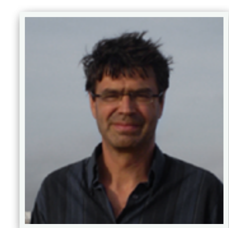
UNIRANDOM 2021 workshop in Rennes

Maximum of Gaussian processes on Stiefel manifold for spiked tensor models

Yohann DE CASTRO (Institut Camille Jordan)
joint work with J.-M. AZAÏS (IMT) and F. DALMAO (UR)



UNIVERSITÉ DE LYON

ÉCOLE CENTRALE LYON

Outline

- Spiked Wigner tensors
- Phase transition for one-spiked model
- Extension to two-spiked model
- Open questions

- k-way n-dimensional symmetric tensors $(k \geq 3, n \geq 4)$

$$\forall \boldsymbol{T}, \boldsymbol{U} \in (\mathbb{R}^n)^{\otimes k}, \quad \langle \boldsymbol{T}, \boldsymbol{U} \rangle = \sum_{i_1, \ldots, i_k \in [n]} T_{i_1, \ldots, i_k} U_{i_1, \ldots, i_k}$$

- Spiked (Wigner) tensors

$$\boxed{\boldsymbol{Y} = \lambda \boldsymbol{\sigma} + \boldsymbol{W}}$$

where $\lambda \geq 0$ *signal-to-noise ratio* (SNR), $\boldsymbol{\sigma} \in (\mathbb{R}^n)^{\otimes k}$ unit norm, $\boldsymbol{W}$ such that

$$\boldsymbol{W} = \frac{1}{k!} \sqrt{\frac{2}{n}} \sum_{\pi \in \mathfrak{S}_n} \boldsymbol{G}^{\pi}$$

with i.i.d standard Gaussians $G_{i_1 \cdots i_k}$ for indices $1 \leq i_1, \ldots, i_k \leq n$, $\mathfrak{S}_n$ is the set of permutations of the set $[n]$, and $\boldsymbol{G}^{\pi}_{i_1 \cdots i_k} = G_{\pi(i_1) \cdots \pi(i_k)}$.

The one-spiked tensor model (Richard et Montanari, 2014) is given by

$$\boxed{\boldsymbol{\sigma} = \boldsymbol{\sigma}_1 := \mathbf{u}^{\otimes k}}$$

and one would like to estimate the unit vector $\mathbf{u} \in \mathbb{S}^{n-1}$.

MLE is the arg maximum of the *tensor form* over the sphere given by

$$\forall \boldsymbol{T} \in (\mathbb{R}^n)^{\otimes k}\,, \quad f_{\boldsymbol{T}} : \mathbf{u} \in \mathbb{S}^{n-1} \mapsto \langle \boldsymbol{T}, \mathbf{u}^{\otimes k} \rangle\,,$$

and one can prove that the tensor form $\boldsymbol{T} \to f_{\boldsymbol{T}}$ is injective on the space of k-way n-dimensional symmetric tensors and the one-spiked MLE is given by:

$$\widehat{\mathbf{u}} = \arg \max_{\mathbf{u} \in \mathbb{S}^{n-1}} f_Y(\mathbf{u})\,.$$

The null hypothesis ($\lambda = 0$) is the **spherical k-spin spin glasses model**.

One-Spiked model: Information goals

- *Weak detection* One can distinguish between $\mathbb{P}_{\lambda\boldsymbol{\sigma}}$ and $\mathbb{P}_0$ with success probability $1/2+\varepsilon$ (the so-called "*better than random*") as n goes to infinity, for some $\varepsilon > 0$ that does not depend on n;

- *Strong detection* One can distinguish between $\mathbb{P}_{\lambda\boldsymbol{\sigma}}$ and $\mathbb{P}_0$ with success probability $1-o_n(1)$ as n goes to infinity;

- *Weak recovery* holds if there exists a measurable function (an estimator) $\widehat{\mathbf{u}} : (\mathbb{R}^n)^{\otimes k} \to \mathbb{S}^{n-1}$ such that

$$\liminf_{n\to\infty} \mathbb{E}|\langle \widehat{\mathbf{u}}, \mathbf{u}\rangle| \geq \varepsilon\,,$$

  for some $\varepsilon > 0$;

- *Strong recovery* holds if

$$\liminf_{n\to\infty} \mathbb{E}|\langle \widehat{\mathbf{u}}, \mathbf{u}\rangle| = 1\,.$$

Theorem 1 *Consider the one-spiked model such that* $\mathbf{u} \in \mathbb{S}^{n-1}$ *is drawn according to the spherical prior* $\mathcal{U}_{\mathrm{sph}}$. *Then, there exist bounds* $\lambda^\star_{\mathrm{low}} < \lambda^\star_{\mathrm{high}}$ *both behaving as* $\sqrt{2\log k} + o_k(1)$ *as* $k \to \infty$ *such that*

- *if* $\lambda < \lambda^\star_{\mathrm{low}}$ *then weak detection and weak recovery are impossible;*
- *if* $\lambda > \lambda^\star_{\mathrm{high}}$ *then strong detection and weak recovery are possible;*

and it holds

$$(\lambda^\star_{\mathrm{low}})^2 = 2\log k + 2\log\log k + o_k(1)$$
$$(\lambda^\star_{\mathrm{high}})^2 = 2\log k + 2\log\log k + 2 - 4\log 2 + o_k(1)\,,$$

in the limit $k \to \infty$.

$$\lambda^\star_{\mathrm{low}} \simeq \lambda^\star_{\mathrm{high}} \simeq \lambda^\star_{\mathrm{mle}} \simeq \sqrt{2\log k}$$

Interlude: Landscape of one-spiked tensor model

- **Computational gap:** no polynomial time algorithm is known to achieve *weak recovery* for

$$\lambda \ll \lambda_{\text{comp}} = \mathcal{O}_n\big(n^{\frac{k-2}{4}}\big)$$

and *power iteration* for MLE estimation fails for

$$\lambda \ll \lambda_{\text{power}} = \mathcal{O}_n\big(n^{\frac{k-2}{2}}\big)$$

- (Ben Arous et al., 2019) showed that *all the local maxima are either very close to the unknown vector* $\mathbf{u}$ *(and to the global maximum) or they are on a narrow spherical annulus orthogonal to* $\mathbf{u}$ *contained in* $|\langle \mathbf{u}, \sigma\rangle| \leq \Theta(\lambda^{-1/(k-2)})$.

- *Population risk* $\mathbb{E} f_Y(\sigma^{\otimes k}) = \lambda \langle \mathbf{u}, \sigma\rangle^k$ has large flat regions on the orthogonal of $\mathbf{u}$. Note that most of the volume concentrates around this hyperplane.

- Random initialization ensures that $|\langle \mathbf{u}, \sigma_0\rangle| = \Theta(n^{-1/2})$ and it will not be trapped in flat regions as soon as $\lambda \gtrsim n^{(k-2)/2}$.

- Multiple spiked model of rank $r \geq 1$:

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}_r := \sum_{i=1}^{r} a_i \mathbf{u}_i^{\otimes k} \,,$$

where $\mathbf{u}_i \in \mathbb{S}^{n-1}$ are orthogonal unit norm vectors and $\mathbf{a} := (a_i) \in \mathbb{S}^{r-1}$ has unit norm (hence the tensor $\boldsymbol{\sigma}_r$ has unit Euclidean norm).

Theorem 1 *MLE $\widehat{\boldsymbol{\sigma}}_r$ is unique a.s., it has rank exactly r, and it holds*

$$\widehat{\boldsymbol{\sigma}}_r = \widehat{\lambda} \times \sum_{\ell=1}^{r} \widehat{\alpha}_\ell \widehat{\mathbf{x}}_\ell^{\otimes k} \,,$$

where $$(\widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{x}}) \in \arg\max \Big\{ \sum_{\ell=1}^{r} \alpha_\ell f_Y(\mathbf{x}_\ell) \Big\} \,,$$

$$\text{subject to} \quad (\boldsymbol{\alpha}, \mathbf{x}) \in \mathbb{S}^{r-1} \times (\mathbb{S}^{n-1})^{\otimes r} \, s.t. \; \mathbf{x}_1 \perp \mathbf{x}_2 \perp \ldots \perp \mathbf{x}_r \,,$$

and $\widehat{\lambda} = \sum_{\ell=1}^{r} \widehat{\alpha}_\ell f_Y(\widehat{\mathbf{x}}_\ell)$.

Denote

$$\boldsymbol{\Theta}_{n,r} := \Big\{(\boldsymbol{\alpha};\mathbf{x}_1,\ldots,\mathbf{x}_r)\in\mathbb{R}^r\times(\mathbb{R}^n)^r \ :\ \boldsymbol{\alpha}^\top\boldsymbol{\alpha}=1 \text{ and } \boldsymbol{U}^\top\boldsymbol{U}=\mathrm{Id}_r\Big\},$$

the manifold given by the Euclidean sphere times the standard Stiefel manifold. Denote $\boldsymbol{\theta}=(\boldsymbol{\alpha};\boldsymbol{U})$ a point of $\boldsymbol{\Theta}_{n,r}$, we define

$$Z_{\boldsymbol{W}}(\boldsymbol{\theta}) := \sum_{\ell=1}^{r}\alpha_\ell\langle \boldsymbol{W},\mathbf{x}_\ell^{\otimes k}\rangle = \sum_{\ell=1}^{r}\alpha_\ell f_{\boldsymbol{W}}(\mathbf{x}_\ell)\,.$$

The null hypothesis is introduced as the **Stiefel k-spin spin glasses model**.

Goal: Non-asymptotic tail bound of the maximum of this Gaussian process

$$M := \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_{n,r}} Z_{\boldsymbol{W}}(\boldsymbol{\theta})$$

Two Spiked model: tractable Kac-Rice formula

- (Edelman et al., 1998) gives the Riemannian toolbox for Stiefel manifolds.
- (Azaïs and Wschebor, 2008) gives the Kac-Rice formula toolbox.
- Stiefel manifold ($k = 2$) is $\boldsymbol{M}_n = \big\{(x, y) \in (\mathbb{S}^{n-1})^2 : x \perp y\big\}$.
- It holds $Z_{\boldsymbol{W}} = \cos\varphi f_{\boldsymbol{W}}(x) + \sin\varphi f_{\boldsymbol{W}}(y)$.

- By Kac-Rice formula $\mathbb{P}_0(M > \lambda_n)$ is upper bounded by

$$\mathrm{vol}(\boldsymbol{M}_n)\int_{\lambda_n}^{+\infty}\int_0^{2\pi}\mathbb{E}\left[\mathrm{Det}\big(\mathrm{Hess}Z_{\boldsymbol{W}}(\varphi,\mathbf{n})\big)\,\Big|\,Z_{\boldsymbol{W}}=v,\mathrm{grad}Z_{\boldsymbol{W}}=0\right]p_{(Z_{\boldsymbol{W}},\mathrm{grad}Z_{\boldsymbol{W}})}(v,0)\mathrm{d}\varphi\mathrm{d}v$$

where $\mathrm{vol}(\boldsymbol{M}_n)$ volume of $\boldsymbol{M}_n$, $\mathrm{grad}Z_{\boldsymbol{W}}, \mathrm{Hess}Z_{\boldsymbol{W}}$ Riemannian gradient and hessian, $p_{(Z_{\boldsymbol{W}},\mathrm{grad}Z_{\boldsymbol{W}})}$ joint distribution.

- For $1 \ll \lambda_n$ one has

$$\mathbb{P}_0(M \leq \lambda_n) = 1 - o_n(1)\,.$$

Multiple Spiked model: open questions and their lead

- **Give the landscape:** compute the exponential growth rate of the number of critical points (and local maxima) of $Z_{\boldsymbol{Y}}$ using Kac-Rice formula on $\mathbb{S}^1 \times \boldsymbol{M}_n$, as done by (Ben Arous et al., 2019) for the sphere.

- **Give the lower and upper bounds** $\lambda_{\mathrm{low,high}}$ using the so-called "second moment method" as in (Montanari, Reichman and Zeitouni, 2015).

- **Study larger models of rank** $r \geq 3$ using Riemannian toolbox of (Edelman et al., 1998).

The Landscape of the Spiked Tensor Mode

GÉRARD BEN AROUS
Courant Institute
SONG MEI
Institute for Computational and Mathematical Engineering, Stanford University
ANDREA MONTANARI
Department of Electrical Engineering and Statistics, Stanford University
AND
MIHAI NICA
University of Toronto

On the Limitation of Spectral Methods: From the Gaussian Hidden Clique Problem to Rank-One Perturbations of Gaussian Tensors

Andrea Montanari
Department of Electrical Engineering and Department of Statistics. Stanford University.
montanari@stanford.edu

Daniel Reichman
Department of Cognitive and Brain Sciences, University of California, Berkeley, CA
daniel.reichman@gmail.com

Ofer Zeitouni
Faculty of Mathematics, Weizmann Institute, Rehovot 76100, Israel
and Courant Institute, New York University
ofer.zeitouni@weizmann.ac.il

THE GEOMETRY OF ALGORITHMS WITH ORTHOGONALITY CONSTRAINTS*

ALAN EDELMAN†, TOMÁS A. ARIAS‡, AND STEVEN T. SMITH§

Abstract. In this paper we develop new Newton and conjugate gradient algorithms on the Grassmann and Stiefel manifolds. These manifolds represent the constraints that arise in such areas as the symmetric eigenvalue problem, nonlinear eigenvalue problems, electronic structures computations, and signal processing. In addition to the new algorithms, we show how the geometrical framework gives penetrating new insights allowing us to create, understand, and compare algorithms. The theory proposed here provides a taxonomy for numerical linear algebra algorithms that provide a top level mathematical view of previously unrelated algorithms. It is our hope that developers of new algorithms and perturbation theories will benefit from the theory, methods, and examples in this paper.

Key words. conjugate gradient, Newton's method, orthogonality constraints, Grassmann manifold, Stiefel manifold, eigenvalues and eigenvectors, invariant subspace, Rayleigh quotient iteration, eigenvalue optimization, sequential quadratic programming, reduced gradient method, electronic structures computation, subspace tracking