Spiked tensor models through the prism of random matrix theory

J. Henrique de M. Goulart

Toulouse INP / IRIT

henrique.goulart@irit.fr

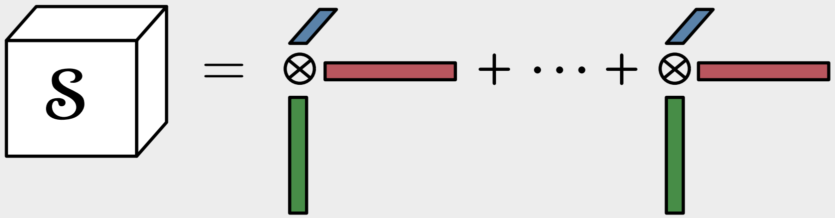
Marseille – May 2024

Canonical polyadic decomposition (CPD)

CPD: Widely used for information extraction & unsupervised learning

$$\boldsymbol{\mathcal{S}} = \sum_{r=1}^{R} \lambda_r \, \boldsymbol{x}_r \otimes \boldsymbol{y}_r \otimes \boldsymbol{z}_r$$

$$S_{ijk} = \sum_{r=1}^{R} \lambda_r \, x_{ir} \, y_{jr} \, z_{kr}$$

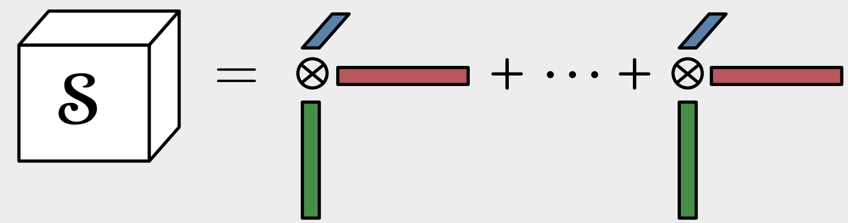


Key property: essential uniqueness (without orthogonality)

Canonical polyadic decomposition (CPD)

CPD: Widely used for information extraction & unsupervised learning

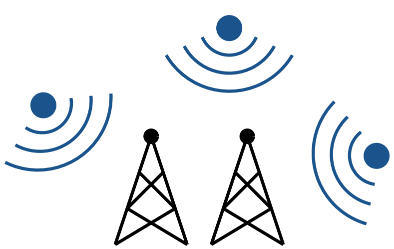
$$\boldsymbol{\mathcal{S}} = \sum_{r=1}^{R} \lambda_r \, \boldsymbol{x}_r \otimes \boldsymbol{y}_r \otimes \boldsymbol{z}_r$$

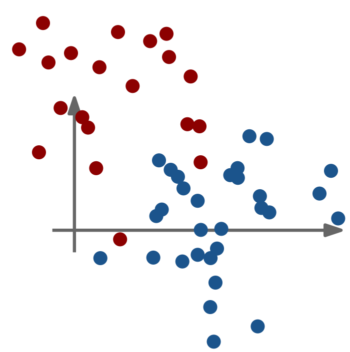


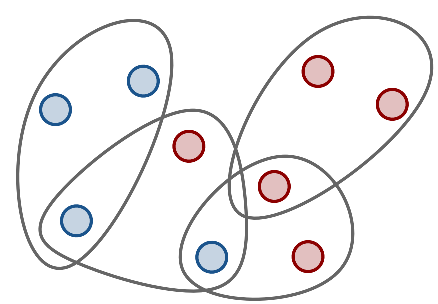
$$S_{ijk} = \sum_{r=1}^{R} \lambda_r \, x_{ir} \, y_{jr} \, z_{kr}$$

Key property: essential uniqueness (without orthogonality)

Applications: source separation,[1] Gaussian mixture estimation,[2] community detection,[3] etc







1. (Comon & Jutten, 2010), 2. (Anandkumar et al., 2014), 3. (Wang et al., 2023)

In practice: approximate CPD

Most ofen, we wish to recover a CPD from noisy observations:

$$\underbrace{\mathcal{Y} = \sum_{r=1}^{R} \lambda_r \, \boldsymbol{x}_r \otimes \boldsymbol{y}_r \otimes \boldsymbol{z}_r}_{\text{signal / information ("spike")}} + \underbrace{\mathcal{W}}_{\text{noise}}$$

- Notoriously difficult:

  No Eckart–Young theorem, may be ill-posed for $R > 1$, NP-hard

- Nevertheless, several algorithms often "work well":

  Algebraic, block coordinate descent, Gauss-Newton, power iteration, ...

In practice: approximate CPD

Most ofen, we wish to recover a CPD from noisy observations:

$$\mathcal{Y} = \underbrace{\sum_{r=1}^{R} \lambda_r \, \boldsymbol{x}_r \otimes \boldsymbol{y}_r \otimes \boldsymbol{z}_r}_{\text{signal / information ("spike")}} + \underbrace{\mathcal{W}}_{\text{noise}}$$

- Notoriously difficult:

  No Eckart–Young theorem, may be ill-posed for $R > 1$, NP-hard

- Nevertheless, several algorithms often "work well":

  Algebraic, block coordinate descent, Gauss-Newton, power iteration, ...

- But how well are $\boldsymbol{x}_r, \boldsymbol{y}_r, \boldsymbol{z}_r$ and their estimates aligned?

$$\mathcal{Y} \approx \sum_{r=1}^{R} \hat{\lambda}_r \, \hat{\boldsymbol{x}}_r \otimes \hat{\boldsymbol{y}}_r \otimes \hat{\boldsymbol{z}}_r \quad \longrightarrow \quad |\langle \hat{\boldsymbol{x}}_r, \boldsymbol{x}_r \rangle|, |\langle \hat{\boldsymbol{y}}_r, \boldsymbol{y}_r \rangle|, |\langle \hat{\boldsymbol{z}}_r, \boldsymbol{z}_r \rangle|?$$

Q: Are there guarantees?

Q: Under which conditions is recovery possible?

Q: Can we do better?

The spiked (rank-one) tensor model

Particular case: Rank-1 symmetric spiked tensor,[1] order d

$$\mathcal{Y} = \lambda\, \boldsymbol{x}^{\otimes d} + \frac{1}{\sqrt{N}}\, \mathcal{W}$$

$$Y_{i_1 \ldots i_d} = \lambda\, x_{i_1} \ldots x_{i_d} + \frac{1}{\sqrt{N}}\, W_{i_1 \ldots i_d}$$

SNR $\lambda > 0$, $\boldsymbol{x} \in \mathbb{S}^{N-1}$, $\mathcal{W} = \frac{1}{d!} \sum_{\pi \in \mathfrak{S}_d} \pi \cdot \mathcal{G}$, $G_{i_1 \ldots i_d} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$

1. (Montanari & Richard, 2014), 2. (Anandkumar et al., 2014), 3. (Duchenne et al., 2011)

The spiked (rank-one) tensor model

Particular case: Rank-1 symmetric spiked tensor,[1] order d

$$\mathcal{Y} = \lambda\, \boldsymbol{x}^{\otimes d} + \frac{1}{\sqrt{N}} \mathcal{W}$$

$$Y_{i_1 \ldots i_d} = \lambda\, x_{i_1} \ldots x_{i_d} + \frac{1}{\sqrt{N}} W_{i_1 \ldots i_d}$$

SNR $\lambda > 0$, $\quad \boldsymbol{x} \in \mathbb{S}^{N-1}$, $\quad \mathcal{W} = \frac{1}{d!} \sum_{\pi \in \mathfrak{S}_d} \pi \cdot \mathcal{G}$, $\quad G_{i_1 \ldots i_d} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$

Problem of interest: estimate x from $\mathcal{Y}$, assumed large $(N \to \infty)$.

Closely related to rank-1 approximation

$$\operatorname*{arg\,min}_{\|\boldsymbol{u}\|=1} \left\| \mathcal{Y} - \lambda\, \boldsymbol{u}^{\otimes d} \right\|_{\mathsf{F}}^2$$

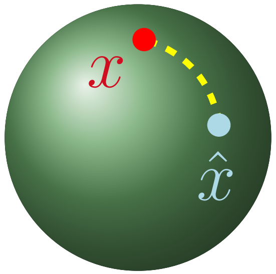
Applications: latent variable analysis,[2] computer vision[3]

1. (Montanari & Richard, 2014), 2. (Anandkumar et al., 2014), 3. (Duchenne et al., 2011)

Asymptotic performance limits?

Given any estimator $\hat{x} : \mathcal{S}^d(N) \to \mathbb{S}^{N-1}$, a natural performance measure is the alignment (or overlap):

$$\alpha_{d,N}(\lambda) := \langle x, \hat{x}(\boldsymbol{\mathcal{Y}}) \rangle \quad \in [-1, 1]$$

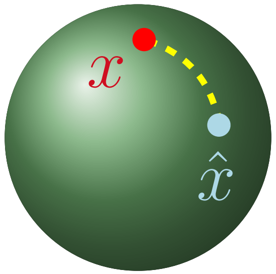


Example: If $\hat{x} \sim \mathcal{U}(\mathbb{S}^{N-1})$, then asymptotically $x \perp \hat{x}$ almost surely

Asymptotic performance limits?

Given any estimator $\hat{x} : \mathcal{S}^d(N) \to \mathbb{S}^{N-1}$, a natural performance measure is the alignment (or overlap):

$$\alpha_{d,N}(\lambda) := \langle x, \hat{x}(\boldsymbol{\mathcal{Y}}) \rangle \quad \in [-1, 1]$$



Example: If $\hat{x} \sim \mathcal{U}(\mathbb{S}^{N-1})$, then asymptotically $x \perp \hat{x}$ almost surely

Central questions:

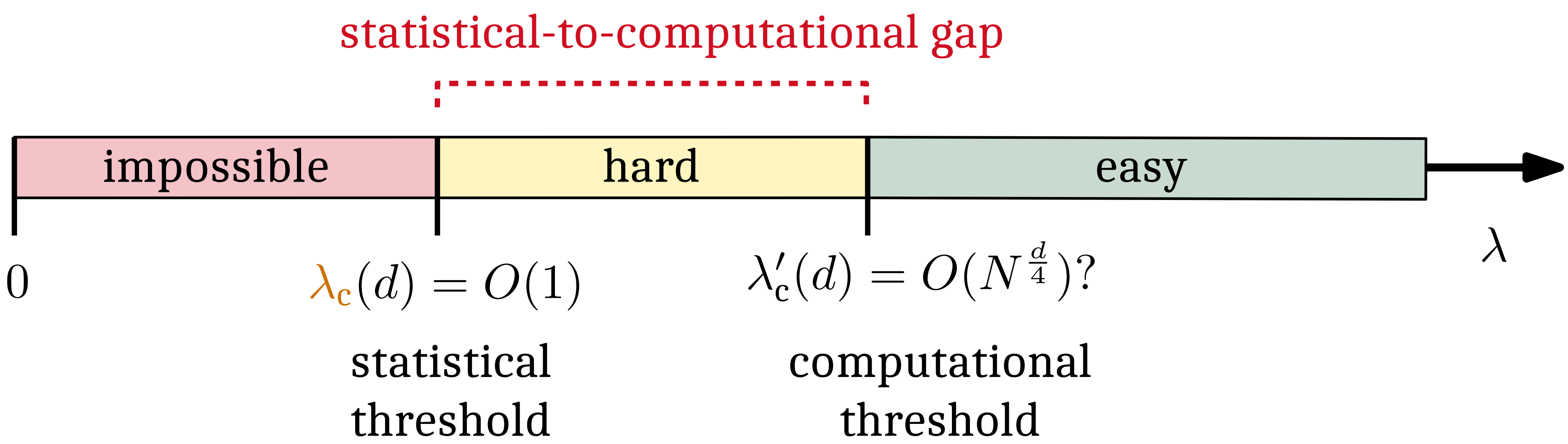
1. Weak recovery: for which range of λ is there a $\hat{x}$ such that

$$\limsup_{N\to\infty} \mathbb{E}\left\{\alpha_{d,N}(\lambda)\right\} > 0 \; ?$$

2. Best asymptotic alignment: what is the largest attainable value of $\limsup_{N\to\infty} \mathbb{E}\left\{\alpha_{d,N}(\lambda)\right\}$ for each λ ?

Answers and conjectured gap

1. Regimes of weak recovery:



2. Maximum likelihood estimation (MLE) attains the information-theoretic bound on the alignment for all λ.

(Richard & Montanari, 2014), (Montanari et al., 2015), (Ben Arous et al., 2019), (Jagannath et al., 2020), (Perry et al., 2020), (Ros et al., 2020)

This talk

1. Performance of maximum likelihood estimation
2. Leveraging RMT tools for MLE performance analysis
3. Rank-R case: performance of deflation
4. Epilogue and summary

Noise model and MLE

$\mathcal{W}$ $\in$ tensor Gaussian orthogonal ensemble

$Q \in \mathsf{O}(N)$

$$p(\mathcal{W}) = \frac{1}{Z_d(N)} \exp\left(-\frac{1}{2}\|\mathcal{W}\|_{\mathsf{F}}^2\right) = p((Q, \ldots, Q) \cdot \mathcal{W})$$

Noise model and MLE

$\mathcal{W} \in$ tensor Gaussian orthogonal ensemble

$Q \in \mathsf{O}(N)$

$$p(\mathcal{W}) = \frac{1}{Z_d(N)} \exp\left(-\frac{1}{2}\,\|\mathcal{W}\|_{\mathsf{F}}^2\right) = p((Q,\ldots,Q)\cdot\mathcal{W})$$

Consequently, $\quad p(\mathcal{Y}\,|\,x) \sim \exp\left(-\frac{N}{2}\left\|\mathcal{Y} - \lambda\, x^{\otimes d}\right\|_{\mathsf{F}}^2\right) \quad$ and:

Maximum likelihood estimator (MLE):

$$\hat{x} := \operatorname*{arg\,min}_{\|u\|=1} \left\|\mathcal{Y} - \lambda\, u^{\otimes d}\right\|_{\mathsf{F}}^2 = \operatorname*{arg\,max}_{\|u\|=1} \left\langle \mathcal{Y}, u^{\otimes d}\right\rangle$$

$$= \max_{\|u\|=1} \sum_{i_1\ldots i_d} Y_{i_1\ldots i_d} \prod_q u_{i_q} = \max_{\|u\|=1} \mathcal{Y}\cdot u^d$$

Noise model and MLE

$\mathcal{W} \in$ tensor Gaussian orthogonal ensemble

$Q \in \mathsf{O}(N)$

$$p(\mathcal{W}) = \frac{1}{Z_d(N)} \exp\left(-\frac{1}{2}\|\mathcal{W}\|_{\mathsf{F}}^2\right) = p((Q, \dots, Q) \cdot \mathcal{W})$$

Consequently, $p(\mathcal{Y} \,|\, x) \sim \exp\left(-\frac{N}{2}\left\|\mathcal{Y} - \lambda\, x^{\otimes d}\right\|_{\mathsf{F}}^2\right)$ and:

Maximum likelihood estimator (MLE):

$$\hat{x} := \underset{\|u\|=1}{\arg\min} \left\|\mathcal{Y} - \lambda\, u^{\otimes d}\right\|_{\mathsf{F}}^2 = \underset{\|u\|=1}{\arg\max} \left\langle \mathcal{Y}, u^{\otimes d}\right\rangle$$

$$= \max_{\|u\|=1} \sum_{i_1 \dots i_d} Y_{i_1 \dots i_d} \prod_q u_{i_q} = \max_{\|u\|=1} \mathcal{Y} \cdot u^d$$

Expected: $\lim_{N\to\infty} \mathbb{E}\left\{\alpha_{d,N}(\lambda)\right\} \approx \begin{cases} 1 & \text{for “large” } \lambda \\ 0 & \text{for “small” } \lambda \end{cases}$

But how exactly does $\alpha_{d,N}(\lambda)$ behave ?

MLE performance

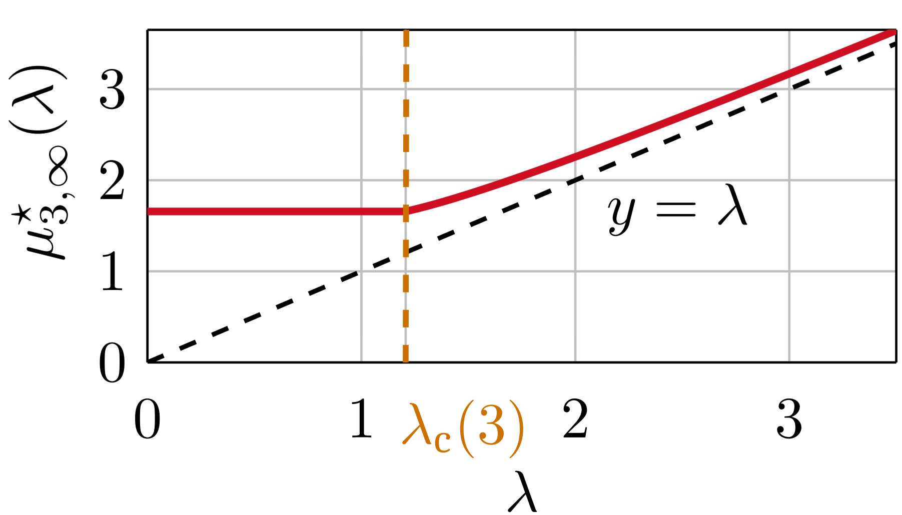
Settled by Jagannath–Lopatto–Miolane (2020), thanks to spin glass theory:

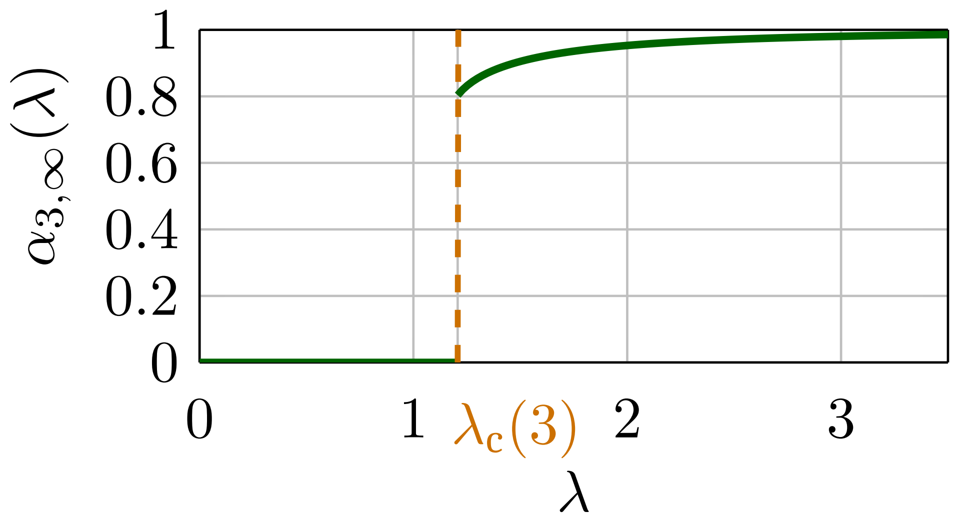
$$\mu^\star_{d,N}(\lambda) = \max_{\|u\|=1} \left\{ \lambda \langle x, u \rangle^d + \frac{1}{\sqrt{N}} \boldsymbol{\mathcal{W}} \cdot u^d \right\} \quad \xrightarrow[N\to\infty]{\text{a.s.}} \quad \mathsf{GS}_d + \int_0^\lambda q_d^*(t)^{d/2}\, dt$$

$$|\alpha_{d,N}(\lambda)| = |\langle x, \hat{x}_{\mathrm{ML}}(\lambda)\rangle| \quad \xrightarrow[N\to\infty]{\text{a.s.}} \quad \sqrt{q_d^*(\lambda)}$$

Explicit expressions exist for $d = 3, 4, 5.$

For all d, these quantites undergo a phase transition at a threshold $\lambda_{\mathrm{c}}(d) = O(1)$.





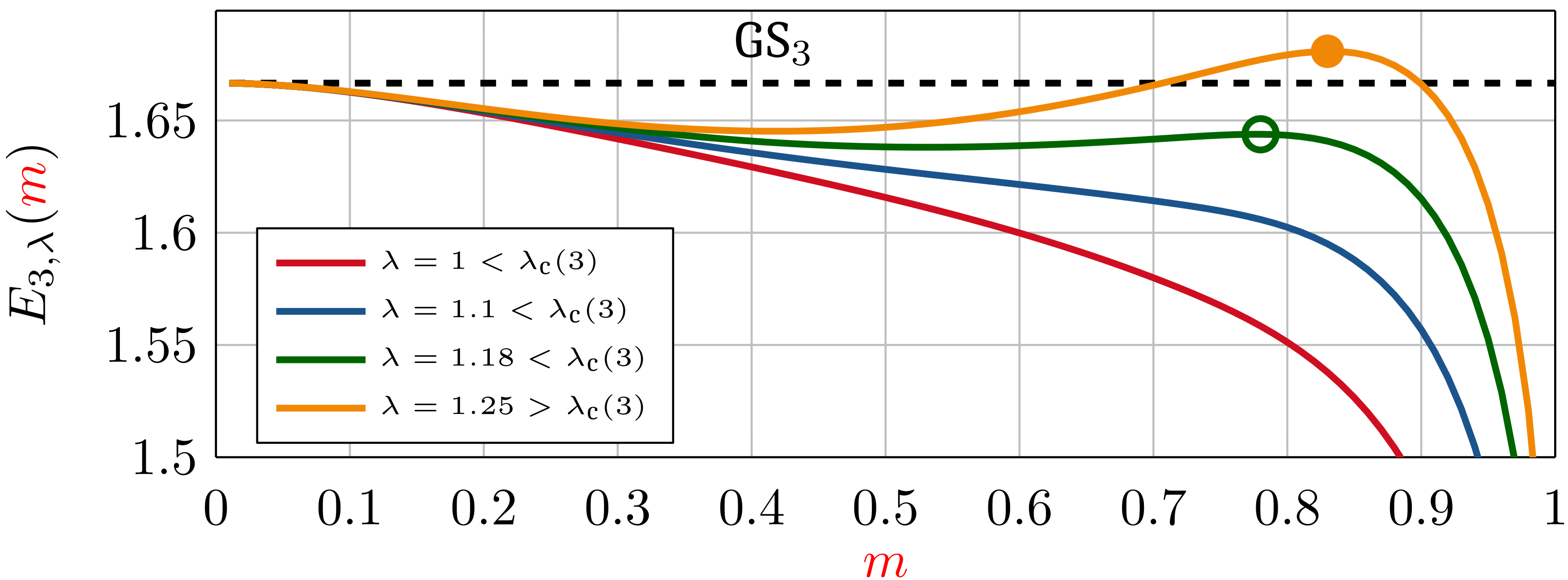
Furthermore, the MLE attains the bound $\limsup_N \mathbb{E}\{|\langle x, \hat{x}\rangle|\} \leq \sqrt{q_d^*(\lambda)}$

Why the discontinuity?

Max of $\boldsymbol{\mathcal{Y}} \cdot u^d$ at a fixed allignment $\langle x, u\rangle = m$ ("lattitude"):

$$E_{d,\lambda}(m) = \lim_{N\to\infty} \mathbb{E}\left\{ \max_{\substack{\|u\|=1 \\ \langle u,x\rangle = m}} \lambda \langle u, x\rangle^d + \frac{1}{\sqrt{N}} \boldsymbol{\mathcal{W}} \cdot u^d \right\}$$

Numerical evaluation for $d = 3$:



This talk

1. Performance of maximum likelihood estimation

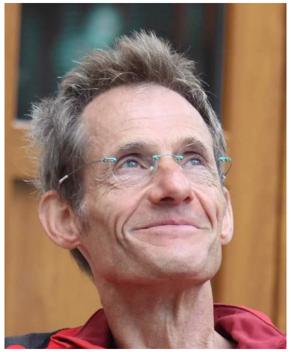
2. Leveraging RMT tools for MLE performance analysis

3. Rank-R case: performance of deflation

4. Epilogue and summary

Next : joint work with R. Couillet and P. Comon (JMLR '22)

LIG, Grenoble



Gipsa, Grenoble

Tensor eigenpairs and MLE

ML problem

$$\max_{\|u\|=1} \mathcal{Y} \cdot u^d$$

Lagrangian

$$L(\mu, u) = \frac{1}{d} \mathcal{Y} \cdot u^d - \frac{\mu}{2} \left(\|u\|^2 - 1\right)$$

Tensor eigenpairs and MLE

ML problem

$$\max_{\|u\|=1} \boldsymbol{\mathcal{Y}} \cdot u^d$$

Lagrangian

$$L(\mu, u) = \frac{1}{d}\boldsymbol{\mathcal{Y}} \cdot u^d - \frac{\mu}{2}\left(\|u\|^2 - 1\right)$$

Critical points satisfy

$$\frac{\partial}{\partial u} L(\mu, u) = \boldsymbol{\mathcal{Y}} \cdot u^{d-1} - \mu\, u = 0, \qquad \|u\| = 1$$

Tensor eigenpairs and MLE

ML problem

$$\max_{\|u\|=1} \boldsymbol{\mathcal{Y}} \cdot u^d$$

Lagrangian

$$L(\mu, u) = \frac{1}{d} \boldsymbol{\mathcal{Y}} \cdot u^d - \frac{\mu}{2} \left(\|u\|^2 - 1 \right)$$

Critical points satisfy

$$\frac{\partial}{\partial u} L(\mu, u) = \boldsymbol{\mathcal{Y}} \cdot u^{d-1} - \mu\, u = 0, \qquad \|u\| = 1$$

Tensor ℓ_2-eigenvalue equations : (Lim, 2005)

$$\boldsymbol{\mathcal{Y}} \cdot u^{d-1} = \mu\, u, \qquad \|u\| = 1$$

In particular :

$$\boldsymbol{\mathcal{Y}} \cdot \hat{x}_{\mathrm{ML}}^{d-1} = \mu_{\max}\, \hat{x}_{\mathrm{ML}}$$

Tensor and matrix eigenpairs

Another characterization of tensor eigenpairs (assuming $\|u\| = 1$):

$$(\mu, u) \text{ eigenpair of } \boldsymbol{\mathcal{Y}} \quad \Leftrightarrow \quad (\mu, u) \text{ eigenpair of } \boldsymbol{\mathcal{Y}} \cdot u^{d-2}$$

Tensor and matrix eigenpairs

Another characterization of tensor eigenpairs (assuming $\|u\| = 1$):

$$(\mu, u) \text{ eigenpair of } \boldsymbol{\mathcal{Y}} \quad \Leftrightarrow \quad (\mu, u) \text{ eigenpair of } \boldsymbol{\mathcal{Y}} \cdot u^{d-2}$$

Proof: $\mu\, u = \boldsymbol{\mathcal{Y}} \cdot u^{d-1} = (\boldsymbol{\mathcal{Y}} \cdot u^{d-2})\, u$

Tensor and matrix eigenpairs

Another characterization of tensor eigenpairs (assuming $\|u\| = 1$):

$$(\mu, u) \text{ eigenpair of } \mathfrak{Y} \quad \Leftrightarrow \quad (\mu, u) \text{ eigenpair of } \mathfrak{Y} \cdot u^{d-2}$$

Proof: $\mu\, u = \mathfrak{Y} \cdot u^{d-1} = (\mathfrak{Y} \cdot u^{d-2})\, u$

In particular, if (μ, u) is a local max, then $\mathrm{Sp}(\mathfrak{Y} \cdot u^{d-2}) \setminus \{\mu\} \subset \;]-\infty, \frac{\mu}{d-1}]$

Proof: Apply the second-order necessary condition

$$\langle \nabla^2_{uu} L(\mu, u)\, w, w \rangle \leq 0, \qquad \forall\, w \in u^{\perp}$$

with $\nabla^2_{uu} L(\mu, u) = \frac{\partial}{\partial u} \left[\mathfrak{Y} \cdot u^{d-1} - \mu u \right] = (d-1)\, \mathfrak{Y} \cdot u^{d-2} - \mu I$ to get

$$\max_{\|w\|=1,\, w \in u^{\perp}} \langle (\mathfrak{Y} \cdot u^{d-2})\, w, w \rangle \leq \frac{\mu}{d-1}$$

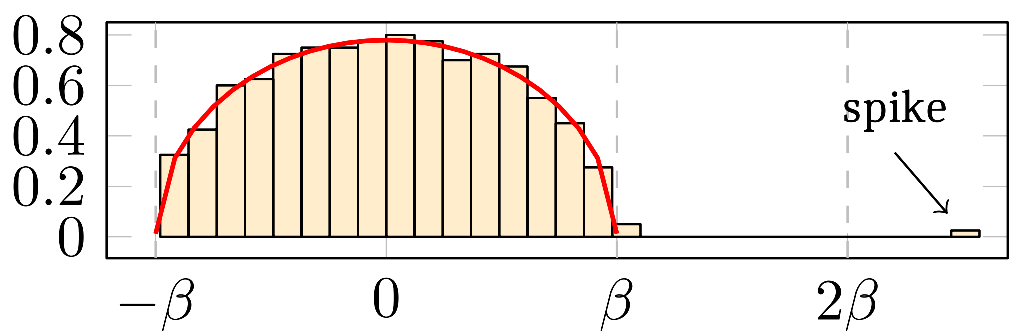
From spiked tensor model to matrix models

Idea : study spiked rank-one matrix models at critical points (μ, u)

$$\boldsymbol{\mathcal{Y}} \cdot u^{d-2} = \underbrace{\lambda \, \langle x, u \rangle^{d-2} \, x \, x^{\mathsf{T}}}_{\text{rank-1 "signal"}} + \underbrace{\frac{1}{\sqrt{N}} \boldsymbol{\mathcal{W}} \cdot u^{d-2}}_{\text{noise}} \qquad \in \mathbb{R}^{N \times N}$$

- SNR weighted by alignment $\langle x, u \rangle$
- $\boldsymbol{\mathcal{W}}$ and u are correlated $\Rightarrow$ "spike" at every local max u regardless of λ

$\mathrm{Sp}(\boldsymbol{\mathcal{Y}} \cdot u^{d-2}) \quad (d = 3)$



Contractions $\boldsymbol{\mathcal{Y}} \cdot u^{d-2}$ at critical points

Recall the tensor eigenvalue equation:

$$\mu\, u = \boldsymbol{\mathcal{Y}} \cdot u^{d-1} = \lambda \, \langle x, u \rangle^{d-1} \, x + \frac{1}{\sqrt{N}} \boldsymbol{\mathcal{W}} \cdot u^{d-1}$$

Our quantities of interest are obtained by taking scalar products with u and x :

$$\begin{aligned} \mu &= \lambda \, \langle x, u \rangle^{d} + \frac{1}{\sqrt{N}} \boldsymbol{\mathcal{W}} \cdot u^{d}, \\ \langle x, u \rangle &= \frac{\lambda}{\mu} \, \langle x, u \rangle^{d-1} + \frac{1}{\mu \sqrt{N}} \left\langle x, \boldsymbol{\mathcal{W}} \cdot u^{d-1} \right\rangle \qquad (\mu \neq 0) \end{aligned}$$

Contractions $\boldsymbol{\mathcal{Y}} \cdot u^{d-2}$ at critical points

Recall the tensor eigenvalue equation:

$$\mu\, u = \boldsymbol{\mathcal{Y}} \cdot u^{d-1} = \lambda\, \langle x, u \rangle^{d-1}\, x + \frac{1}{\sqrt{N}} \boldsymbol{\mathcal{W}} \cdot u^{d-1}$$

Our quantities of interest are obtained by taking scalar products with u and x :

$$\begin{aligned} \mu &= \lambda\, \langle x, u \rangle^{d} + \frac{1}{\sqrt{N}} \boldsymbol{\mathcal{W}} \cdot u^{d}, \\ \langle x, u \rangle &= \frac{\lambda}{\mu}\, \langle x, u \rangle^{d-1} + \frac{1}{\mu\sqrt{N}} \left\langle x, \boldsymbol{\mathcal{W}} \cdot u^{d-1} \right\rangle \qquad (\mu \neq 0) \end{aligned}$$

Q : Assuming $\mu \xrightarrow[N\to\infty]{\text{a.s.}} \mu_{d,\infty}$ and $\langle x, u \rangle \xrightarrow[N\to\infty]{\text{a.s.}} \alpha_{d,\infty}$, which solutions do we get for $(\mu_{d,\infty}, \alpha_{d,\infty})$?

Contractions $\boldsymbol{\mathcal{Y}} \cdot u^{d-2}$ at critical points

Recall the tensor eigenvalue equation:

$$\mu\, u = \boldsymbol{\mathcal{Y}} \cdot u^{d-1} = \lambda\, \langle x, u \rangle^{d-1}\, x + \frac{1}{\sqrt{N}} \boldsymbol{\mathcal{W}} \cdot u^{d-1}$$

Our quantities of interest are obtained by taking scalar products with u and x :

$$\mathbb{E}\left\{\mu\right\} = \lambda\, \mathbb{E}\left\{\langle x, u \rangle^{d}\right\} + \frac{1}{\sqrt{N}}\, \mathbb{E}\left\{\boldsymbol{\mathcal{W}} \cdot u^{d}\right\}$$

$$\mathbb{E}\left\{\langle x, u \rangle\right\} = \lambda\, \mathbb{E}\left\{\frac{1}{\mu}\, \langle x, u \rangle^{d-1}\right\} + \frac{1}{\sqrt{N}}\, \mathbb{E}\left\{\frac{1}{\mu}\left\langle x, \boldsymbol{\mathcal{W}} \cdot u^{d-1}\right\rangle\right\} \qquad (\mu \neq 0)$$

Q : Assuming $\mu \xrightarrow[N\to\infty]{\text{a.s.}} \mu_{d,\infty}$ and $\langle x, u \rangle \xrightarrow[N\to\infty]{\text{a.s.}} \alpha_{d,\infty}$, which solutions do we get for $(\mu_{d,\infty}, \alpha_{d,\infty})$?

Contractions $\boldsymbol{\mathcal{Y}} \cdot u^{d-2}$ at critical points

Recall the tensor eigenvalue equation:

$$\mu\, u = \boldsymbol{\mathcal{Y}} \cdot u^{d-1} = \lambda \, \langle x, u \rangle^{d-1} \, x + \frac{1}{\sqrt{N}} \boldsymbol{\mathcal{W}} \cdot u^{d-1}$$

Our quantities of interest are obtained by taking scalar products with u and x :

$$\mathbb{E}\left\{\mu\right\} = \lambda\, \mathbb{E}\left\{\langle x, u\rangle^{d}\right\} + \frac{1}{\sqrt{N}} \mathbb{E}\left\{\boldsymbol{\mathcal{W}} \cdot u^{d}\right\}$$

$$\mathbb{E}\left\{\langle x, u\rangle\right\} = \lambda\, \mathbb{E}\left\{\frac{1}{\mu} \, \langle x, u\rangle^{d-1}\right\} + \frac{1}{\sqrt{N}} \mathbb{E}\left\{\frac{1}{\mu} \left\langle x, \boldsymbol{\mathcal{W}} \cdot u^{d-1}\right\rangle\right\} \qquad (\mu \neq 0)$$

Q : Assuming $\mu \xrightarrow[N\to\infty]{\text{a.s.}} \mu_{d,\infty}$ and $\langle x, u\rangle \xrightarrow[N\to\infty]{\text{a.s.}} \alpha_{d,\infty}$, which solutions do we get for $\left(\mu_{d,\infty}, \alpha_{d,\infty}\right)$?

Technical tools

Computation of $\mathbb{E}\left\{\mathcal{W} \cdot u^d\right\}$ and $\mathbb{E}\left\{\frac{1}{\mu}\left\langle x, \mathcal{W} \cdot u^{d-1}\right\rangle\right\}$:

1. Gaussian integration-by-parts

$$z \sim \mathcal{N}(0, \sigma^2) \quad \Rightarrow \quad \mathbb{E}\left\{z\, f(z)\right\} = \sigma^2\, \mathbb{E}\left\{f'(z)\right\}$$

$$\mathbb{E}\left\{\mathcal{W} \cdot u^d\right\} = \sum_{\boldsymbol{i}} \mathbb{E}\left\{W_{\boldsymbol{i}}\, u_{i_1} \ldots u_{i_d}\right\} = \sum_{\boldsymbol{i}} \sigma_{\boldsymbol{i}}^2 \sum_{j=1}^{d} \mathbb{E}\left\{\frac{\partial u_{i_j}}{\partial W_{\boldsymbol{i}}} \prod_{k \neq j}^{d} u_{i_k}\right\}$$

$$\mathbb{E}\left\{\frac{1}{\mu}\left\langle x, \mathcal{W} \cdot u^{d-1}\right\rangle\right\} = \sum_{\boldsymbol{i}} \sigma_{\boldsymbol{i}}^2\, \mathbb{E}\left\{-\frac{1}{\mu^2} \frac{\partial \mu}{\partial W_{\boldsymbol{i}}} \left(\prod_{k=1}^{d-1} u_{i_k}\right) x_{i_d} + \frac{1}{\mu} x_{i_d} \sum_{j=1}^{d-1} \frac{\partial u_{i_j}}{\partial W_{\boldsymbol{i}}} \prod_{k \neq j}^{d-1} u_{i_k}\right\}.$$

2. Implicit function theorem to compute the required derivatives :

$$\begin{pmatrix} \frac{\partial u}{\partial W_{\boldsymbol{i}}} \\ \frac{\partial \mu}{\partial W_{\boldsymbol{i}}} \end{pmatrix} = \begin{pmatrix} -\frac{1}{(d-1)\sqrt{N}} R\left(\frac{\mu}{d-1}\right) \phi + \frac{1}{(d-2)\,\mu} \frac{\partial \mu}{\partial W_{\boldsymbol{i}}} u \\ \frac{1}{\sigma_{\boldsymbol{i}}^2 \sqrt{N}} \prod_{j=1}^{d} u_{i_j} \end{pmatrix}$$

Main result

Theorem. Let (μ, u) be a sequence of critical points of the MLE problem s.t.

$$\langle x, u\rangle \xrightarrow[N\to\infty]{\text{a.s.}} \alpha_{d,\infty}(\lambda) > 0,$$
$$\mu \xrightarrow[N\to\infty]{\text{a.s.}} \mu_{d,\infty}(\lambda) > (d-1)\beta_d.$$

Then, $\mu_{d,\infty}(\lambda)$ satisfies the fixed-point equation

$$\mu_{d,\infty}(\lambda) = \phi_d(\mu_{d,\infty}(\lambda), \lambda),$$

where

$$\phi_d(z,\lambda) := \lambda\, \omega_d^d(z,\lambda) - \frac{1}{d-1}\, m_d\left(\frac{z}{d-1}\right),$$

$$\omega_d(z,\lambda) := \left[\frac{1}{\lambda}\left(z + \frac{1}{d}\, m_d\left(\frac{z}{d-1}\right)\right)\right]^{\frac{1}{d-2}},$$

$$m_d(z) = \frac{2}{\beta_d^2}\left(-z + z\sqrt{1 - \frac{\beta_d^2}{z^2}}\right)$$

(Stieltjes transform of limiting spectral distribution)

Furthermore, $\alpha_{d,\infty}(\lambda) = \omega_d(\mu_{d,\infty}(\lambda), \lambda)$.

Relation to known results & beyond

For $d = 3$: the only positive solution to the fixed-point equation is:

$$\mu_{3,\infty}(\lambda) = \frac{3\lambda^2 + \lambda\sqrt{9\lambda^2 - 12} + 4}{\sqrt{18\lambda^2 + 6\lambda\sqrt{9\lambda^2 - 12}}}, \qquad \lambda \geq 2/\sqrt{3},$$

which gives

$$\alpha_{3,\infty}(\lambda) = \sqrt{\frac{1}{2} + \sqrt{\frac{3\lambda^2 - 4}{12\lambda^2}}}.$$

These equations precisely describe the local max of $E_{d,\lambda}(m)$ that becomes global (MLE) for $\lambda > \lambda_c(d)$.

Relation to known results & beyond

For $d = 3$: the only positive solution to the fixed-point equation is:

$$\mu_{3,\infty}(\lambda) = \frac{3\lambda^2 + \lambda\sqrt{9\lambda^2 - 12} + 4}{\sqrt{18\lambda^2 + 6\lambda\sqrt{9\lambda^2 - 12}}}, \qquad \lambda \geq 2/\sqrt{3},$$

which gives

$$\alpha_{3,\infty}(\lambda) = \sqrt{\frac{1}{2} + \sqrt{\frac{3\lambda^2 - 4}{12\lambda^2}}}.$$

These equations precisely describe the local max of $E_{d,\lambda}(m)$ that becomes global (MLE) for $\lambda > \lambda_{\mathrm{c}}(d)$.

For $d = 4, 5$: same is true according to numerical verification.

Conjecture: true for all $d \geq 3$.

Relation to known results & beyond

For $d = 3$: the only positive solution to the fixed-point equation is:

$$\mu_{3,\infty}(\lambda) = \frac{3\lambda^2 + \lambda\sqrt{9\lambda^2 - 12} + 4}{\sqrt{18\lambda^2 + 6\lambda\sqrt{9\lambda^2 - 12}}}, \qquad \lambda \geq 2/\sqrt{3},$$

which gives

$$\alpha_{3,\infty}(\lambda) = \sqrt{\frac{1}{2} + \sqrt{\frac{3\lambda^2 - 4}{12\lambda^2}}}.$$

These equations precisely describe the local max of $E_{d,\lambda}(m)$ that becomes global (MLE) for $\lambda > \lambda_{\mathrm{c}}(d)$.

For $d = 4, 5$: same is true according to numerical verification.

Conjecture: true for all $d \geq 3$.

Even more important: how to extend this approach to more general models ?

Byproduct: limiting spectrum of contractions

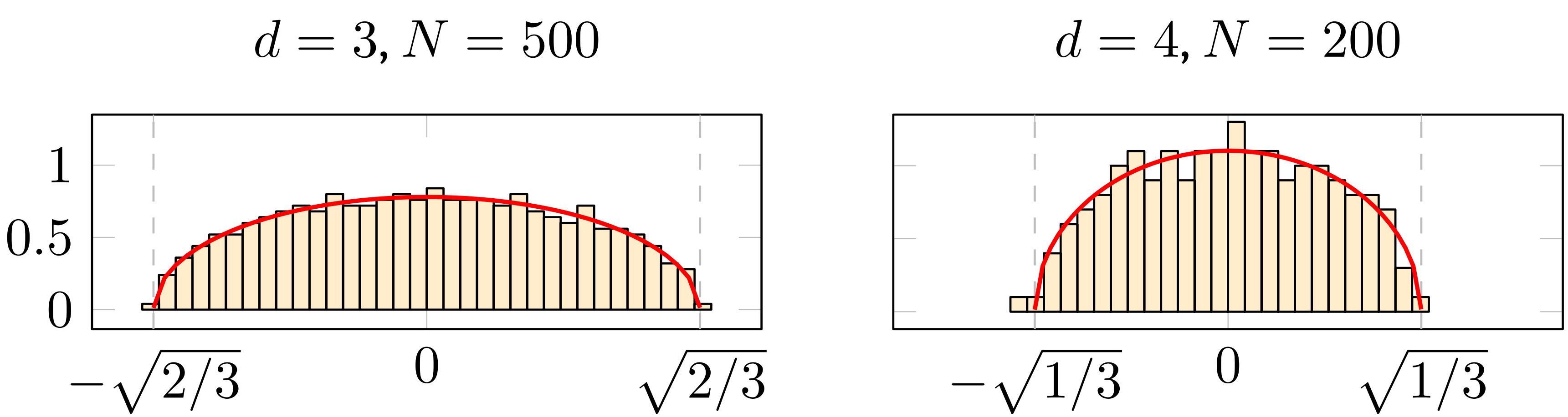
Theorem. Take a sequence of vectors $v \in \mathbb{S}^{N-1}$ and of Gaussian tensors $\mathcal{W}$.

Define the empirical spectral measure $\rho_N = \frac{1}{N}\sum_{i=1}^N \delta_{\nu_i}$, where

$$\{\nu_i\}_{i=1}^N = \mathrm{Sp}\left(\frac{1}{\sqrt{N}}\mathcal{W}\cdot v^{d-2}\right)$$

Then, ρ_N converges a.s. weakly to a semicircle law ρ with density

$$\rho(dx) = \frac{2}{\pi\,\beta_d^2}\sqrt{\left(\beta_d^2 - x^2\right)^+}\,dx, \qquad \beta_d := \frac{2}{\sqrt{d(d-1)}}.$$



But why does this work?

Ingredients: tensor eigenvalue equation satsfied by all critical points and a.s. convergence

$$\langle x, u \rangle \xrightarrow[N\to\infty]{\text{a.s.}} \alpha_{d,\infty}(\lambda) > 0$$
$$\mu \xrightarrow[N\to\infty]{\text{a.s.}} \mu_{d,\infty}(\lambda) > (d-1)\beta_d$$

Output: description of global maximizer(s) for $\lambda > \lambda_{\mathrm{c}}(d)$.

But why does this work?

Ingredients: tensor eigenvalue equation satsfied by all critical points and a.s. convergence

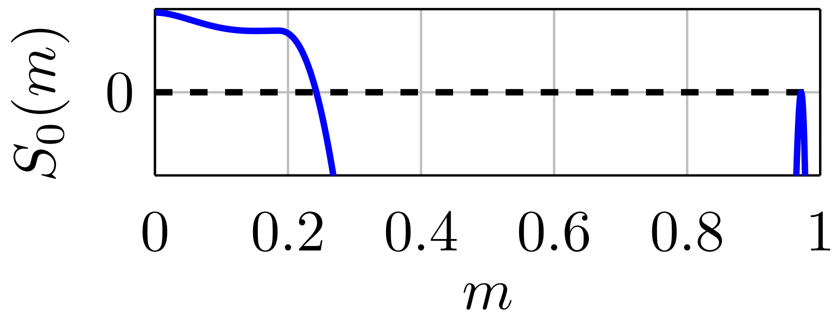
$$\langle x, u\rangle \xrightarrow[N\to\infty]{\text{a.s.}} \alpha_{d,\infty}(\lambda) > 0$$

$$\mu \xrightarrow[N\to\infty]{\text{a.s.}} \mu_{d,\infty}(\lambda) > (d-1)\beta_d$$

Output: description of global maximizer(s) for $\lambda > \lambda_{\mathrm{c}}(d)$.

The conditions on $\alpha_{d,\infty}$ and $\mu_{d,\infty}$ could be met by some strictly local max when $\lambda > \lambda_{\mathrm{c}}(d)$, as per Ben Arous–Mei–Montanari (2019)

expected # local max with alignment m $\sim \exp(N\, S_0(m))$



Our take: The "selectivity" probably comes from the a.s. conv. assumption.

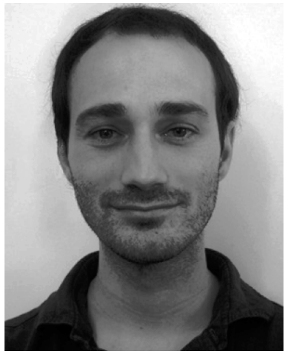
This talk

1. Performance of maximum likelihood estimation
2. Leveraging RMT tools for MLE performance analysis
3. Rank-R case: performance of deflation
4. Epilogue and summary

Next : joint work with M. Seddik, M. Guillaud & A. Decurninge (GRETSI'23)

TII, Abu Dhabi

INRIA Lyon



Huawei Paris

"But you promised to talk about the rank-R case!"

If we wanted an exact (symmetric) orthogonal decomposition (odeco)

$$\mathcal{S} = \sum_{r=1}^{R} \lambda_r \, \boldsymbol{x}_r^{\otimes d}, \qquad \langle \boldsymbol{x}_r, \boldsymbol{x}_s \rangle = \delta_{rs},$$

Hotelling-type deflation[1]

1. Let $\mathcal{S}_1 = \mathcal{S}$
2. For $i = 1, \ldots, R$ do
   - (i) Solve $\boldsymbol{u}_i \equiv \operatorname*{arg\,max}_{\|\boldsymbol{u}\|=1} \mathcal{S}_i \cdot \boldsymbol{u}^d = \operatorname*{arg\,max}_{\|\boldsymbol{u}\|=1} \langle \mathcal{S}_i, \boldsymbol{u}^{\otimes d} \rangle$
   - (ii) Compute $\hat{\mu}_i = \mathcal{S}_i \cdot \boldsymbol{u}_i^d$ rank-1 approx.
   - (iii) "Deflate": $\mathcal{S}_{i+1} = \mathcal{S}_i - \hat{\mu}_i \, \boldsymbol{u}_i^{\otimes d}$
3. return $\{(\hat{\mu}_i, \boldsymbol{u}_i)\}_{i=1}^{r}$

1. (Hotelling, 1933), 2. (Stegeman & Comon, 2010)

"But you promised to talk about the rank-R case!"

If we wanted an exact (symmetric) orthogonal decomposition (odeco)

$$\mathcal{S} = \sum_{r=1}^{R} \lambda_r \, \boldsymbol{x}_r^{\otimes d}, \qquad \langle \boldsymbol{x}_r, \boldsymbol{x}_s \rangle = \delta_{rs},$$

Hotelling-type deflation[1]

1. Let $\mathcal{S}_1 = \mathcal{S}$
2. For $i = 1, \ldots, R$ do
   - (i) Solve $\boldsymbol{u}_i \equiv \underset{\|\boldsymbol{u}\|=1}{\arg\max}\; \mathcal{S}_i \cdot \boldsymbol{u}^d = \underset{\|\boldsymbol{u}\|=1}{\arg\max}\; \langle \mathcal{S}_i, \boldsymbol{u}^{\otimes d} \rangle$
   - (ii) Compute $\hat{\mu}_i = \mathcal{S}_i \cdot \boldsymbol{u}_i^d$ rank-1 approx.
   - (iii) "Deflate": $\mathcal{S}_{i+1} = \mathcal{S}_i - \hat{\mu}_i \, \boldsymbol{u}_i^{\otimes d}$
3. return $\{(\hat{\mu}_i, \boldsymbol{u}_i)\}_{i=1}^{r}$

Problem: Doesn't work for non-orthogonal (and/or noisy) decomposition

Even worse: Subtracting the best rank-1 approx. of $\mathcal{S}_i$ may increase rank[2]

1. (Hotelling, 1933), 2. (Stegeman & Comon, 2010)

Is it that bad? Can we improve it?

Goal: Characterize asymptotic performance of deflation as $N \to \infty$ assuming

$$\mathcal{S} = \sum_{r=1}^{R} \lambda_r \, \boldsymbol{x}_r^{\otimes d} + \frac{1}{\sqrt{N}} \mathcal{W},$$

rank-R spiked model

$$\left[\begin{array}{l} \langle \boldsymbol{x}_r, \boldsymbol{x}_s \rangle = \begin{cases} 1, & r = s, \\ \alpha_{rs} \neq 0, & r \neq s, \end{cases} \\ \lambda_r > 0 \\ \mathcal{W} = \dfrac{1}{d!} \displaystyle\sum_{\pi \in \mathfrak{S}_d} \pi \cdot \mathcal{G}, \quad G_{i_1 \ldots i_d} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \end{array}\right.$$

Methodology: Analysis of contractions $\mathcal{S}_i \cdot \boldsymbol{u}^{d-2}$ using tools from RMT[1,2]

1. (Goulart et al., 2022), 2. (Seddik et al., 2023)

Is it that bad? Can we improve it?

Goal: Characterize asymptotic performance of deflation as $N \to \infty$ assuming

$$\mathcal{S} = \sum_{r=1}^{R} \lambda_r \, \boldsymbol{x}_r^{\otimes d} + \frac{1}{\sqrt{N}} \mathcal{W},$$

rank-R spiked model

$$\left[\begin{array}{l} \langle \boldsymbol{x}_r, \boldsymbol{x}_s \rangle = \begin{cases} 1, & r = s, \\ \alpha_{rs} \neq 0, & r \neq s, \end{cases} \\ \lambda_r > 0 \\ \mathcal{W} = \frac{1}{d!} \sum_{\pi \in \mathfrak{S}_d} \pi \cdot \mathcal{G}, \quad G_{i_1 \dots i_d} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \end{array}\right.$$

Methodology: Analysis of contractions $\mathcal{S}_i \cdot \boldsymbol{u}^{d-2}$ using tools from RMT[1,2]

TL;DR: Asymptotic analysis $\Rightarrow$ system of eqs. in model params λ_r, α_{rs} and in

summary statistics: $\hat{\mu}_i, \quad \hat{\eta}_{ij} \equiv \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle, \quad \hat{\rho}_{ir} \equiv \langle \boldsymbol{u}_i, \boldsymbol{x}_r \rangle$

$\Rightarrow$ numerical computation of sum. stat. given model params (theoretical tool)

$\Rightarrow$ numerical computation of params and $\hat{\rho}_{ir}$ given $\hat{\mu}_i, \hat{\eta}_{ij}$ (practical tool?)

1. (Goulart et al., 2022), 2. (Seddik et al., 2023)

Limiting summary statistics

Technical hypotheses:

$$\hat{\mu}_i \xrightarrow[N\to\infty]{\text{a.s.}} \mu_i > \gamma_d(d-1), \qquad \underbrace{\hat{\rho}_{ir}}^{\langle \boldsymbol{u}_i, \boldsymbol{x}_r \rangle} \xrightarrow[N\to\infty]{\text{a.s.}} \rho_{ir} \neq 0, \qquad \underbrace{\hat{\eta}_{ij}}^{\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle} \xrightarrow[N\to\infty]{\text{a.s.}} \eta_{ij} \neq 0$$

Limiting summary statistics

Technical hypotheses:

$$\hat{\mu}_i \xrightarrow[N\to\infty]{\text{a.s.}} \mu_i > \gamma_d(d-1), \qquad \overbrace{\hat{\rho}_{ir}}^{\langle \boldsymbol{u}_i, \boldsymbol{x}_r \rangle} \xrightarrow[N\to\infty]{\text{a.s.}} \rho_{ir} \neq 0, \qquad \overbrace{\hat{\eta}_{ij}}^{\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle} \xrightarrow[N\to\infty]{\text{a.s.}} \eta_{ij} \neq 0$$

Starting point: From $\boldsymbol{\mathcal{S}}_i = \overbrace{\sum_{r=1}^{R} \lambda_r \, \boldsymbol{x}_r^{\otimes d} + \frac{1}{\sqrt{N}} \boldsymbol{\mathcal{W}}}^{\boldsymbol{\mathcal{S}} \text{ (model)}} - \overbrace{\sum_{j=1}^{i-1} \hat{\mu}_j \, \boldsymbol{u}_j^{\otimes d}}^{\text{deflation}}$ follows

$$\underbrace{\hat{\mu}_i \, \boldsymbol{u}_i = \boldsymbol{\mathcal{S}}_i \cdot \boldsymbol{u}_i^{d-1}}_{\text{eigenvalue eqn}} = \sum_{r=1}^{R} \lambda_r \langle \boldsymbol{u}_i, \boldsymbol{x}_r \rangle^{d-1} \boldsymbol{x}_r + \frac{1}{\sqrt{N}} \boldsymbol{\mathcal{W}} \cdot \boldsymbol{u}_i^{d-1} - \sum_{j=1}^{i-1} \hat{\mu}_j \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle^{d-1} \boldsymbol{u}_j$$

Limiting summary statistics

Technical hypotheses:

$$\hat{\mu}_i \xrightarrow[N\to\infty]{\text{a.s.}} \mu_i > \gamma_d(d-1), \qquad \overset{\langle \boldsymbol{u}_i, \boldsymbol{x}_r \rangle}{\hat{\rho}_{ir}} \xrightarrow[N\to\infty]{\text{a.s.}} \rho_{ir} \neq 0, \qquad \overset{\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle}{\hat{\eta}_{ij}} \xrightarrow[N\to\infty]{\text{a.s.}} \eta_{ij} \neq 0$$

Starting point: From $\mathcal{S}_i = \overbrace{\sum_{r=1}^{R} \lambda_r \, \boldsymbol{x}_r^{\otimes d} + \frac{1}{\sqrt{N}} \mathcal{W}}^{\mathcal{S}\ \text{(model)}} - \overbrace{\sum_{j=1}^{i-1} \hat{\mu}_j \, \boldsymbol{u}_j^{\otimes d}}^{\text{deflation}}$ follows

$$\underbrace{\hat{\mu}_i \, \boldsymbol{u}_i = \mathcal{S}_i \cdot \boldsymbol{u}_i^{d-1}}_{\text{eigenvalue eqn}} = \sum_{r=1}^{R} \lambda_r \langle \boldsymbol{u}_i, \boldsymbol{x}_r \rangle^{d-1} \boldsymbol{x}_r + \frac{1}{\sqrt{N}} \mathcal{W} \cdot \boldsymbol{u}_i^{d-1} - \sum_{j=1}^{i-1} \hat{\mu}_j \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle^{d-1} \boldsymbol{u}_j$$

Bottom line: By contracting with $\boldsymbol{u}_j$ or $\boldsymbol{x}_r$ and using Stein's lemma, we get an asymptotic system of equations in the summary statistics.

Main result

Theorem. As $N \to \infty$, under our technical hypotheses,

$$
\begin{array}{ll}
\times\, \boldsymbol{u}_i \;\to & \\
\times\, \boldsymbol{x}_r \;\to & \\
\times\, \boldsymbol{u}_j \;\to &
\end{array}
\begin{cases}
\mu_i + \dfrac{1}{d-1}\, g\left(\dfrac{\mu_i}{d-1}\right) = \displaystyle\sum_{r=1}^{R} \lambda_r \rho_{ir}^d - \sum_{j=1}^{i-1} \mu_j \eta_{ij}^d, \\[2ex]
h(\mu_i)\, \rho_{ir} = \displaystyle\sum_{s=1}^{R} \lambda_s\, \alpha_{rs}\, \rho_{is}^{d-1} - \sum_{j=1}^{i-1} \mu_j\, \rho_{jr}\, \eta_{ij}^{d-1}, \\[2ex]
\left[h(\mu_i) + q(\mu_j)\, \eta_{ij}^{d-2}\right] \eta_{ij} = \displaystyle\sum_{r=1}^{R} \lambda_r\, \rho_{jr}\, \rho_{ir}^{d-1} - \sum_{k=1}^{i-1} \mu_k\, \eta_{kj}\, \eta_{ik}^{d-1},
\end{cases}
$$

for $(i, j) \in \{1, \ldots, R\}^2$, with

$$
g(z) \equiv \frac{2}{\gamma_d^2}\left(-z + \sqrt{z^2 - \gamma_d^2}\right), \qquad \text{(limiting Stieltjes transform)}
$$

$$
h(z) \equiv z + \frac{1}{d}\, g\left(\frac{z}{d-1}\right) \qquad \text{and} \qquad q(z) \equiv \frac{1}{d(d-1)}\, g\left(\frac{z}{d-1}\right).
$$

Example: $R = 2, \ d = 3$

step 1
$\mathcal{S}_1$

$$\sum_{r=1}^{2} \lambda_r \, \rho_{1r}^3 - \mu_1 - \frac{1}{d-1} g\left(\frac{\mu_1}{d-1}\right) = 0$$

$$\sum_{r=1}^{2} \lambda_r \, \alpha_{1r} \, \rho_{1r}^2 - h(\mu_1) \, \rho_{11} = 0$$

$$\sum_{r=1}^{2} \lambda_r \, \alpha_{2r} \, \rho_{1r}^2 - h(\mu_1) \, \rho_{12} = 0$$

$$\sum_{r=1}^{2} \lambda_r \, \rho_{2r}^3 - \mu_2 - \frac{1}{d-1} g\left(\frac{\mu_2}{d-1}\right) - \mu_1 \, \eta_{12}^3 = 0$$

$$\sum_{r=1}^{2} \lambda_r \, \alpha_{1r} \, \rho_{2j}^2 - h(\mu_2) \, \rho_{21} - \mu_1 \, \rho_{11} \, \eta_{12}^2 = 0$$

$$\sum_{r=1}^{2} \lambda_r \, \alpha_{2r} \, \rho_{2r}^2 - h(\mu_2) \, \rho_{22} - \mu_1 \, \rho_{12} \, \eta_{12}^2 = 0$$

$$\sum_{r=1}^{2} \lambda_r \, \rho_{1r} \, \rho_{2r}^2 - h(\mu_2) \, \eta_{12} - [\mu_1 + q(\mu_1)] \, \eta_{12}^2 = 0$$

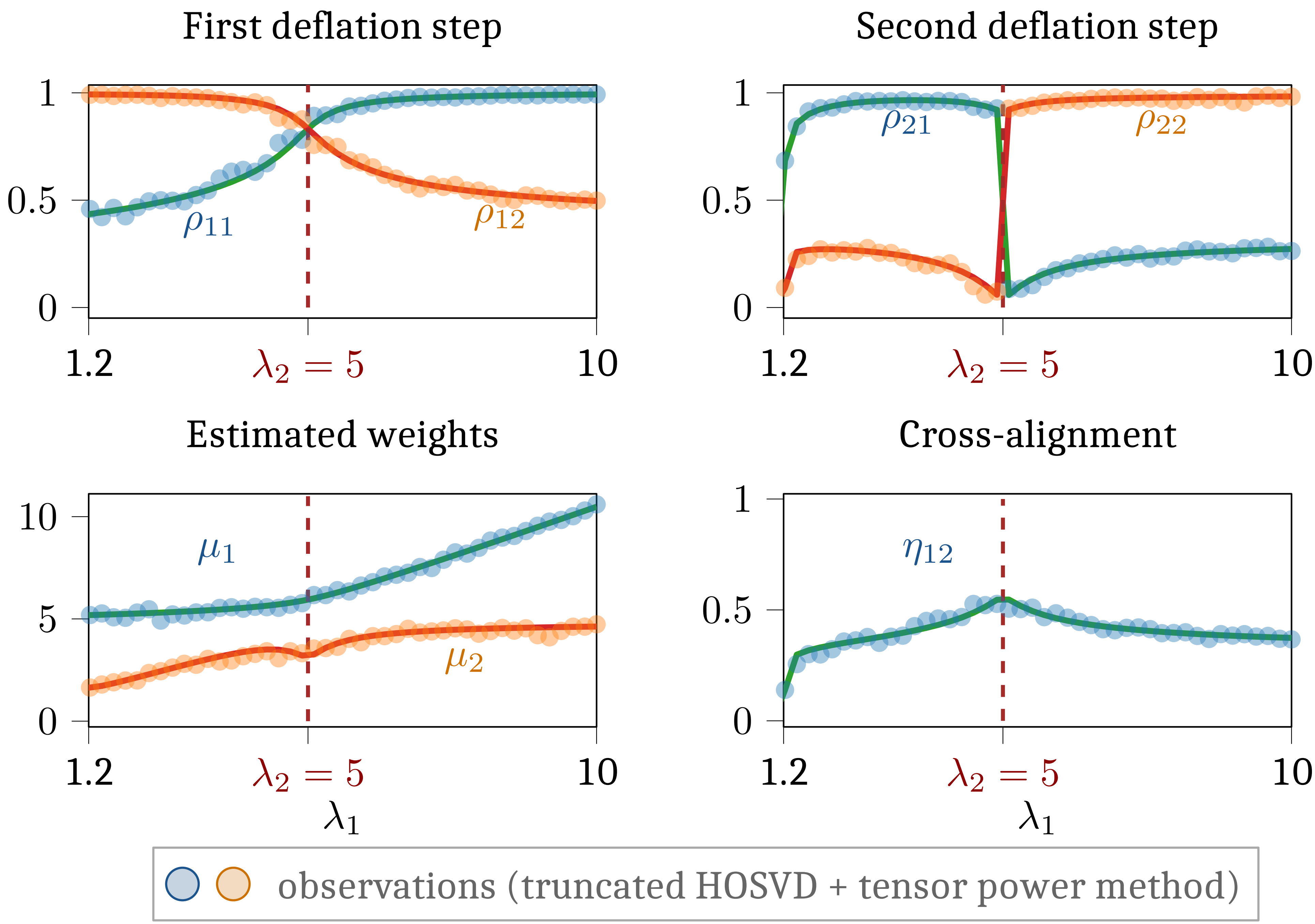
step 2
$\mathcal{S}_2$

model params:
$\lambda_1, \lambda_2, \alpha_{12}$

deflation result:
μ_1, μ_2, η_{12}

deflation performance:
$\rho_{11}, \rho_{12}, \rho_{21}, \rho_{22}$

Numerical results ($\alpha = 0.4$, $R = 2$, $d = 3$, $N = 100$)



This talk

1. Performance of maximum likelihood estimation
2. Leveraging RMT tools for MLE performance analysis
3. Rank-R case: performance of deflation
4. Epilogue and summary

Epilogue: extensions

1. Asymmetric CPD

$$\mathcal{Y} = \lambda\, x^{(1)} \otimes \cdots \otimes x^{(d)} + \frac{1}{\sqrt{N_1 + \cdots + N_d}}\, \mathcal{W}$$

Asymptotic alignment of MLE, rank-one case: Seddik–Guillaud-Couillet, 2024

Asymptotic analysis of deflation: Seddik–G.–Guillaud–Decurninge, presented at CAMSAP 2023

2. Nested tensor-matrix model[2,3], with application to multiview clustering:

$$\mathcal{Y} = \lambda\, (x \otimes y + M) \otimes z + \frac{1}{\sqrt{N}}\, \mathcal{W} \qquad \text{(ICLR'24)}$$

Summary

Rank-one symmetric tensor model: simple but quite rich

$$Y_{i_1\ldots i_d} = \lambda\, x_{i_1} \ldots x_{i_d} + \frac{1}{\sqrt{N}}\, W_{i_1\ldots i_d}$$

Statistical thresholds, MLE landscape and performance now well understood, largely thanks to spin glass theory.

Standard RMT tools can be leveraged by studying contractions and

- bring additional insights
- provide more elementary means of reaching some of those results
- are flexible and accessible for extensions/generalization

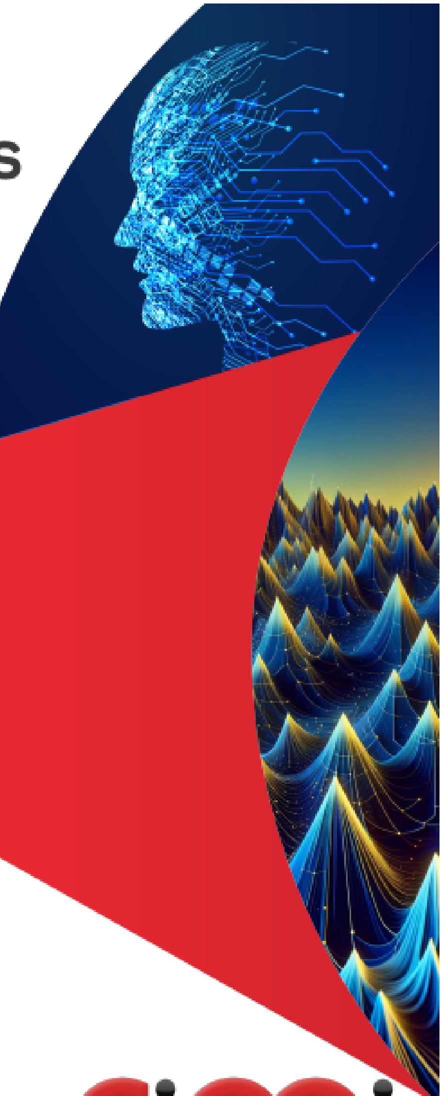
In particular, the RMT approach lends itself well to the asymptotic analysis of a simple deflation algorithm

Thanks!

See you next fall !

Thematic Trimester

Beyond classical regimes in statistical inference and machine learning



September to November 2024

Institut de Mathématiques de Toulouse - France

1 Colloquium
2 Thematic Schools &
1 Workshop,

- **Opening Colloquium**
  September 11th, 2024
- **Thematic School: Optimization & algorithms for high-dimensional machine learning and inference**
  October 7th to 11th, 2024
- **Thematic School: Models & methods for high-dimensional machine learning and inference**
  October 14th to 18th, 2024
- **Workshop**
  November 4th to 8th, 2024

cimi
Centre International de Mathématiques et d'Informatique
TOULOUSE

Organizers: Henrique Goulart (IRIT/Toulouse INP), Vanessa Kientz (CEA List), Vincent Lahoche (CEA List), Xiaoyi Mai (IMT/UT2J), Mohamed Tamaazousti (CEA List)

https://indico.math.cnrs.fr/category/682

For more details: JMLR 21-1038, arxiv:2304.10248

henrique.goulart @ irit.fr

Extensions (by others)

- Extension to asymmetric models by Seddik–Guillaud–Couillet (2024):

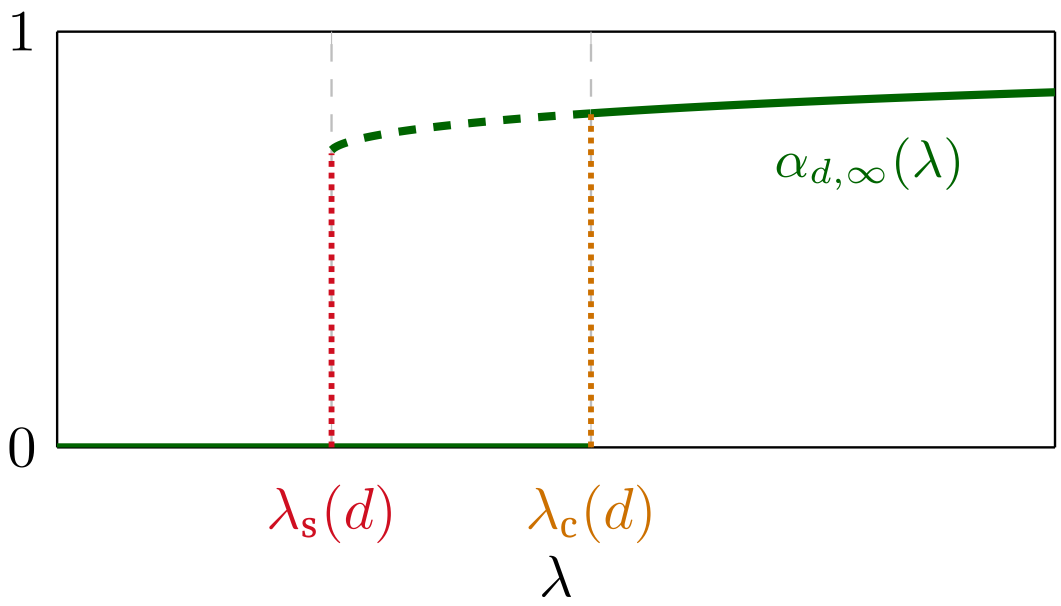
$$\mathcal{Y} = \lambda\, x^{(1)} \otimes \cdots \otimes x^{(d)} + \frac{1}{\sqrt{N_1 + \cdots + N_d}}\, \mathcal{W}$$

  with $W_{i_1 \ldots i_d} \sim \mathcal{N}(0, 1)$

- Joint spectral law of contracted ensembles for Wigner-type tensors: Au & Garza-Vargas (2021)

Open questions

- Proof that limiting equations describe MLE?
- Can we determine the phase transition with this approach?



- Can we extend it to higher-rank tensors?

Last but not least, the bibliography (1/2)

- Montanari, A., & Richard, E. A statistical model for tensor PCA. NIPS, 2014.
- Montanari, A., Reichman, D., & Zeitouni, O. On the Limitation of Spectral Methods: From the Gaussian Hidden Clique Problem to Rank One Perturbations of Gaussian Tensors. NIPS, 2015.
- Perry, A., Wein, A. S., & Bandeira, A. S. Statistical limits of spiked tensor models. AIHP-PS, 2020.
- Ben Arous, G., Mei, S., Montanari, A., & Nica, M. The Landscape of the Spiked Tensor Model. CPAM, 2019.
- Ros, V., Ben Arous, G., Biroli, G., & Cammarota, C. Complex Energy Landscapes in Spiked-Tensor and Simple Glassy Models: Ruggedness, Arrangements of Local Minima, and Phase Transitions. Physical Review X, 2019.
- Jagannath, A., Lopatto, P., & Miolane, L. Statistical thresholds for Tensor PCA. AAP, 2020.

Last but not least, the bibliography (2/2)

- Seddik, M. E. A., Guillaud, M., & Couillet, R. When Random Tensors meet Random Matrices. arXiv:2112.12348, 2021.
- Lim, L. H. Singular values and eigenvalues of tensors: A variational approach. CAMSAP, 2005.
- Pastur, L. & Shchrebina, M. Eigenvalue distribution of large random matrices. AMS, 2011.
- Au, B. & Garza-Vargas, J. Spectral asymptotics for contracted tensor ensembles. arXiv:2110.01652, 2021.

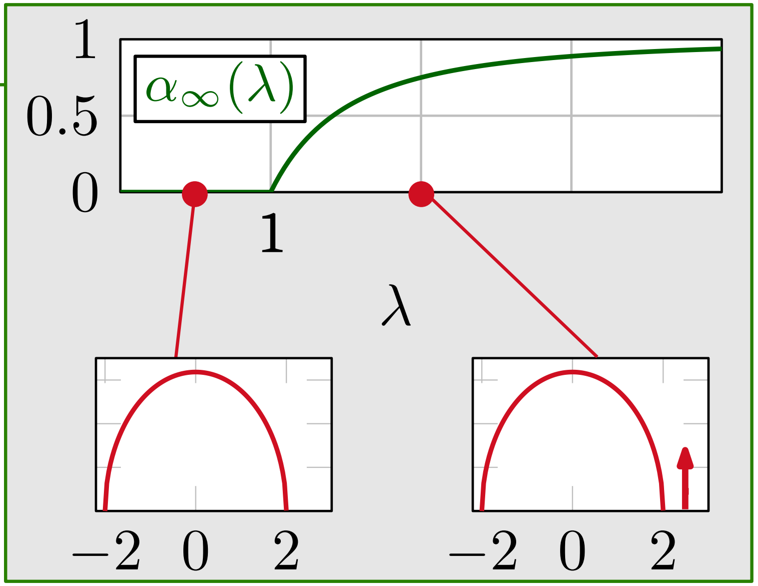
Random optimization landscape

Behavior reminiscent of "BBP phase transition" known for spiked matrix model

$$Y = \lambda\, x\, x^{\mathsf{T}} + \frac{1}{\sqrt{N}}\, W$$

(Benaych-Georges & Nadakuditi, 2011)

But why the discontinuity ?



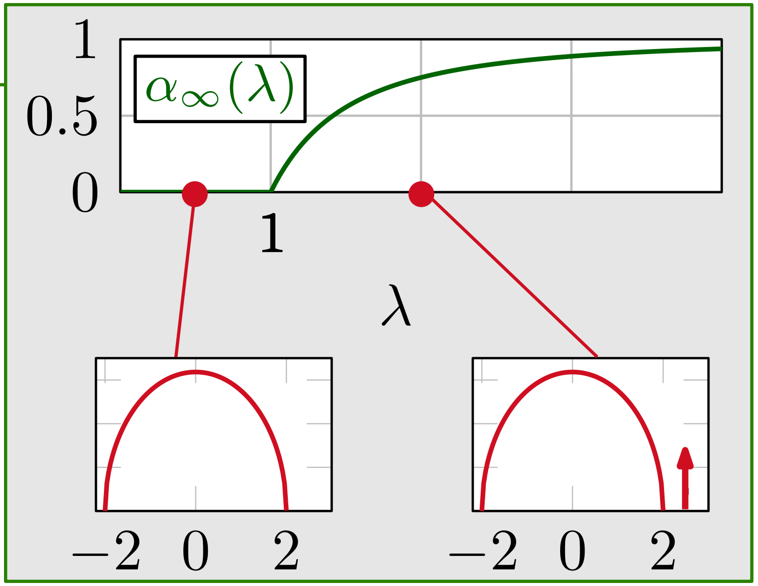
Random optimization landscape

Behavior reminiscent of "BBP phase transition" known for spiked matrix model

$$Y = \lambda\, x\, x^{\mathsf{T}} + \frac{1}{\sqrt{N}}\, W$$

(Benaych-Georges & Nadakuditi, 2011)

But why the discontinuity ?

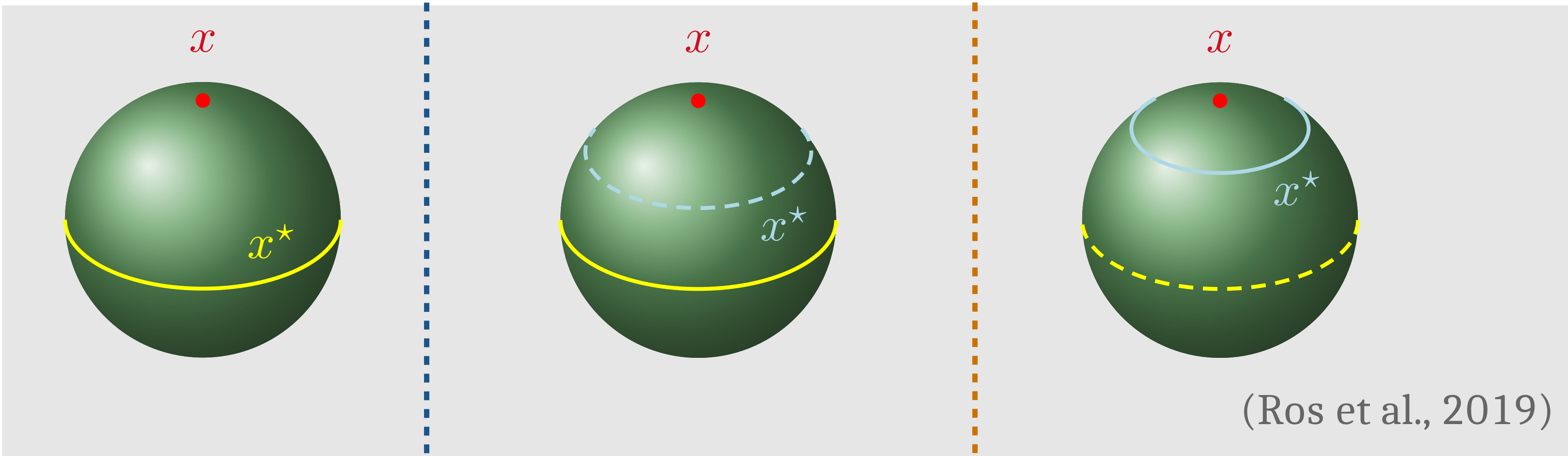


Insight found in study of the (random) ML landscape
(Ros et al., 2019) (Ben Arous et al., 2019)

- Quantification of "landscape complexity" (# of critical pts/local max)
- Connection with (spin) glasses and "rough energy landscapes"
- Configuration encoding signal competes with random ones

$$\mathcal{Y}(u,u,u) = \lambda\, \langle u, x\rangle^3 + \frac{1}{\sqrt{N}}\mathcal{W}(u,u,u)$$

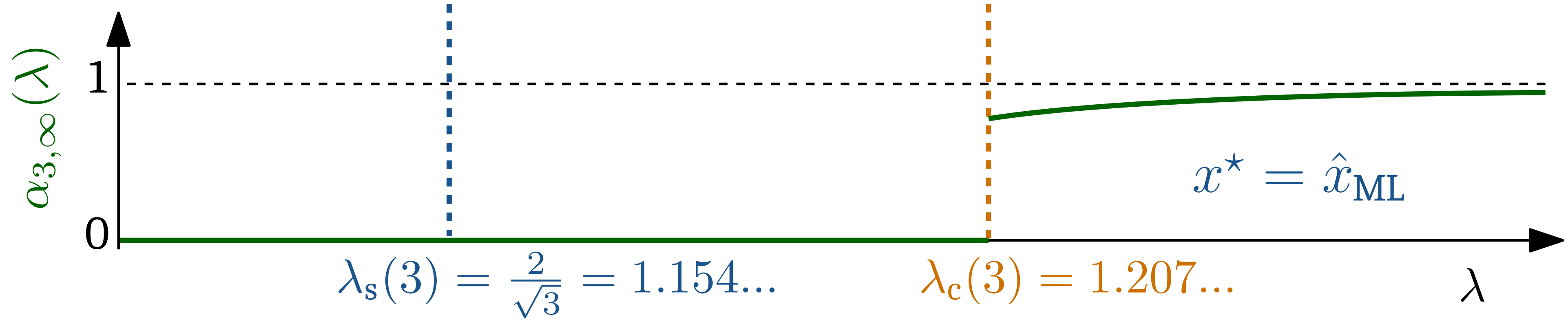
Geometric phase transitions



All max around equator

Global max on equator, some "informative" local max $x^{\star}$ "escape" towards north

Local max on equator, $x^{\star}$ become global and approach north



RMT to the rescue

Problem: Given a local max (μ, u) and

spectral dec.

$$\mathcal{Y}(u) = \lambda \langle x, u \rangle \, x \, x^\top + \frac{1}{\sqrt{N}} \mathcal{W}(u) \qquad = \sum_i \nu_i \, v_i v_i^\top$$

compute the limiting values of $\langle x, u \rangle$ and μ (if any).

Key tool: Resolvent of $\mathcal{Y}(u)$

$$R(z) := (\mathcal{Y}(u) - zI)^{-1} \qquad = \sum_i \frac{1}{\nu_i - z} v_i v_i^\top$$

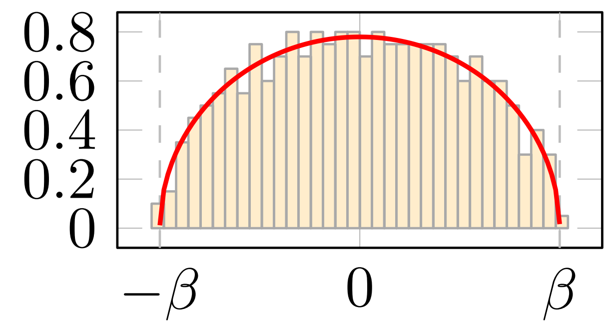
- Analytic on $\mathbb{C} \setminus \mathrm{Sp}(\mathcal{Y}(u))$
- For ν_i of multiplicity one, $\langle v_i, x \rangle^2 = -\frac{1}{2\pi i} \oint_{C_{\nu_i}} x^\top R(z) \, x \, dz$
- Encodes (random) spectral measure of $\mathcal{Y}(u)$

$$\frac{1}{N} \operatorname{tr} R(z) = \int \frac{1}{\nu - z} \rho_{\mathcal{Y}(u)}(d\nu), \qquad \rho_{\mathcal{Y}(u)} = \frac{1}{N} \sum_i \delta_{\nu_i}$$

Spectral measure of contraction ensemble $\{\mathbf{\mathcal{Y}}(v)\}$

“Byproduct”: limiting spectrum of $\mathcal{Y}(v)$, $v \in \mathbb{S}^{N-1}$

$$\rho(dx) = \frac{3}{\pi}\sqrt{\left[\frac{2}{3} - x^2\right]_+}\,dx$$



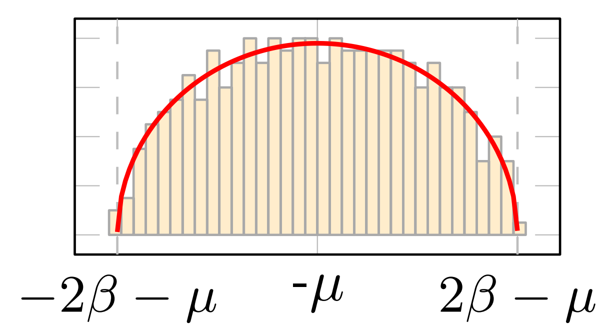
$\beta = \sqrt{2/3} = 0.8164...$

Seems trivial (Gaussian model), but symmetry induces dependencies

Consequences :

- At critical points, Hessian $2\,\mathcal{Y}(u) - \mu I$ behaves as a shifted GOE (Ros et al., 2019)

- At local maxima : $\mu \geq 2\beta$ (and $\mu \leq 1.657...$ for $\lambda < \lambda_c$)



The matrix case : predictions by RMT

Symmetric $N \times N$ rank-one spiked matrix model :

$$Y = \lambda\, x x^{\mathsf{T}} + \frac{1}{\sqrt{N}} W \qquad \|x\| = 1,\ \lambda > 0,\ W = W^{\mathsf{T}},$$

$$f(Y) \sim e^{-\frac{N}{4\sigma^2}\|Y - \lambda\, x x^{\mathsf{T}}\|_{\mathrm{F}}^2}$$

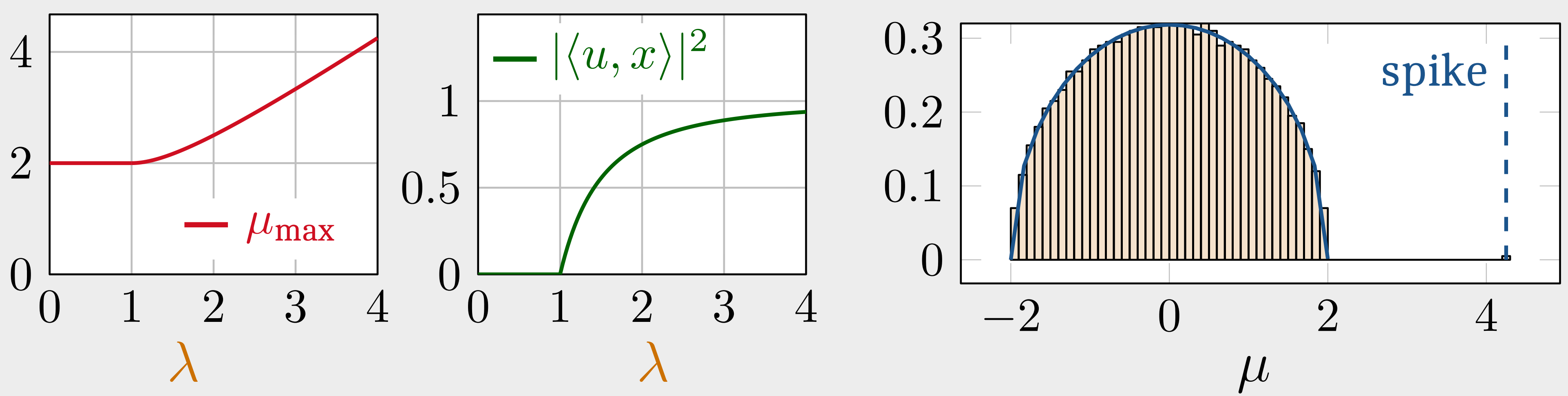
$\Rightarrow$ Maximum likelihood (ML) estimator of x $=$ dominant eigenvector of Y

Theorem. The largest eigenvalue of Y and its associated eigenvector satisfy

$$\mu_{\max}(\lambda) \xrightarrow[N\to\infty]{\text{a.s.}} \begin{cases} \lambda + \frac{\sigma^2}{\lambda}, & \lambda > \sigma, \\ 2\sigma, & \text{otherwise}, \end{cases} \qquad |\langle u, x\rangle|^2 \xrightarrow[N\to\infty]{\text{a.s.}} \begin{cases} 1 - \frac{\sigma^2}{\lambda^2}, & \lambda \geq \sigma, \\ 0, & \text{otherwise}. \end{cases}$$

Ex.: $\sigma = 1$, $N = 2000$

Emp. spectrum & semi-circle law



Stieltjes transform

Definition. The Stieltjes transform of a probability measure ρ is

$$s_\rho(z) := \int \frac{1}{t-z}\,\rho(dt), \qquad z \in \mathbb{C} \setminus \mathrm{supp}(\rho)$$

Theorem. (*Stieltjes*) The sequence of random probability measures ρ_N converges almost surely weakly to the deterministtic measure ρ iff

$$s_{\rho_N}(z) \xrightarrow[N\to\infty]{\text{a.s.}} s_\rho(z) \quad \text{for all} \quad z \in \mathbb{R} + i\,\mathbb{R}_+^*$$

Properties :

- s_ρ is analytic on its domain
- If $\mathrm{supp}(\rho)$ is bounded, then $s_\rho(z) \underset{z\to\pm\infty}{\longrightarrow} 0$
- If ρ has a density f, then by the Stieltjes-Perron inversion formula :

$$f(x) = \lim_{\epsilon\to 0^+} (2i\pi)^{-1}\,\left[s_\rho(x-i\epsilon) - s_\rho(x+i\epsilon)\right]$$

Resolvent matrix

In particular :

$$s_{\rho_N}(z) = \int \frac{1}{t-z}\,\rho_N(dt) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\mu_i(Y_N)-z} = \frac{1}{N}\,\mathrm{tr}\,(Y_N - zI)^{-1}$$

Definition. Resolvent of a symmetric matrix Y :

$$R(z) := (Y - zI)^{-1}, \qquad z \in \mathbb{C} \setminus \sigma(Y)$$

Properties :

- Spectral decomposition : $R_N = \sum_{i=1}^{N} \frac{1}{\mu_i(Y_N)-z}\, u_i\, u_i^\top$
- tr $R_N(z)$ analytic on $\mathbb{C} \setminus \sigma(Y_N)$
- If $\mu_i(Y_N)$ is an isolated eigenvalue, then

$$P_{\mu_i} = -\frac{1}{2\pi i}\oint_{C_{\mu_i(Y_N)}} R_N(z)\,dz$$

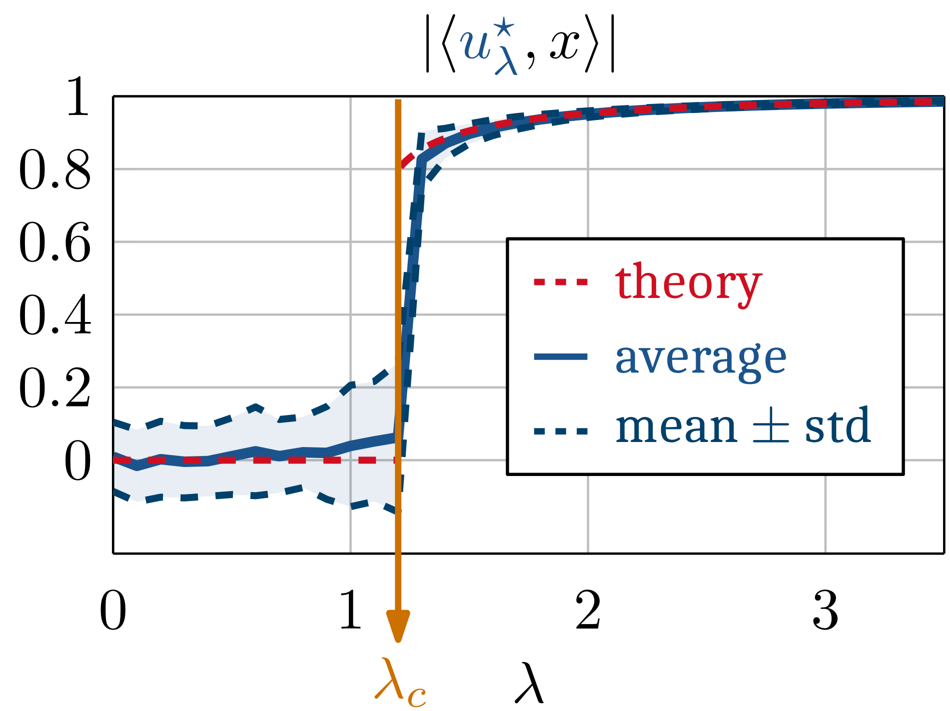
is an orthogonal projector onto the eigenspace associated with $\mu_i(Y_N)$

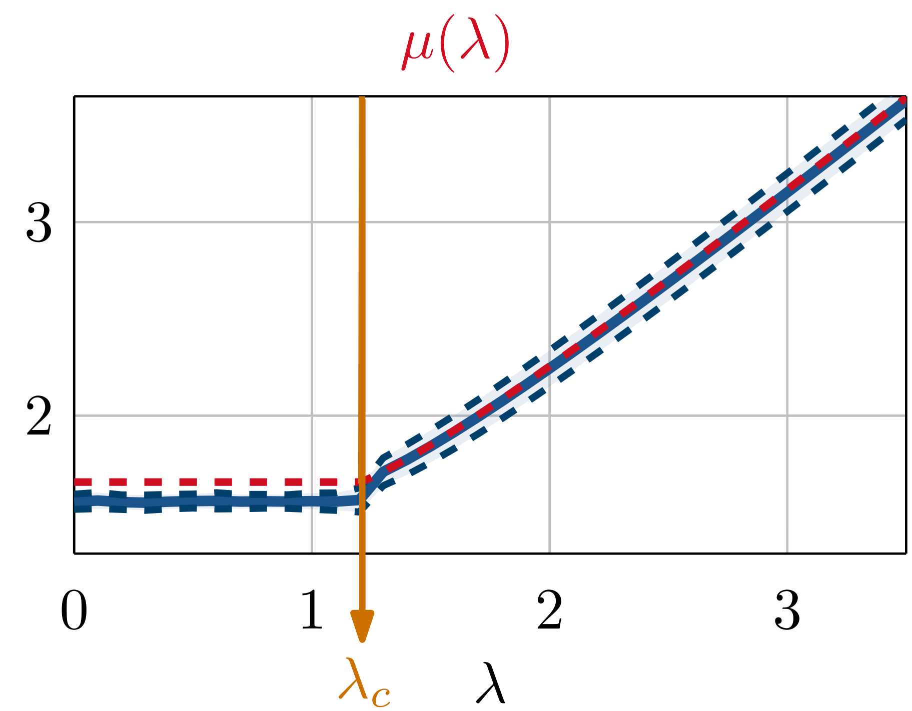
Numerical experiment, $N = 100$

Shifted high-order power method : (Kolda & Mayo, 2011)

1. $z^{t+1} = \boldsymbol{\mathcal{Y}}(u^t, u^t) + \alpha\, u^t = \sum_{jk} y_{:jk}\, u_j^t\, u_k^t + \alpha\, u^t$
2. $u^{t+1} = \|z^{t+1}\|^{-1}\, z^{t+1} \qquad \left(\mu^{t+1} = \boldsymbol{\mathcal{Y}}(u^{t+1}, u^{t+1}, u^{t+1}) \right)$

- Local search for $\arg\max_{u \in \mathbb{S}^{N-1}} \boldsymbol{\mathcal{Y}}(u, u, u)$ (exponentially many local maxima)
- Initialization : $u^0 = x$ for $\lambda > \lambda_c$ and u^0 random on $\mathbb{S}^{N-1} \cap x^\perp$ otherwise





$\mathcal{W} \in$ tensor Gaussian orthogonal ensemble $\qquad Q \in \mathsf{O}(N)$

$$p(\mathcal{W}) = \frac{1}{Z_d(N)} \exp\left(-\frac{1}{2}\|\mathcal{W}\|_{\mathsf{F}}^2\right) = p((Q, \ldots, Q) \cdot \mathcal{W})$$

Consequences:

1. $\mathrm{Var}(W_{i_1 \ldots i_d})$ depends on the pattern of repetitions in $(i_1, \ldots, i_d)$, e.g.:

$$\text{for } d = 3, \quad \|\mathcal{W}\|_{\mathsf{F}}^2 = \sum_i W_{iii}^2 + 3 \sum_{i<j} (W_{iij}^2 + W_{ijj}^2) + 6 \sum_{i<j<k} W_{ijk}^2$$

2. Law of $\mathcal{Y}$: $\quad p(\mathcal{Y} \mid x) \sim \exp\left(-\frac{N}{2}\left\|\mathcal{Y} - \lambda\, x^{\otimes d}\right\|_{\mathsf{F}}^2\right)$

Maximum likelihood estimator (MLE):

$$\hat{x} := \operatorname*{arg\,min}_{\|u\|=1} \left\|\mathcal{Y} - \lambda\, u^{\otimes d}\right\|_{\mathsf{F}}^2 = \operatorname*{arg\,max}_{\|u\|=1} \left\langle \mathcal{Y}, u^{\otimes d} \right\rangle$$

Technical tools

Computation of $\mathbb{E}\left\{\mathcal{W}\cdot u^d\right\}$ and $\mathbb{E}\left\{\frac{1}{\mu}\left\langle x, \mathcal{W}\cdot u^{d-1}\right\rangle\right\}$:

1. Gaussian integration-by-parts

$$z \sim \mathcal{N}(0,\sigma^2) \quad \Rightarrow \quad \mathbb{E}\left\{z\, f(z)\right\} = \sigma^2\,\mathbb{E}\left\{f'(z)\right\}$$

$$\mathbb{E}\left\{\mathcal{W}\cdot u^d\right\} = \sum_{\boldsymbol{i}} \mathbb{E}\left\{W_{\boldsymbol{i}}\, u_{i_1}\dots u_{i_d}\right\} = \sum_{\boldsymbol{i}} \sigma_{\boldsymbol{i}}^2 \sum_{j=1}^{d} \mathbb{E}\left\{\frac{\partial u_{i_j}}{\partial W_{\boldsymbol{i}}} \prod_{k\neq j}^{d} u_{i_k}\right\}$$

$$\mathbb{E}\left\{\frac{1}{\mu}\left\langle x, \mathcal{W}\cdot u^{d-1}\right\rangle\right\} = \sum_{\boldsymbol{i}} \sigma_{\boldsymbol{i}}^2\,\mathbb{E}\left\{-\frac{1}{\mu^2}\frac{\partial \mu}{\partial W_{\boldsymbol{i}}}\left(\prod_{k=1}^{d-1} u_{i_k}\right) x_{i_d} + \frac{1}{\mu} x_{i_d} \sum_{j=1}^{d-1} \frac{\partial u_{i_j}}{\partial W_{\boldsymbol{i}}} \prod_{k\neq j}^{d-1} u_{i_k}\right\}.$$

2. Implicit function theorem to compute the required derivatives :

$$\begin{pmatrix} \dfrac{\partial u}{\partial W_{\boldsymbol{i}}} \\[2ex] \dfrac{\partial \mu}{\partial W_{\boldsymbol{i}}} \end{pmatrix} = \begin{pmatrix} -\dfrac{1}{(d-1)\sqrt{N}}\, R\left(\dfrac{\mu}{d-1}\right)\phi + \dfrac{1}{(d-2)\,\mu}\dfrac{\partial \mu}{\partial W_{\boldsymbol{i}}}\, u \\[2ex] \dfrac{1}{\sigma_{\boldsymbol{i}}^2\sqrt{N}} \displaystyle\prod_{j=1}^{d} u_{i_j} \end{pmatrix}$$

Byproduct: limiting spectrum of contractions

Theorem. Take a sequence of vectors $v \in \mathbb{S}^{N-1}$ and of Gaussian tensors $\mathcal{W}$.

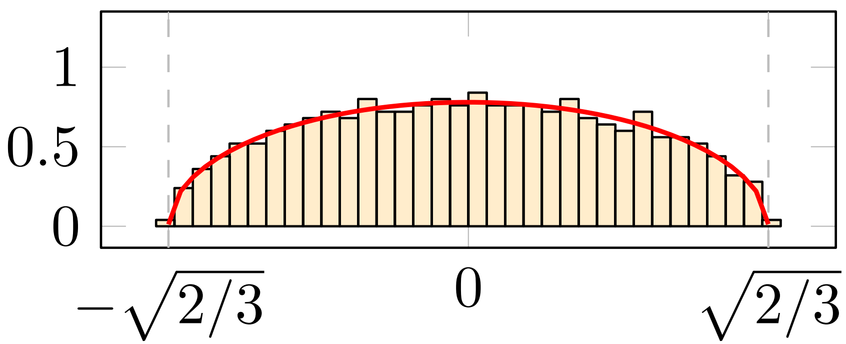
Define the empirical spectral measure $\rho_N = \frac{1}{N}\sum_{i=1}^N \delta_{\nu_i}$, where

$$\{\nu_i\}_{i=1}^N = \mathrm{Sp}\left(\frac{1}{\sqrt{N}}\mathcal{W}\cdot v^{d-2}\right)$$

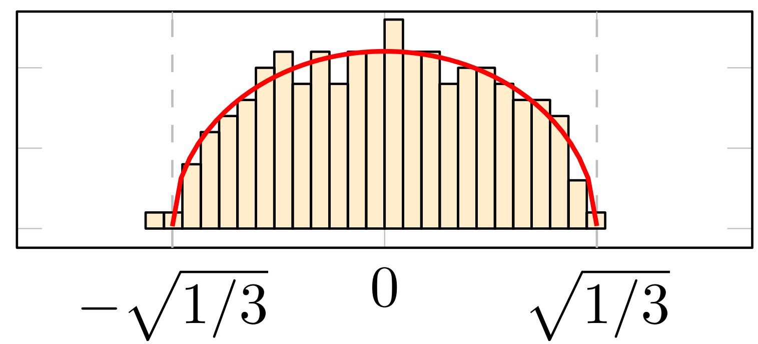
Then, ρ_N converges a.s. weakly to a semicircle law ρ with density

$$\rho(dx) = \frac{2}{\pi\,\beta_d^2}\sqrt{\left(\beta_d^2 - x^2\right)^+}\,dx, \qquad \beta_d := \frac{2}{\sqrt{d(d-1)}}.$$

$d = 3$, $N = 500$



$d = 4$, $N = 200$



But why does this work?

Ingredients: tensor eigenvalue equation satsfied by all critical points and a.s. convergence

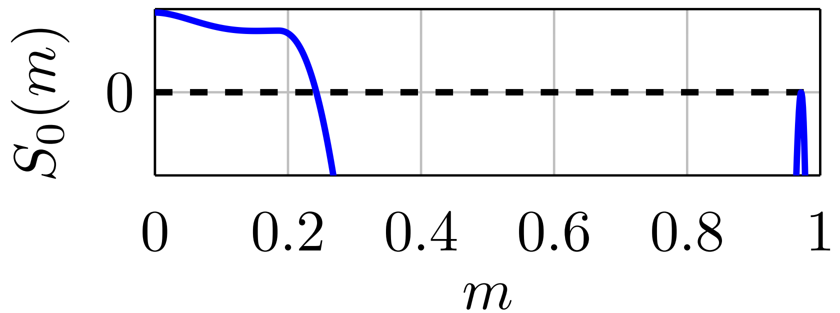
$$\langle x, u \rangle \xrightarrow[N\to\infty]{\text{a.s.}} \alpha_{d,\infty}(\lambda) > 0$$

$$\mu \xrightarrow[N\to\infty]{\text{a.s.}} \mu_{d,\infty}(\lambda) > (d-1)\beta_d$$

Output: description of global maximizer(s) for $\lambda > \lambda_{\text{c}}(d)$.

The conditions on $\alpha_{d,\infty}$ and $\mu_{d,\infty}$ could be met by some strictly local max when $\lambda > \lambda_{\text{c}}(d)$, as per Ben Arous–Mei–Montanari (2019)

expected # local max with alignment m $\sim \exp(N\, S_0(m))$



Our take: The "selectivity" probably comes from the a.s. conv. assumption.