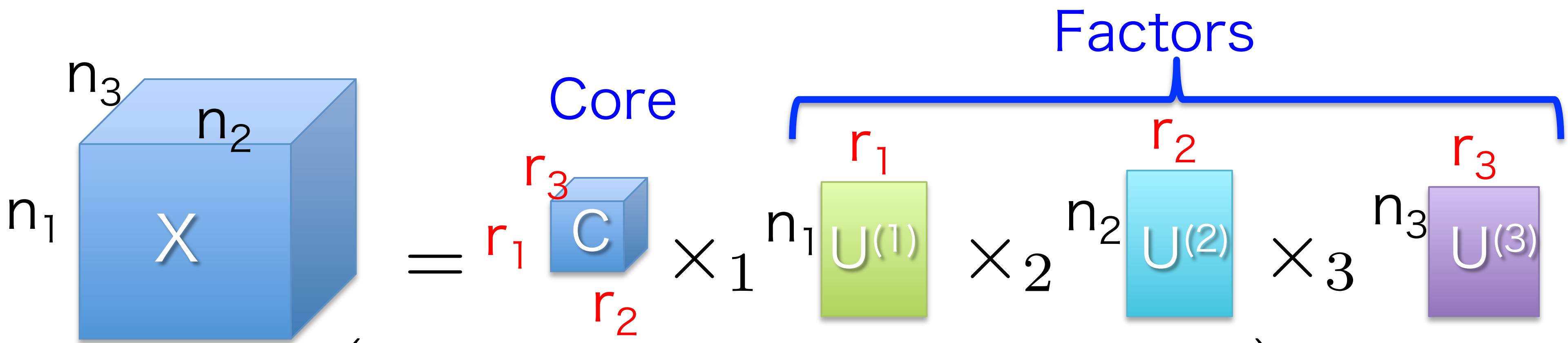
$$\left(X_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} \right)$$ (for 3rd order tensor)

- Known as CP (canonical polyadic) decomposition
  [Hitchcock 27; Carroll & Chang 70; Harshman 70]

- Comutation of the above decomposition is NP hard!

Tucker decomposition

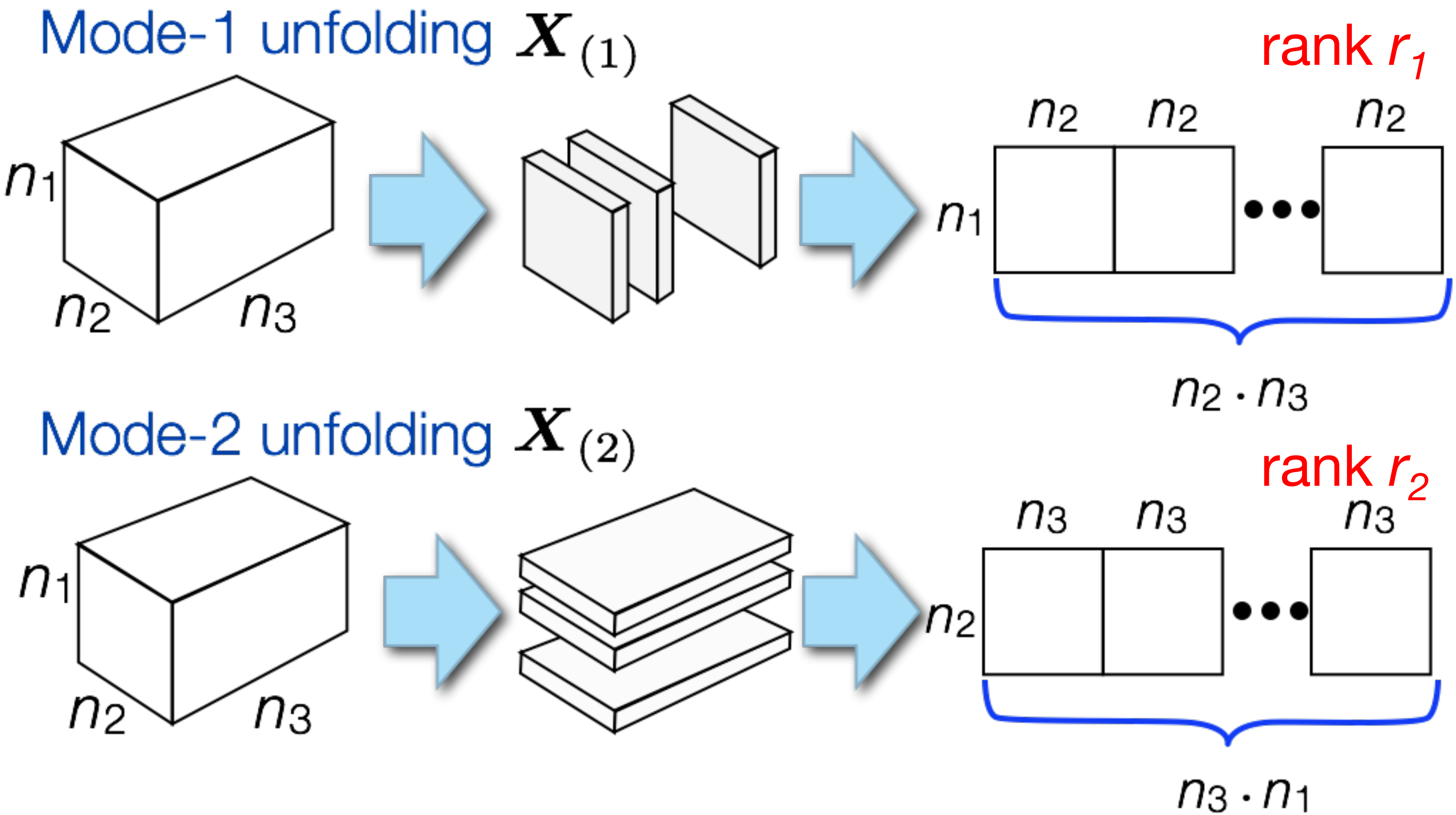
[Tucker 66; De Lathauwer+00]



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

- Factors can be obtained by unfolding operation+SVD
- In practice no unfolding is low-rank --- Common solution: iterate truncated SVD (HOSVD, HOOI); non-convex

Unfolding (matricization)



Core idea

Unfolding
(Matricization)

Tensor X is low rank
$\exists\, k,\ r_k < n_k$
(in the sense of Tucker decomposition)

Unfolding $X_{(k)}$
is low-rank
(as a matrix)

Tensorization

Overlapped trace norm

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

- Convex optimization problem

$$\underset{\mathcal{W}\in\mathbb{R}^{n_1\times\cdots\times n_K}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y}-\mathfrak{X}(\mathcal{W})\|^2+\lambda_M |||\mathcal{W}|||_{\underline{S_1/1}}$$

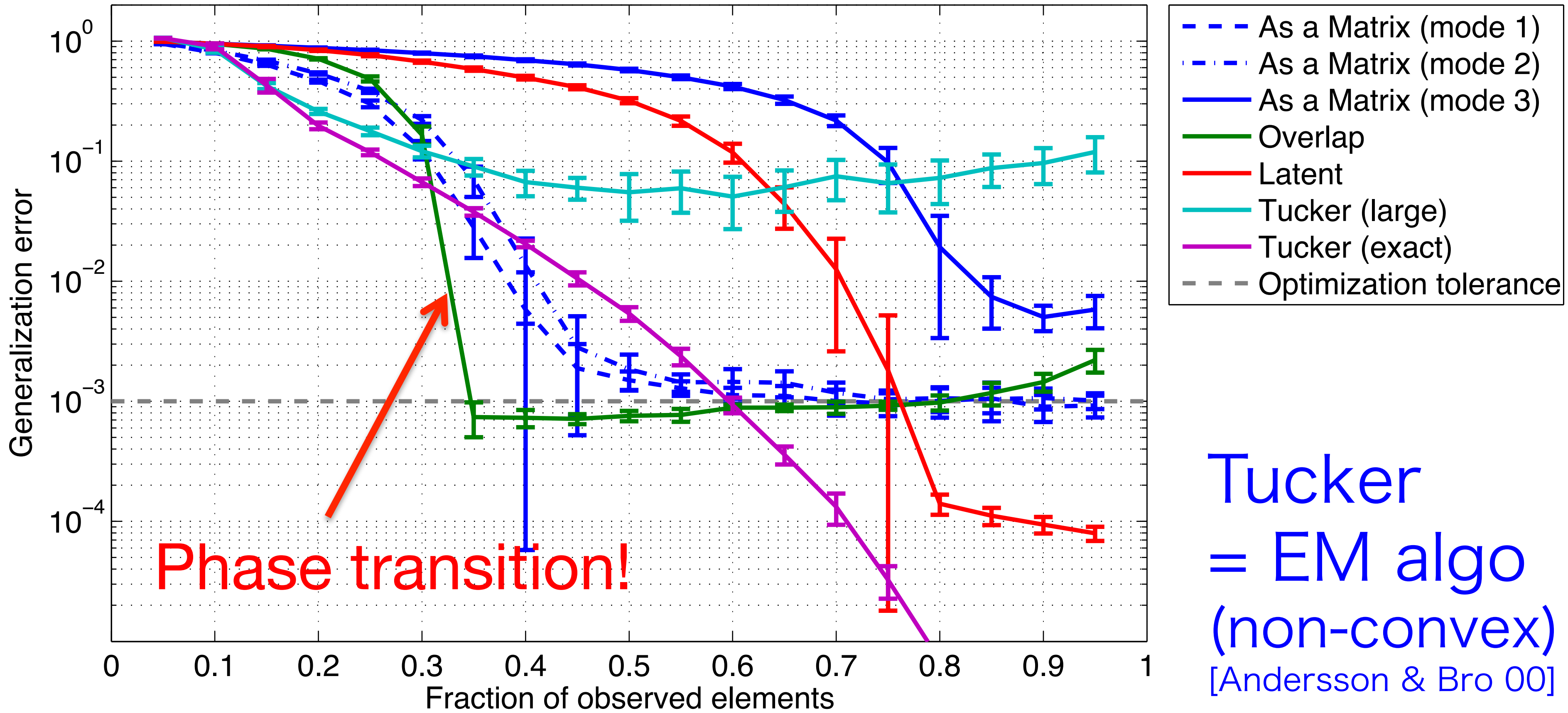
where $$|||\mathcal{W}|||_{\underline{S_1/1}} := \sum_{k=1}^{K}\|\boldsymbol{W}_{(k)}\|_{S_1}$$

mode-*k* unfolding

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

Empirical performance

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



Tucker
= EM algo
(non-convex)
[Andersson & Bro 00]

Analysis: Problem setting

Observation

$\mathcal{W}^*$: true tensor with rank ($r_1,\ldots,r_K$)

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

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

Optimization

Likelihood Regularization

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

Reg. constant

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

Observation operator $\mathfrak{X} : \mathbb{R}^N \to \mathbb{R}^M$

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

Theorem ("overlapped" approach)

[T, Suzuki, Hayashi, Kashima 11]

Assume that the elements of the design X are independently and identically Gaussian distributed. Moreover, if

$$\frac{\#\text{samples } (M)}{\#\text{variables } (N)} \geq c_1 \underbrace{\|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}}_{\text{normalized rank}} \approx \frac{r}{n}$$

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

Theorem (random Gauss design)

[T, Suzuki, Hayashi, Kashima 11]

Assume that the elements of the design X are independently and identically Gaussian distributed. Moreover, if

$$\frac{\#\text{samples } (M)}{\#\text{variables } (N)} \geq c_1 \underbrace{\|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}}_{\text{normalized rank}} \approx \frac{r}{n}$$

Convergence!

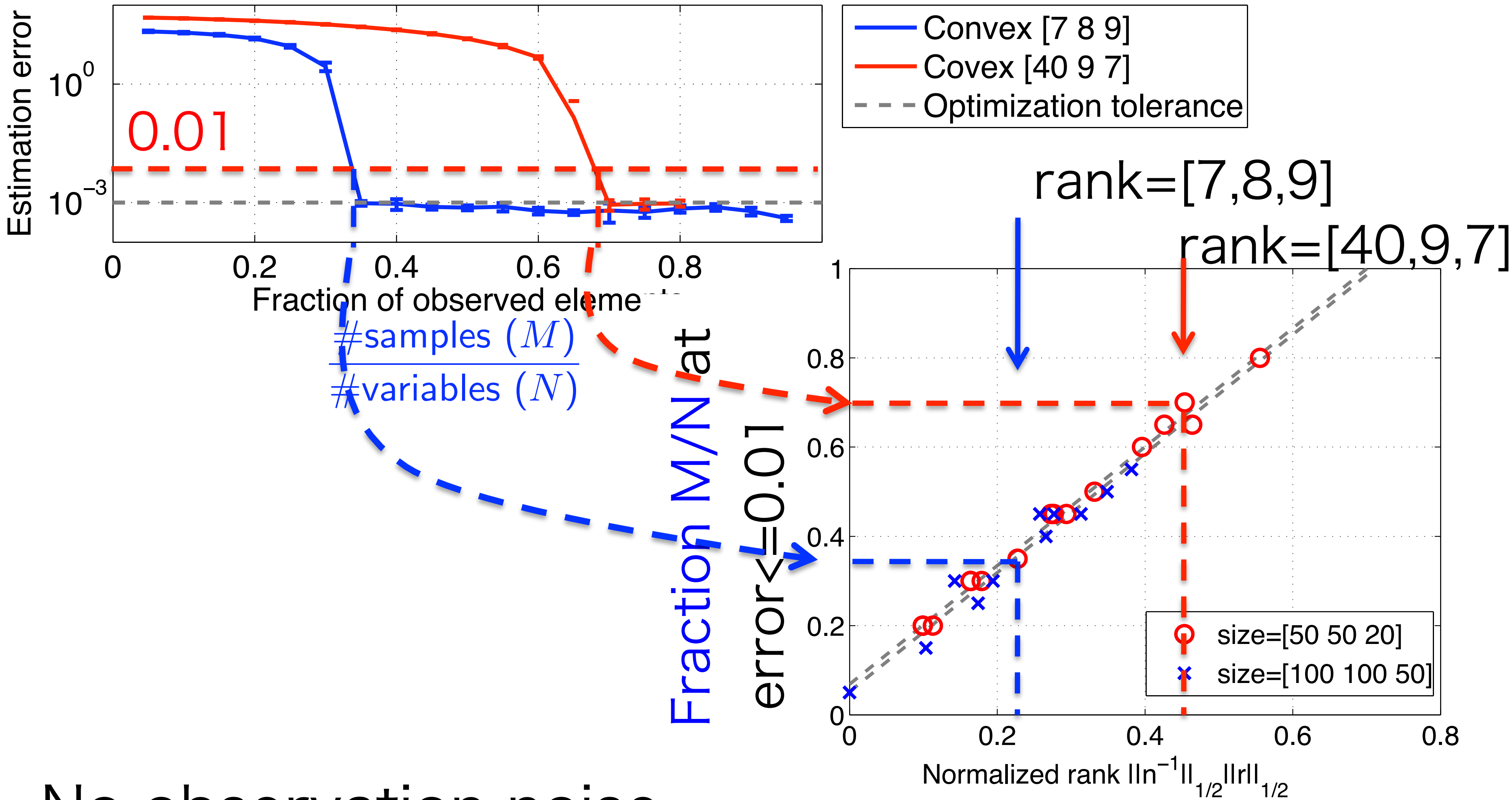
$$\frac{\left\|\hat{\mathcal{W}} - \mathcal{W}^*\right\|_F^2}{N} \leq O_p\left(\frac{\sigma^2\|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}}{M}\right)$$

(with appropriate choice of λ_M)

$$\|\boldsymbol{n}^{-1}\|_{1/2} := \left(\tfrac{1}{K}\textstyle\sum_{k=1}^{K}\sqrt{1/n_k}\right)^2, \quad \|\boldsymbol{r}\|_{1/2} := \left(\tfrac{1}{K}\textstyle\sum_{k=1}^{K}\sqrt{r_k}\right)^2$$

Tensor completion

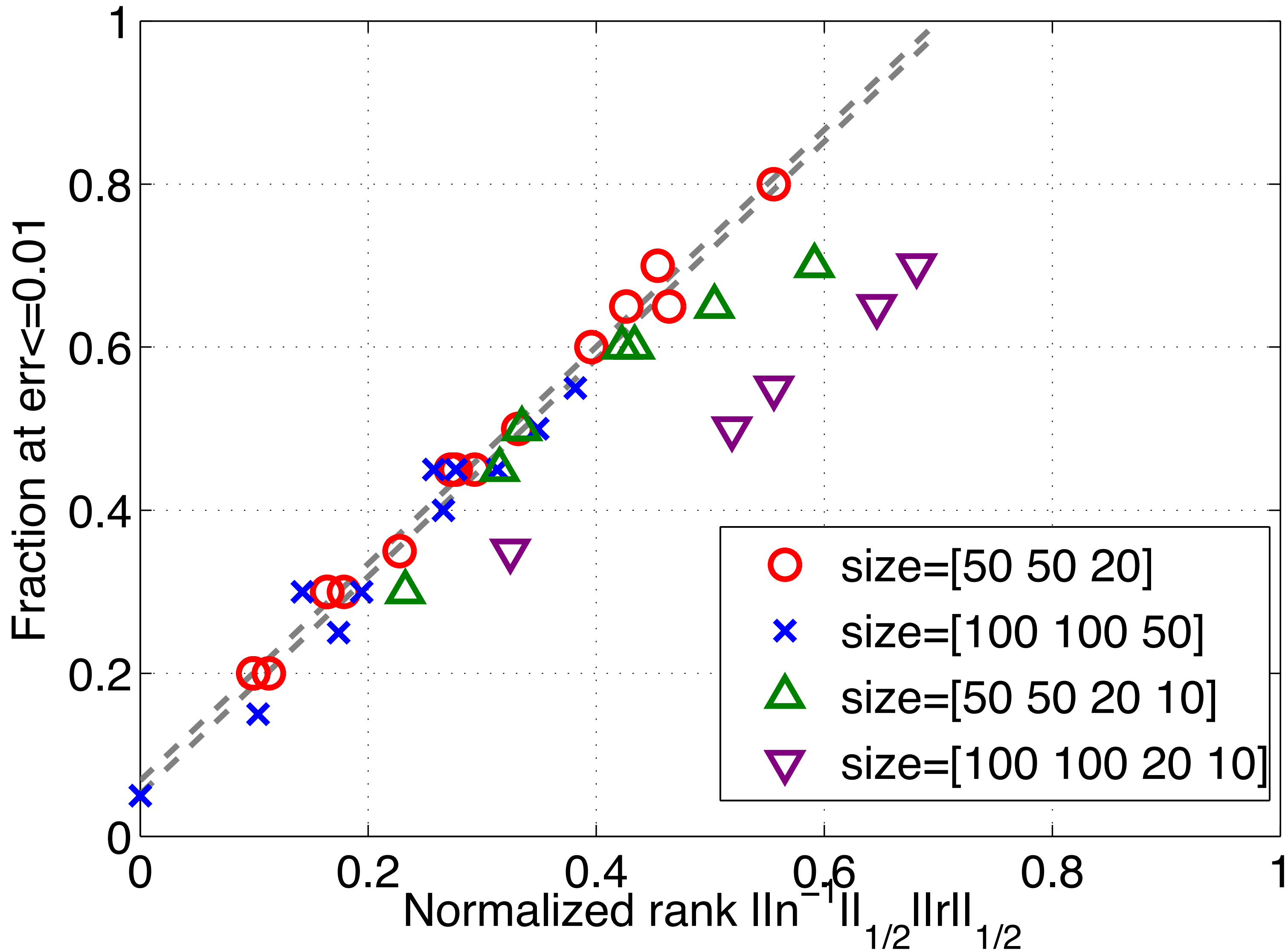
size = 50x50x20 true rank 7x8x9 or 40x9x7



No observation noise

Normalized rank

Theory vs. Experiments (4th order)



Limitation: exponentially many samples required!

- Simplify by setting $n_k = n$ and $r_k = r$
- Then there are constants c0, c1, c2 such that
  - #samples $M \geq c_1 n^{K-1} r$
  - reg. const. $\lambda_M = c_0 \sigma \sqrt{n^{K-1}/M}$

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

with high probability.

Why?

- Key steps in the analysis
  - Relation between the norm and the rank

$$\left\|\!\left|\mathcal{W}\right|\!\right\|_{\underline{S_1/1}} \leq K\sqrt{r}\left\|\!\left|\mathcal{W}\right|\!\right\|_F \qquad \text{(OK)}$$

  - Dual norm of noise tensor

$$\mathbb{E}\left\|\!\left|\mathfrak{X}^\top(\boldsymbol{\epsilon})\right|\!\right\|_{(\underline{S_1/1})^*} \leq \frac{\sigma\sqrt{M}}{K}\left(\sqrt{n^{K-1}} + \sqrt{n}\right)$$

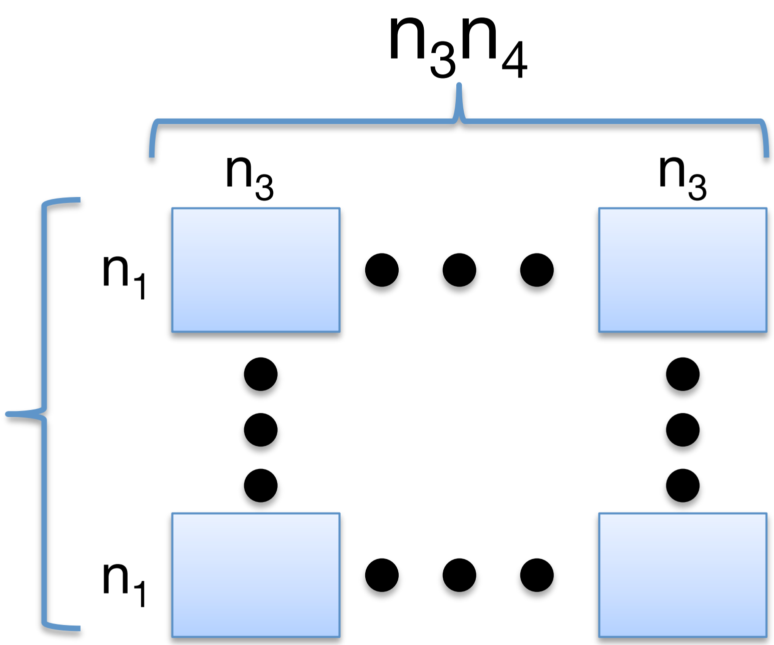
unbalanced (Bad)

where $\mathfrak{X}^\top(\boldsymbol{\epsilon}) := \sum_{i=1}^{M} \epsilon_i \mathcal{X}_i$

Balanced unfolding

- For K>3, there are 2^{K-1}-1 > K ways to unfold a tensor. For example,

$$\boldsymbol{X}_{(1,2;3,4)} = n_1 n_2$$



(See also Mu et al. 2013)

Balanced trace norm (for K=4)

- Definition

$$|||\mathcal{W}|||_{\mathrm{balanced}} := \|\boldsymbol{W}_{(1,2;3,4)}\|_{S_1} + \|\boldsymbol{W}_{(1,3;2,4)}\|_{S_1} + \|\boldsymbol{W}_{(1,4;2,3)}\|_{S_1}$$

– Relation between the norm and the rank

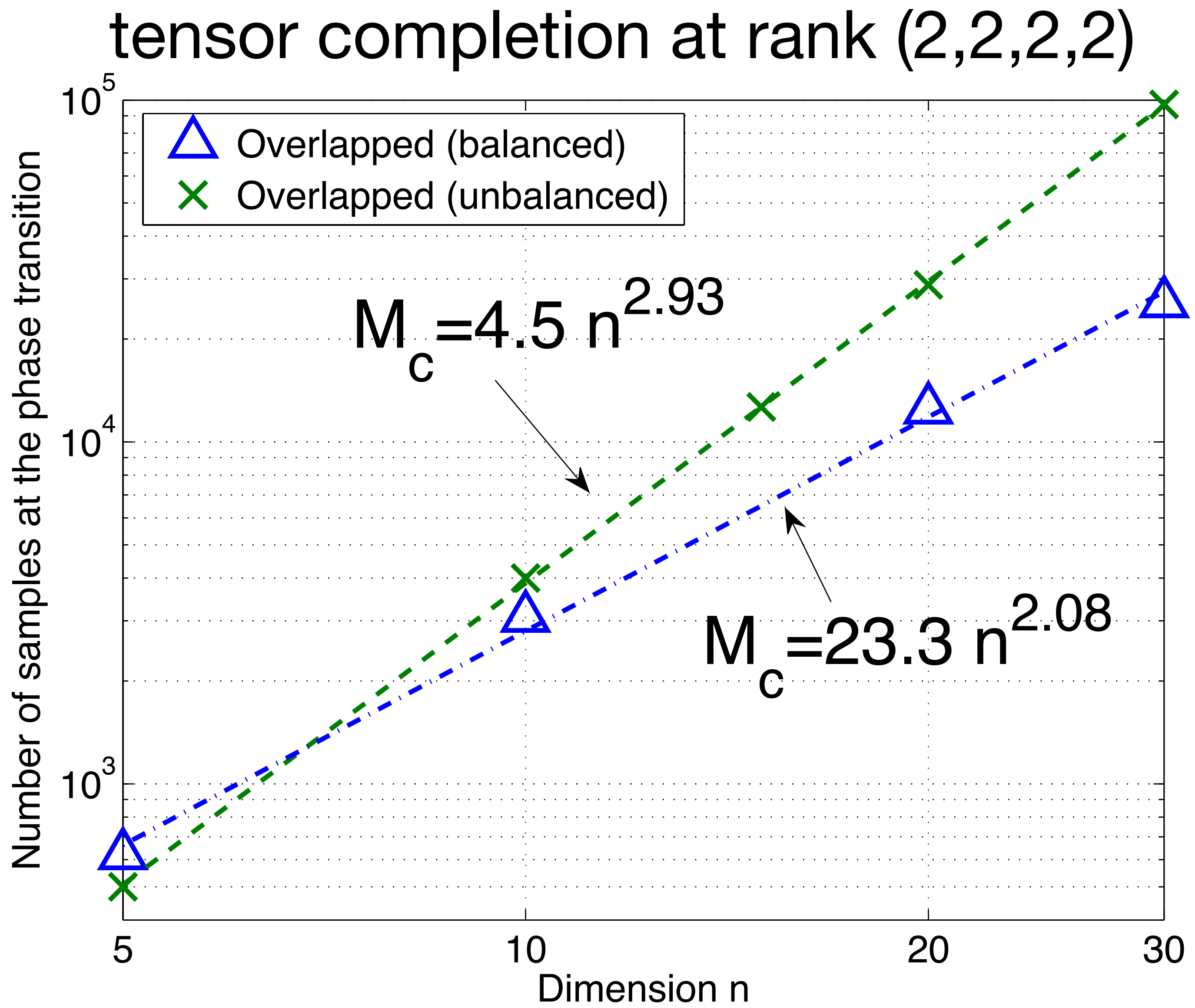
$$|||\mathcal{W}|||_{\mathrm{balanced}} \leq 3\sqrt{r^2}|||\mathcal{W}|||_F$$

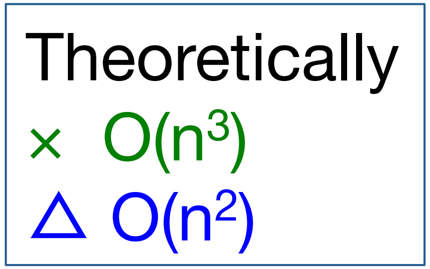
– Dual norm of noise tensor

$$\mathbb{E}|||\mathfrak{X}^\top(\boldsymbol{\epsilon})|||_{\mathrm{balanced}^*} \leq \frac{\sigma\sqrt{M}}{3} \cdot 2\sqrt{n^2}$$

Sample complexity $O(r^2 n^2)$

Experiment (K=4)





Comparison of computational complexity

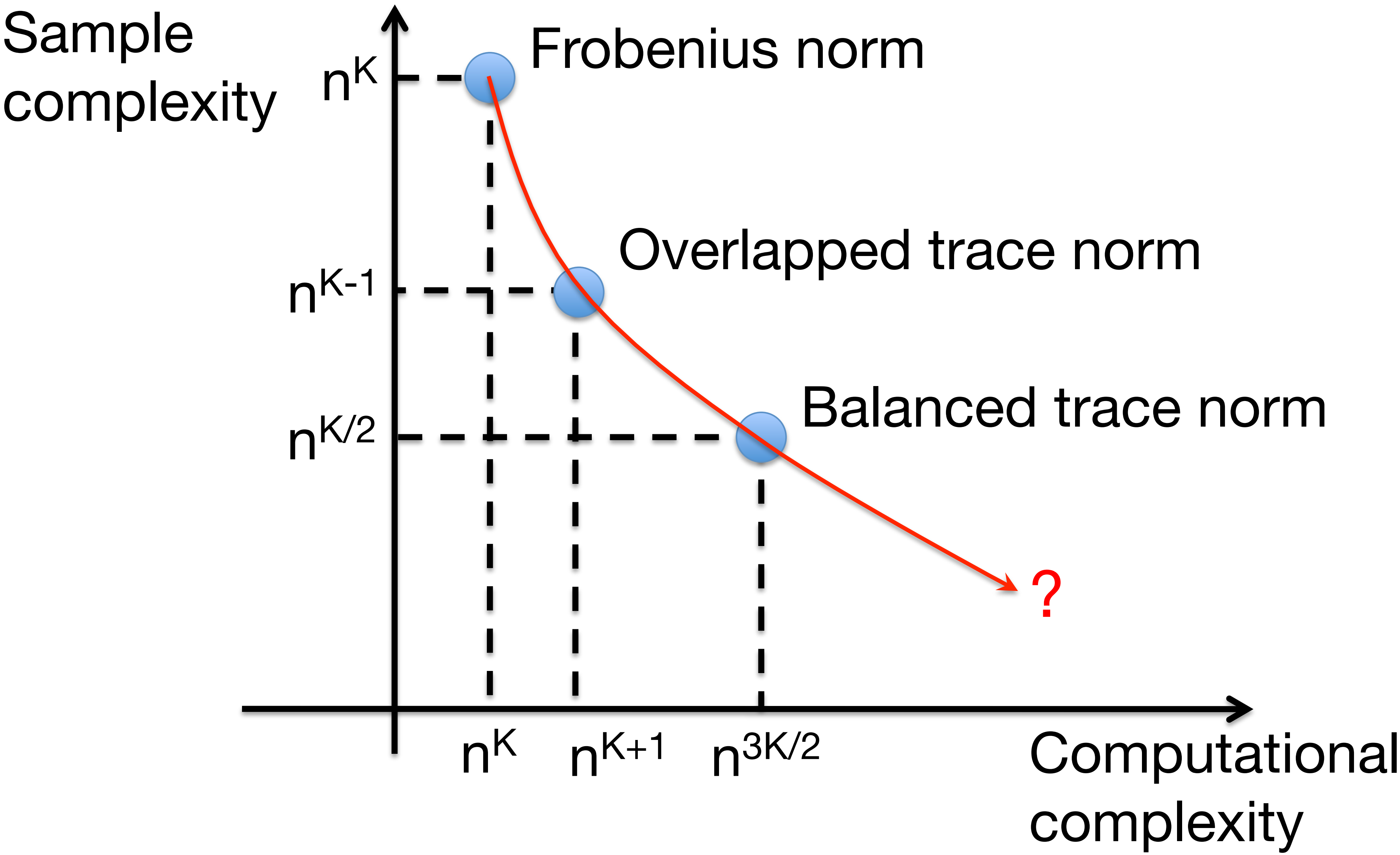
- Overlapped trace norm (Sample Complex. $O(rn^{K-1})$)
  - requires SVD of n^{K-1} x n matrix: Large!

    $O(n^{K+1}+n^3)$ $\Rightarrow$ $O(n^5)$ for $K=4$ OK
- Balanced trace norm (Sample Complex. $O(r^{K/2}n^{K/2})$)
  - requires SVD of $n^{K/2}$ x $n^{K/2}$ matrix: OK

    $O(n^{1.5K})$ $\Rightarrow$ $O(n^6)$ for $K=4$ Large!

statistically more efficient, computationally more challenging!

Computation-statistics trade-off



Tensor trace norm

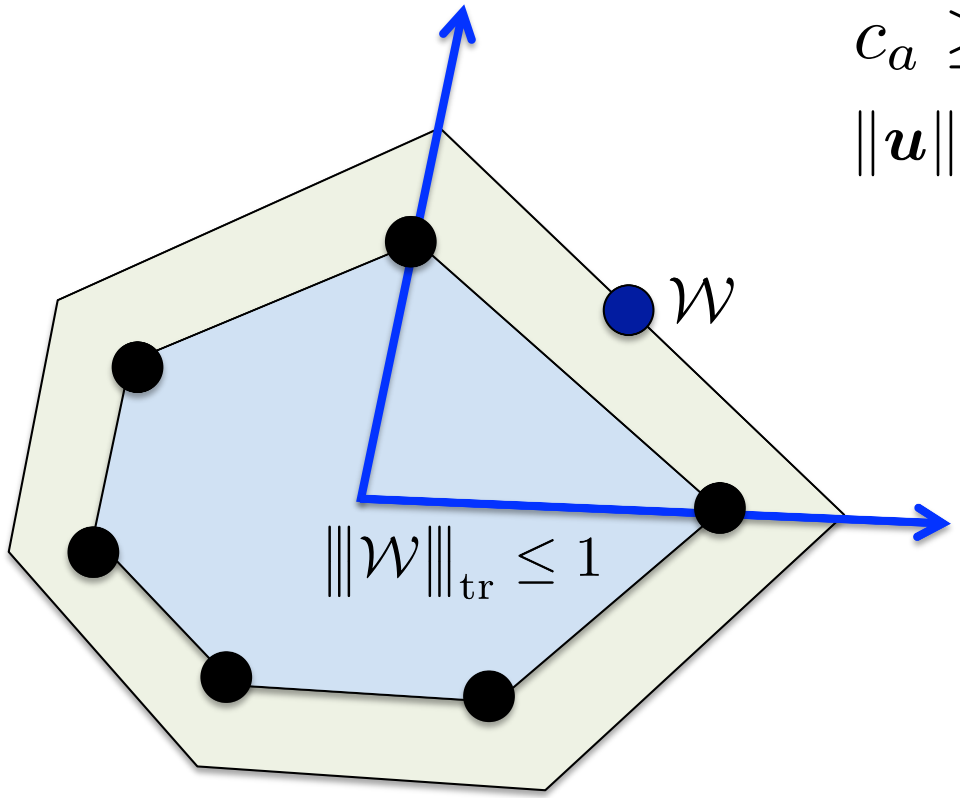
For K=3

$$\|\!|\mathcal{W}|\!\|_{\mathrm{tr}} = \inf \sum_{a \in \mathcal{A}} c_a \quad \text{s.t.} \quad \mathcal{W} = \sum_{a \in \mathcal{A}} c_a \boldsymbol{u}_a \circ \boldsymbol{v}_a \circ \boldsymbol{w}_a$$

$$c_a \geq 0$$

$$\|\boldsymbol{u}\| \leq 1,\ \|\boldsymbol{v}\| \leq 1,\ \|\boldsymbol{w}\| \leq 1$$

rank-1 tensor (outer prod. of vectors)



can be seen as an atomic norm [Chandrasekaran 12] with atomic set = set of rank-1 tensors

Tensor trace norm

For K=3

$$\left\|\!\left|\mathcal{W}\right|\!\right\|_{\mathrm{tr}} = \inf \sum_{a\in\mathcal{A}} c_a \quad \text{s.t.} \quad \mathcal{W} = \sum_{a\in\mathcal{A}} c_a \boldsymbol{u}_a \circ \boldsymbol{v}_a \circ \boldsymbol{w}_a$$

$$c_a \geq 0$$

$$\|\boldsymbol{u}\| \leq 1,\ \|\boldsymbol{v}\| \leq 1,\ \|\boldsymbol{w}\| \leq 1$$

Relation between the norm and the orthogonal CP rank

(Kolda 2001)

$$\left\|\!\left|\mathcal{W}\right|\!\right\|_{\mathrm{tr}} \leq \sqrt{R}\left\|\!\left|\mathcal{W}\right|\!\right\|_{F}$$

Dual norm of the noise tensor

$$\mathbb{E}\left\|\!\left|\mathfrak{X}^{\top}(\boldsymbol{\epsilon})\right|\!\right\|_{\mathrm{tr}^*} \leq C\sigma\sqrt{M}\sqrt{n}$$

Sample complexity O(Rn)

Dual of the trace norm is the *tensor operator norm*

$$\left\|\!\left|\mathcal{Y}\right|\!\right\|_{\mathrm{tr}^*} = \left\|\!\left|\mathcal{Y}\right|\!\right\|_{\mathrm{op}} := \sup_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}} \sum_{i,j,k} Y_{ijk} u_i v_j w_k$$
$$\text{s.t. } \|\boldsymbol{u}\| \leq 1,\ \|\boldsymbol{v}\| \leq 1,\ \|\boldsymbol{w}\| \leq 1$$

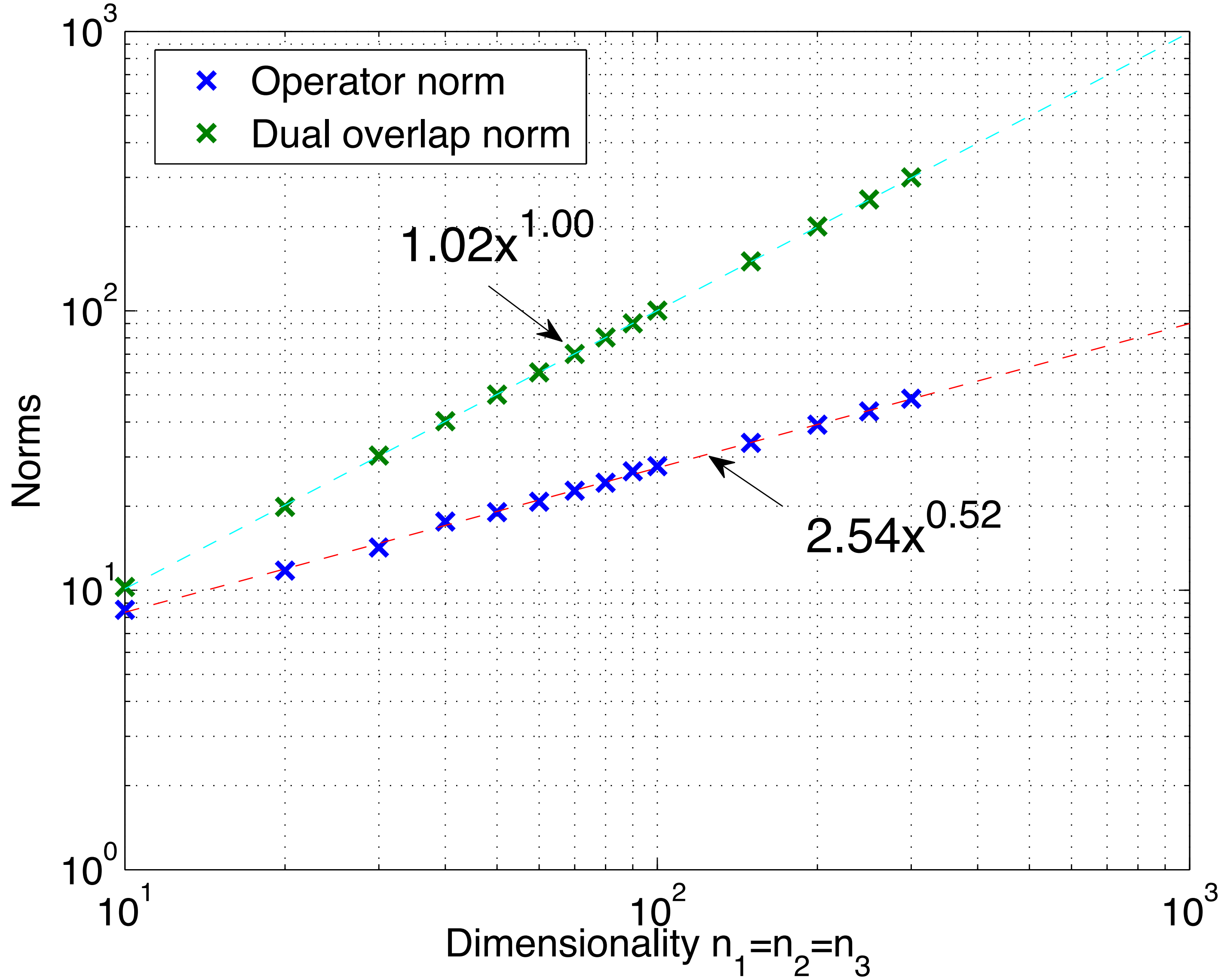
Greedy algorithm for computing the operator norm

1. Initialize u, v, w.
2. Fix u, maximize over v and w (matrix operator norm)
3. Cycle over v, w, u, … until convergence

(can be improved by incorporating gradient)

Empirical scaling (K=3)

Theoretically
× O(n)
x O(√n)



Low-rank tensor estimation with the *tensor trace norm*

Likelihood Regularization

$$\underset{\mathcal{W}\in\mathbb{R}^{n_1\times\cdots\times n_K}}{\text{minimize}} \quad \boxed{\frac{1}{2}\|\boldsymbol{y}-\mathfrak{X}(\mathcal{W})\|^2} + \boxed{\lambda_M \||\mathcal{W}\||_{\mathrm{tr}}}$$

Key operation: prox operator

$$\mathrm{prox}_\lambda(\mathcal{W}) = \underset{\mathcal{Y}}{\mathrm{argmin}}\left(\lambda\||\mathcal{Y}\||_{\mathrm{tr}} + \frac{1}{2}\||\mathcal{Y}-\mathcal{W}\||_F^2\right)$$

$$= \mathcal{W} - \mathrm{proj}_\lambda(\mathcal{W})$$ (Moreau's theorem)

$$\mathrm{proj}_\lambda(\mathcal{W}) = \underset{\mathcal{Y}}{\mathrm{argmin}}\||\mathcal{W}-\mathcal{Y}\||_F \quad \text{s.t.} \quad \||\mathcal{Y}\||_{\mathrm{op}} \le \lambda$$

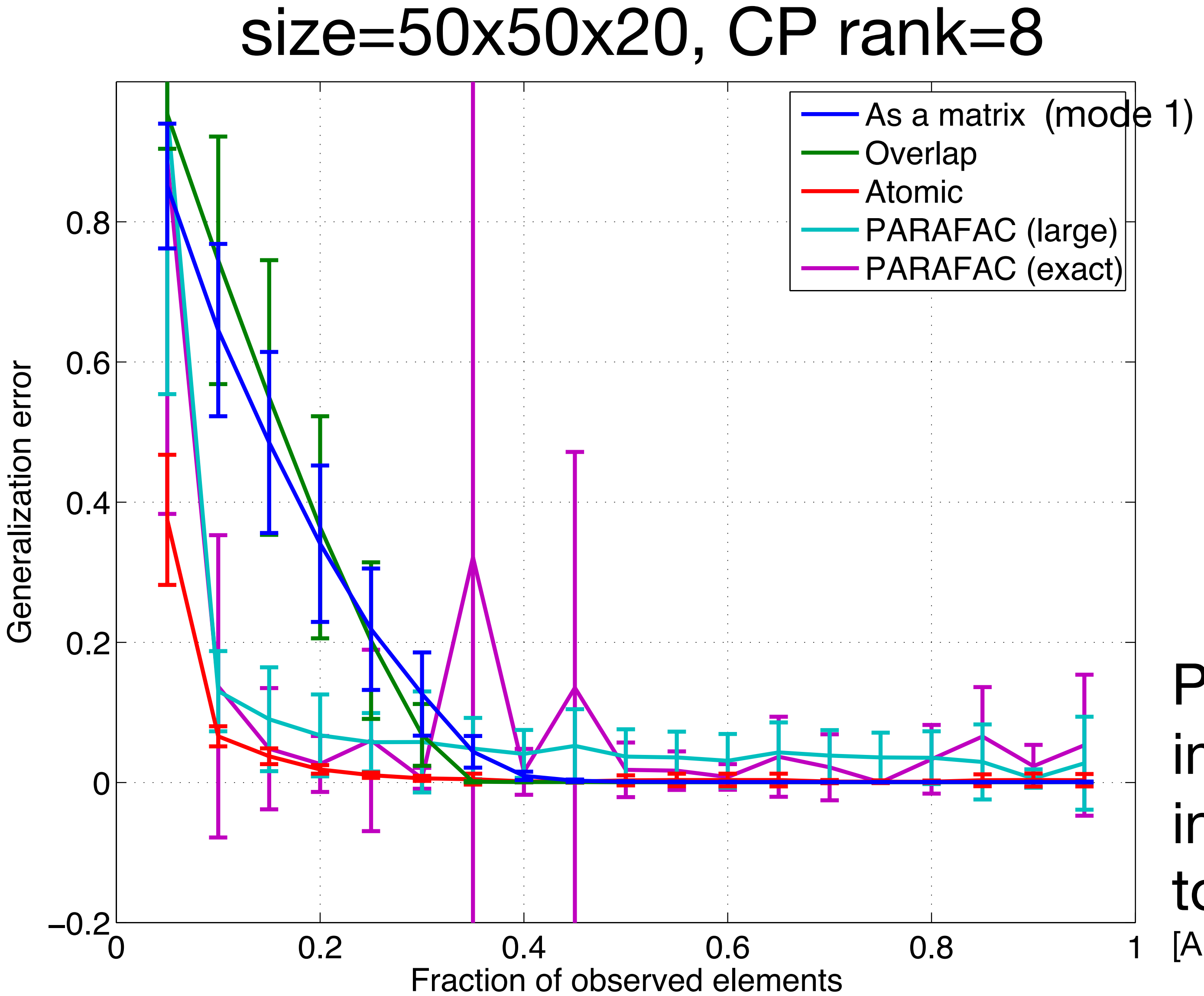
Tensor operator norm

Greedy algorithm for $\text{prox}_{\lambda}(W)$

1. Let R=W.
2. Compute $||R||_{op}$
   if $||R||_{op} \leq \lambda$, done. Return W-R
   otherwise, $R=R+(\lambda-||R||_{op})\, u \cdot v \cdot w$
3. Go to 2.

Tensor completion experiment

(λ→0)



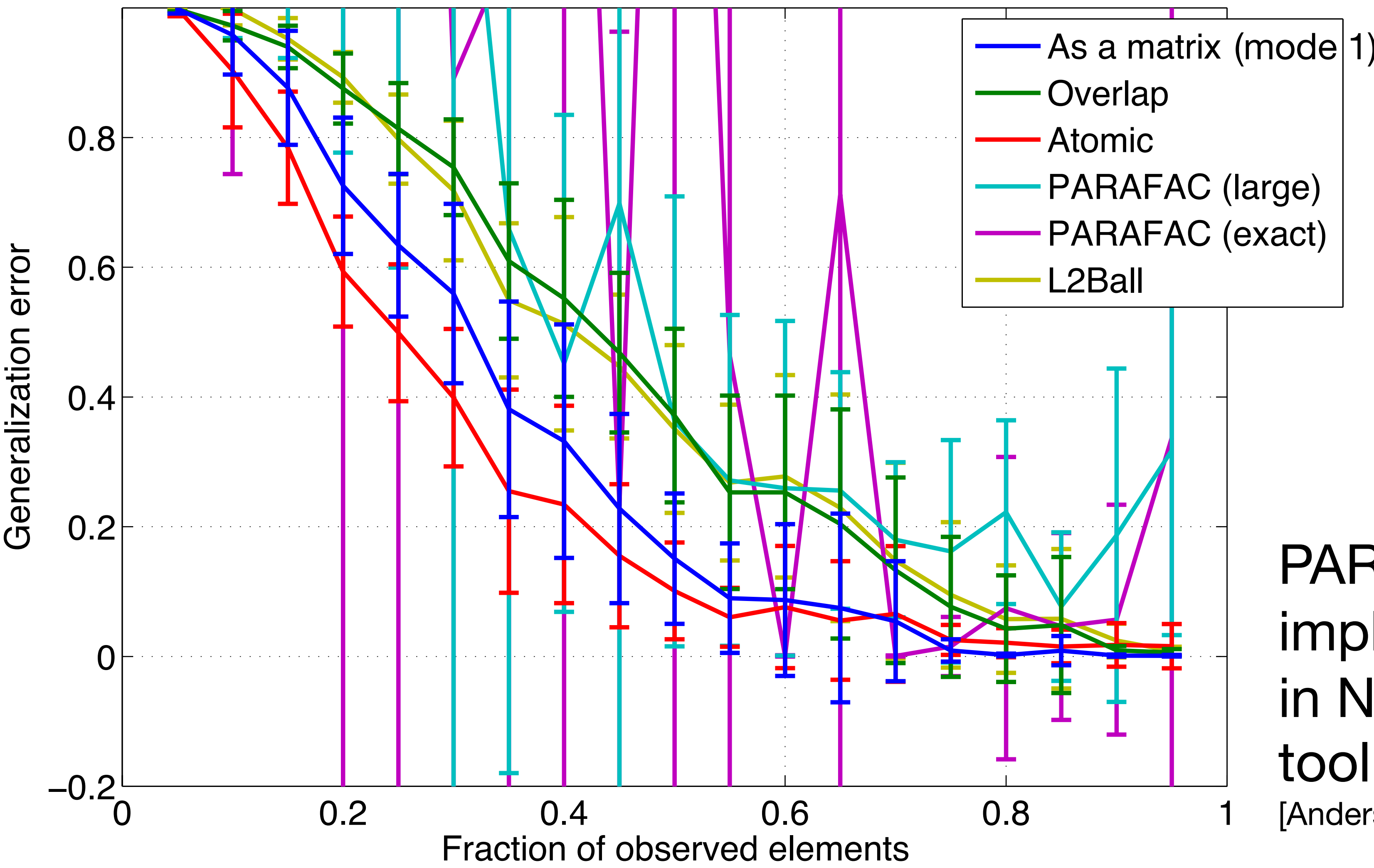
PARAFAC implemented in N-way toolbox

[Andersson & Bro 00]

Balanced vs. unbalanced

($\lambda \to 0$)

size=25x5x5, CP rank=3



balanced 25x25

PARAFAC implemented in N-way toolbox

[Andersson & Bro 00]

Summary

- Tensor decomposition via convex optimization
  - Fast and stable algorithm for tensor decomposition
  - Rank selection is replaced by regularization parameter selection
- Limitation of the overlapped trace norm
  - unbalancedness of the unfolding
  - balanced unfolding
- Optimization statistics trade-off
  - balanced trace norm requires less samples but more computation
  - tensor trace norm requires only O(n) samples but seems intractable

References

- Andersson and Bro. (2000) The n-way toolbox for matlab. *Chemometrics & Intelligent Laboratory Systems*, 52(1):1–4, 2000. http://www.models.life.ku.dk/source/nwaytoolbox/.
- Chandrasekaran, Recht, Parrilo, and Willsky. (2012) The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849.
- Kolda & Bader (2009) Tensor Decompositions and Applications. *SIAM Review*.
- Gandy, Recht, and Yamada. (2011) Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27:025010.
- Håstad. (1990) Tensor rank is NP-complete. Journal of Algorithms, 11(4):644– 654.
- Mu, Huang, Wright, and Goldfarb. (2013) Square deal: Lower bounds and improved relaxations for tensor recovery. arXiv preprint arXiv:1307.5870.
- Signoretto, De Lathauwer, and Suykens. (2010) Nuclear norms for tensors and their use for convex multilinear estimation. Technical Report 10-186, ESAT-SISTA, K.U.Leuven.
- Tomioka, Suzuki, Hayashi, and Kashima. (2011) Statistical performance of convex tensor decomposition. In *Advances in NIPS* 24, pages 972–980.
- Tomioka and Suzuki. (2013) Convex tensor decomposition via structured schatten norm regularization. In *Advances in NIPS* 26, pages 1331–1339.
- Tomioka, Suzuki, Hayashi, & Kashima. (2014) Low-Rank Tensor Denoising and Recovery via Convex Optimization. In Suykens, Signoretto, & Argyriou, editors, *Regularization, Optimization, Kernels, and Support Vector Machines.* To be published from CRC Press.