High-dimensional optimization for the multi-spike tensor PCA problem.

Cédric Gerbelot (ENS Lyon).
Based on joint work with Gérard Ben Arous (Courant)
and Vanessa Piccolo (ENS Lyon → EPFL).

Goal: recover r orthogonal spikes $v_1, \dots, v_r \in \mathbb{S}^{N-1}(1)$ for M noisy observations of the form

$$Y^\ell = W^\ell + \sum_{i=1}^{r} \sqrt{N}\lambda_i v_i^{\otimes p} \quad 1 \leq \ell \leq M \quad , \quad r \text{ fixed, independent on } N.$$

where $W^\ell \in (\mathbb{R}^N)^{\otimes p}$ have i.i.d. subGaussian entries, $p \geq 2$ and $\lambda_1 \geq \dots \geq \lambda_r$ are the signal-to-noise ratios (SNRs). Assume p and r are known.

To solve this problem: gradient flow, (Langevin dynamics) and online SGD on the objective function defined by Gaussian maximum likelihood:

$$\mathcal{E}(X) = \frac{1}{M}\sum_{\ell=1}^{M} \| Y^\ell - \sum_{i=1}^{r}\lambda_i x_i^{\otimes p} \|_F^2 \; , \quad \left(\text{if choose } \sum_{i=1}^{r} x_i^{\otimes p} \text{ doesn't change much}\right).$$

where $X = [x_1 | \dots | x_r] \in \mathbb{R}^{N \times r}$ is constrained to the Stiefel manifold

$$St_N(1,r) = \{ X \in \mathbb{R}^{N\times r} : X^T X = I_r \}.$$

Expanding $\mathcal{E}(X)$, we obtain

$$\mathcal{E}(X) = \underbrace{\frac{1}{M}\sum_{\ell=1}^{M}\sum_{i=1}^{r} \lambda_i \langle W^\ell, x_i^{\otimes p}\rangle}_{\text{noise part } H_0(X)} - \underbrace{\sum_{1 \leq i,j \leq r} \sqrt{N}\lambda_i\lambda_j m_{ij}^{p-1}(X)}_{\text{signal part } \phi(X)},$$

where

$$m_{ij}(X) = \langle v_i, x_j \rangle \text{ is the correlation between } v_i \text{ and } x_j.$$

Upon appropriate control of the noise, can study autonomous, low-dimensional (r^2) dynamics on the $\{m_{ij}(X)\}_{1 \le i,j \le r}$ $\Longrightarrow$ effective dynamics.

Large body of recent works on single and multi-index models used as templates to understand high-dimensional, non-convex optimization. We had wonderful talks by Aukosh, summary statistics with Reza & Gerard, - - - etc - - -, talks by members of Florent's group, talk by Eshaan and collaborators, wonderful lecture by Bruno, etc...

In multi-index case, many works use various "oracle" modifications of dynamics and focus on achieving positive correlation with the target subspace.

Here, simple model but try to understand entire dynamics precisely: (more modest than two layer neural network)

- understand fixed points, no "oracle" modification:
  - perfect recovery: $x_i = (1 - o(1)) v_i$
  - recovery of a permutation $x_i = (1 - o(1)) v_{\sigma(i)}$, what permutation?
  - recovery of the good subspace: $\mathrm{dist}(XX^T, VV^T) = o(1)$.
  - recovery of a rank deficient subspace.

- with what time and sample complexity, starting from uninformative initialization.

Background: the single spike tensor PCA problem.

Recover $v \in \mathbb{S}^{d-1}(1)$ from noisy observations of the form

$$Y^\ell = W^\ell + \lambda \sqrt{N} v^{\otimes p} \quad 1 \le \ell \le M.$$

For $p = 2$, matrix PCA [Johnstone 2002], for $p \geq 3$ introduced by [Montanari & Richard 2014]. For first order methods curvature of landscape is the main obstacle:
↳ $M \gg N^{p-1}$ is enough, conjectured N^{p-2}. (full batch).

- theoretical computer science & statistics : low degree polynomials, sum-of-squares ...
what is hard to compute, with what algorithm, what method is optimal, in what sense
[Hopkins, Shi & Steurer 2015], [Perry, Wein & Bandeira 2020]. Mention talks by Alex, Theo

- probability and mathematical physics : static and dynamical questions on high-dimensional random landscapes : complexity (# of critical points, ...), behaviour of Langevin dynamics on such landscapes, etc ... links with spin glass theory ...
[Ben Arous, Mei, Montanari and Nica '19], [Ben Arous, Jagannath & Gheissari 2020+]
↳ proved N^{p-2} for (full-batch) gradient flow.

- statistical physics : similar to the above using different (heuristic) methods
[Sarao, Urbani, Krzakala & Zdeborova 2020+].
important to mention multispike extension, harder.

⟶ Here, no search for an optimal algorithm, i.e. low-degree polynomials or spectral method, again, sandbox to understand high-dimensional, non-convex optimization.

Back to the multispike tensor PCA problem.

Gradient flow: Maybe do this part in the end.

Recall
$$\inf_{X \in St_N(1,r)} H_0(X) + \phi(X)$$

Gradient flow :
$$\begin{cases} \dot{X}(t) = -\nabla_{St} \mathcal{E}(X(t)) \\ X(0) = X_0 \sim \mathcal{U}(St_N(1,r)) \end{cases}$$ invariant measure on Stiefel.

where
$$\nabla_{St} \mathcal{E}(X(t)) = \nabla \mathcal{E}(X(t)) - \frac{1}{2} X (X^T \nabla \mathcal{E}(X(t)) + X^T \nabla \mathcal{E}(X)),$$

is the orthogonal projection of the Euclidian gradient $\nabla \mathcal{E}(X(t))$ on the tangent space $T_X St(1,r)$. The main focus of this talk will be on online stochastic gradient descent, but let's give a few precisions on gradient flow :

$$\dot{X}(t) = -\nabla_{St} H_0(X) - \nabla_{St} \phi(X),$$

then

$$\dot{m}_{ij}(t) = -\langle v_i, \nabla_{St} H_0(X)_j \rangle - \langle v_i, \nabla_{St} \phi(X)_j \rangle.$$

Standard approach is to bound the noise by using uniform concentration on the gradient to bound $\sup_{x \in S^{N-1}(1)} \|\nabla_{St} H_0(x)\|_2$ gives the wrong exponent : N^{p-1}. To prove N^{p-2} for full-batch gradient flow, need time dependent bound on $\langle v_i, \nabla_{St} H_0(X)_j \rangle$: adapt "bounding flow" method of [BAGJ 20, 20+] to multi-spike case.

↳ first paper with Gerard & Vanessa ⟶ Langevin & gradient flow.

↳ Interesting links with DMFT, more of interest to probabilists, happy to talk about it more after.

⟶ Focus now on online SGD for rest of the talk.

Online SGD:

Single sample cost function $\mathcal{L}(X(t))$: $\mathcal{E}(X(t))$ with $M = 1$.

$$\begin{cases} X(t+1) = R_{X(t)}\left(-\delta_N \nabla_{St} \mathcal{L}(X(t))\right) \\ X(0) = X_0 \sim \mathcal{U}(St(1,r)) \end{cases}$$

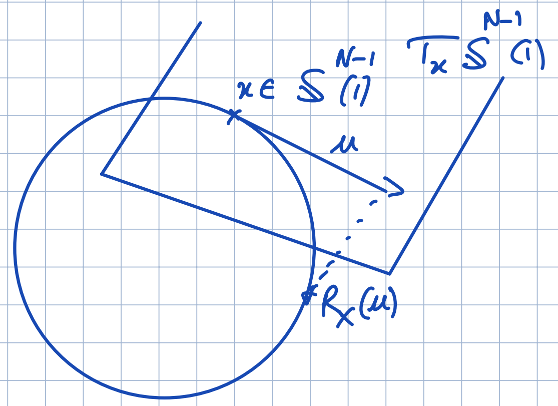
where

$$\nabla_{St} \mathcal{E}(X(t)) = \nabla \mathcal{E}(X(t)) - \frac{1}{2} X\left(X^T \nabla \mathcal{E}(X(t)) + X^T \nabla \mathcal{E}(X)\right),$$

is the orthogonal projection of the Euclidian gradient $\nabla \mathcal{E}(X(t))$ on the tangent space $T_X St(1,r)$, and $R_X(U)$ is the polar retraction, i.e.

$$R_X(U) = (X+U)(I_r + UU^T)^{-1/2} \text{ for any } (X,U) \in (St_N(1,r), T_X St_N(1,r)).$$

<u>Remark</u>: similar to Claire's talk, remember on the sphere



Project Euclidian gradient on tangent space, exit the manifold by taking a finite step, back to the manifold with the retraction. Book by N. Boumal "Optimization on smooth manifolds".

For clarity of exposition, focus on the initial quadrant where $m_{ij}(0) \geq 0$ for all $1 \leq i, j \leq r$. Denote this normalized volume measure $\mathcal{U}_+(St_N(1,r))$.

<u>Remark</u>: can generate a sample from $\mathcal{U}(St_N(1,r))$ with

$$Y = X(X^TX)^{-1/2} \quad \text{and } X \text{ has i.i.d. entries.}$$

In high dimension, can roughly consider the overlap matrix

$$M_0 = V^T X_0 \quad \text{as i.i.d. } \mathcal{N}(0, \tfrac{1}{N}) \text{ random variables.}$$

To state our main result, we first need a definition:

<u>Definition (Greedy maximum selection).</u>

Let $A = (\lambda_i \lambda_j m_{ij}^{p-2}(X_0))_{1 \leq i,j \leq r} \in \mathbb{R}^{r \times r}$. We define the pairs $\{(i_k^*, j_k^*)\}_{k=1}^{r}$, recursively as

$$(i_k^*, j_k^*) = \underset{1 \le i,j \le r-(k-1)}{\operatorname{argmax}} \; [A^{(k-1)}]_{ij},$$

where $A^{(k-1)}$ is obtained by removing the rows $i_1^*, \ldots, i_{k-1}^*$ and the columns $j_1^*, \ldots, j_{k-1}^*$ from A. Greedy maximum selection of A: $(i_1^*, j_1^*) \ldots (i_r^*, j_r^*)$.

For $p \geq 3$, we then have the following result: (will do $p=2$ if time permits).

Theorem: $X_0 \sim \mathcal{U}_+(St_N(r,n))$. If $M \gg \log(N) N^{p-2}$, the online SGD with step size $\delta_N \ll \log(N)^{-1} N^{-\frac{p-1}{2}}$ produces an estimator X_M s.t., for all $k \in [r]$

$$|m_{i_k^* j_k^*}(X_M)| \xrightarrow[N \to \infty]{P} 1 .$$

Remark:

- always recover a permutation
- theorem is asymptotic, but very robust to finite size effects (finite size in paper).
- perfect recovery if the SNRs are well separated.

Sketch of proof:

Output of online SGD at time t:

$$X_t = R_{X_{t-1}}\left(-\delta_N \nabla_{St} \mathcal{L}(X_{t-1}; Y^t)\right)$$

$$= \left(X_{t-1} - \delta_N \nabla_{St} \mathcal{L}(X_{t-1}; Y^t)\right)\left(I_d + \delta_N^2 \nabla_{St} \mathcal{L}(X_{t-1}; Y^t)^T \nabla_{St} \mathcal{L}(X_{t-1}; Y^t)\right)^{-1/2}$$

$$\simeq X_{t-1} - \delta_N \nabla_{St} \mathcal{L}(X_{t-1}; Y^t) + \text{higher order terms in } \delta_N .$$

↳ need to be careful to control these.

Correlations evolve according to

$$m_{ij}(t) \simeq \langle v_i, (X_{t-1})_j \rangle - \delta_N \langle v_i, \nabla_{St} \mathcal{L}(X_{t-1}; y^t) \rangle$$

$$m_{ij}(t) \simeq m_{ij}(0) - \delta_N \sum_{\ell=1}^{t} \langle v_i, (\nabla_{St} \mathcal{L}(X_{\ell-1}; y^\ell)) \rangle$$

$$= \underbrace{m_{ij}(0)}_{\text{init.}} - \underbrace{\delta_N \sum_{\ell=1}^{t} \langle v_i, \nabla_{St} H^\ell (X_{\ell-1})_j \rangle}_{\text{martingale error term}} - \underbrace{\delta_N \sum_{\ell=1}^{t} \langle v_i, (\nabla_{St} \Phi(X_{\ell-1}))_j \rangle}_{\text{signal}}$$

martingale error term of order $\delta_N \left(\sum_{\ell=1}^{t} \mathrm{Var}[\langle v_i, \nabla_{St} H^\ell (X_{\ell-1})_j \rangle] \right)^{1/2}$

⟶ governs the sample complexity by balancing with init + signal (proof method pioneered in Tan & Vershynin '19, BAGJ '21.)

Upon controlling the noise, can focus on population dynamics

Here, pop. dyn. reads:

$$\dot{m}_{ij}(t) = p \lambda_i \lambda_j m_{ij}^{p-1} - \overbrace{\frac{p}{2} \sum_{1 \le k, \ell \le r} \lambda_k m_{kj} m_{k\ell} m_{i\ell} \left(\lambda_j m_{kj}^{p-2} + \lambda_\ell m_{k\ell}^{p-2} \right)}^{\text{orthogonal correction}}.$$

How do we analyze this? In similar fashion to $\mathcal{U}(\mathbb{S}^{d-1}(1))$, one can show that for $X_0 \sim \mathcal{U}(St(1,r))$, we have for all $1 \le i,j \le r$:

$$m_{ij}(0) \simeq \frac{1}{\sqrt{N}} \quad \text{w.h.p.}$$

(mention restriction to $\mathcal{U}_+(St(1,r))$.)

Then, mean initialization, we can write

$$\dot{m}_{ij}(t) \simeq p \lambda_i \lambda_j m_{ij}^{p-1}(t)$$

so that, for $p \geq 3$

$$m_{ij}(t) \simeq m_{ij}(0)\left(1-(p-2)\lambda_i\lambda_j m_{ij}(0)^{p-2}t\right)^{-\frac{1}{p-2}}.$$

We now write

$$m_{ij}(0) = \frac{\gamma_{ij}}{\sqrt{N}} \quad \text{for some } \gamma_{ij} \text{ of order one.}$$

Then, the first hitting time of $\{m_{ij} \geq \varepsilon\}$ is given by

$$T_\varepsilon^{(ij)} \simeq \frac{1-\left(\frac{\gamma_{ij}}{\sqrt{N}}\right)^{p-2}}{\lambda_i\lambda_j(p-2)\gamma_{ij}^{p-2}} N^{\frac{p-2}{2}}.$$

The key observation is to see that, for any i, j, i', j', if $\lambda_i\lambda_j m_{ij}(0)^{p-2} > (1+\delta)\lambda_{i'}\lambda_{j'} m_{i'j'}(0)^{p-2}$ for some δ of order one, then m_{ij} reaches ε before $m_{i'j'}$ can escape its original scale, i.e.

$$m_{i'j'}\left(T_\varepsilon^{(ij)}\right) \leq \frac{C}{\sqrt{N}}.$$

Thus m_{ij} will trigger the correction term on $m_{ij'}, m_{i'j}$ before it can move too much and send it decreasing near zero. Since the γ_{ij} are roughly i.i.d. standard normal, can always find an ordering of the $\lambda_i\lambda_j m_{ij}(0)^{p-2}$ verifying a sufficient separation. Thus, largest $\lambda_i\lambda_j m_{ij}(0)^{p-2}$ will rise, eliminate all those sharing a line and column index, and so on and so forth. Drawing on blackboard. + small fluctuations.

<u>Remark:</u> partitioning $\frac{1}{\sqrt{N}}, \varepsilon, 1-\varepsilon$ insufficient here, need sequence of hitting times

$$T_{ij}^{(m)} = \inf t \geq 0 : m_{ij}(X) \geq \frac{C^m}{\sqrt{N}} \quad \text{for suitable sequence } m(N).$$

⟶ Sharper control on all error terms, show stability of ordering defined by GITs.

⟶ adds a logarithmic factor from strong Markov + union bound.

→ show simulation, explain that there are smaller fluctuations that need to be controlled, remind presence of the noise. Simulations for two directions, then for more.

For $p = 2$:

Theorem:

Assume $X_0 \sim \mathcal{U}_{+}(St_N(1,r))$ and $\lambda_i = \lambda_{i+1}(1+k_i)$ for $k_i > 0$ of order 1. If $M \gg \log(N)^2 N^{\frac{1}{2}(1-\frac{\lambda_r}{\lambda_1})}$, $\delta_N \ll \log(N)^{-1} N^{-1+\frac{\lambda_r}{2\lambda_1}}$, then for all $i \in [r]$:

$$|m_{ii}(X_T)| \xrightarrow[N\to\infty]{P} 1.$$

- reach the global minimizer (mention Brockett cost function and stable manifold).
- sequential elimination harder to show: all m_{ij} escape the scale of $\frac{1}{\sqrt{N}}$.
- more sensitive to finite size effect.
- if all λ's are equal, monotone dynamics on eigenvalues of $G = M^T M$, subspace recovery.

If time, show simulations for $p = 2$ and give intuition on the bounding flow.