Stochastic gradient descent in high dimensions for multi-spiked tensor PCA

Program “Random tensors and related topics”, IHP Paris

VANESSA PICCOLO (ENS de Lyon)
Based on joint work with GÉRARD BEN AROUS (Courant Institute, NYU) and CÉDRIC GERBELOT (Courant Institute, NYU)

October, 2 2024

Multi-spiked tensor PCA problem

Multi-spiked tensor PCA problem

Definition (Spiked tensor model [Johnstone 2001; Richard, Montanari 2014])

Observe a p-tensor $\boldsymbol{Y} \in (\mathbb{R}^N)^{\otimes p}$ of the form

$$\boldsymbol{Y} = \sum_{i=1}^{r} \sqrt{N} \lambda_i \boldsymbol{v}_i^{\otimes p} + \boldsymbol{W}, \tag{1}$$

where

- $p \geq 2$ and r is fixed,
- $\boldsymbol{W} \in (\mathbb{R}^N)^{\otimes p}$ is a p-tensor with i.i.d. Gaussian entries $W_{i_1,\ldots,i_p} \sim \mathcal{N}(0,1)$,
- $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$ are the signal-to-noise ratios (SNRs),
- $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r \in \mathbb{S}^{N-1}$ are unknown, orthogonal signal vectors.

Multi-spiked tensor PCA problem

Definition (Spiked tensor model [Johnstone 2001; Richard, Montanari 2014])

Observe a p-tensor $\boldsymbol{Y} \in (\mathbb{R}^N)^{\otimes p}$ of the form

$$\boldsymbol{Y} = \sum_{i=1}^{r} \sqrt{N} \lambda_i \boldsymbol{v}_i^{\otimes p} + \boldsymbol{W}, \tag{1}$$

where

- $p \geq 2$ and r is fixed,
- $\boldsymbol{W} \in (\mathbb{R}^N)^{\otimes p}$ is a p-tensor with i.i.d. Gaussian entries $W_{i_1,\ldots,i_p} \sim \mathcal{N}(0,1)$,
- $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$ are the signal-to-noise ratios (SNRs),
- $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r \in \mathbb{S}^{N-1}$ are unknown, orthogonal signal vectors.

Goal: Given M i.i.d. samples $(\boldsymbol{Y}^\ell)_{\ell \leq M}$ of the form (1), estimate $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$ (with high probability as $N \to \infty$).

Multi-spiked tensor PCA problem

Definition (Spiked tensor model [Johnstone 2001; Richard, Montanari 2014])

Observe a p-tensor $\boldsymbol{Y} \in (\mathbb{R}^N)^{\otimes p}$ of the form

$$\boldsymbol{Y} = \sum_{i=1}^{r} \sqrt{N} \lambda_i \boldsymbol{v}_i^{\otimes p} + \boldsymbol{W}, \tag{1}$$

where

- $p \geq 2$ and r is fixed,
- $\boldsymbol{W} \in (\mathbb{R}^N)^{\otimes p}$ is a p-tensor with i.i.d. Gaussian entries $W_{i_1,\ldots,i_p} \sim \mathcal{N}(0,1)$,
- $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$ are the signal-to-noise ratios (SNRs),
- $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r \in \mathbb{S}^{N-1}$ are unknown, orthogonal signal vectors.

Goal: Given M i.i.d. samples $(\boldsymbol{Y}^\ell)_{\ell \leq M}$ of the form (1), estimate $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$ (with high probability as $N \to \infty$).

Estimation task: Produce estimators $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r$ attaining

- exact recovery of the spikes: $\boldsymbol{x}_i = (1 - o(1)) \boldsymbol{v}_i$ for all $1 \leq i \leq r$,
- recovery of a permutation of the spikes: there exists a permutation $\sigma \in S_r$ such that $\boldsymbol{x}_i = (1 - o(1)) \boldsymbol{v}_{\sigma(i)}$ for all $1 \leq i \leq r$,

The multi-spiked tensor PCA problem

Statistical procedure: Maximum Likelihood Estimator of $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r]$ is given by a solution of

$$\begin{aligned} \text{minimize} \quad & \mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}) = \sum_{i=1}^{r} \lambda_i \langle \boldsymbol{W}, \boldsymbol{x}_i^{\otimes p} \rangle - \sum_{1 \leq i,j \leq r} \sqrt{N} \lambda_i \lambda_j \langle \boldsymbol{v}_i, \boldsymbol{x}_j \rangle^{p-1} \\ \text{subject to} \quad & \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_r, \end{aligned} \tag{2}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r] \in \mathbb{R}^{N \times r}$. The set $\mathrm{St}(N, r) = \{\boldsymbol{X} \in \mathbb{R}^{N \times r} \colon \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_r\}$ is known as the Stiefel manifold.

The multi-spiked tensor PCA problem

Statistical procedure: Maximum Likelihood Estimator of $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r]$ is given by a solution of

$$\begin{aligned} &\text{minimize} \quad \mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}) = \sum_{i=1}^{r} \lambda_i \langle \boldsymbol{W}, \boldsymbol{x}_i^{\otimes p} \rangle - \sum_{1 \leq i,j \leq r} \sqrt{N} \lambda_i \lambda_j \langle \boldsymbol{v}_i, \boldsymbol{x}_j \rangle^{p-1} \\ &\text{subject to} \quad \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_r, \end{aligned} \tag{2}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r] \in \mathbb{R}^{N \times r}$. The set $\mathrm{St}(N, r) = \{\boldsymbol{X} \in \mathbb{R}^{N \times r} \colon \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_r\}$ is known as the Stiefel manifold.

Algorithmic approach: We need an algorithm for outputting this estimator $\hat{\boldsymbol{X}} = \mathrm{argmin}_{\boldsymbol{X} \in \mathrm{St}(N,r)} \mathcal{L}(\boldsymbol{X}; \boldsymbol{Y})$.

The multi-spiked tensor PCA problem

Statistical procedure: Maximum Likelihood Estimator of $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r]$ is given by a solution of

$$\begin{aligned} &\text{minimize} \quad \mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}) = \sum_{i=1}^{r} \lambda_i \langle \boldsymbol{W}, \boldsymbol{x}_i^{\otimes p} \rangle - \sum_{1 \le i,j \le r} \sqrt{N} \lambda_i \lambda_j \langle \boldsymbol{v}_i, \boldsymbol{x}_j \rangle^{p-1} \\ &\text{subject to} \quad \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_r, \end{aligned} \tag{2}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r] \in \mathbb{R}^{N \times r}$. The set $\mathrm{St}(N, r) = \{\boldsymbol{X} \in \mathbb{R}^{N \times r} : \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_r\}$ is known as the Stiefel manifold.

Algorithmic approach: We need an algorithm for outputting this estimator $\hat{\boldsymbol{X}} = \operatorname{argmin}_{\boldsymbol{X} \in \mathrm{St}(N,r)} \mathcal{L}(\boldsymbol{X}; \boldsymbol{Y})$.

Goal today: Understand thresholds (number of samples / steps needed) for SGD from random initializations to recover.

Online stochastic gradient descent (SGD)

Online SGD algorithm

Online SGD algorithm

Input: i.i.d. samples $(\boldsymbol{Y}^\ell)_{\ell \le M}$, loss function $\mathcal{L}(\boldsymbol{X};\boldsymbol{Y}^\ell)$, initial guess $\boldsymbol{X}_0$, and step size $\delta_N > 0$.
Update:

$$\boldsymbol{X}_t = \mathcal{R}_{\boldsymbol{X}_{t-1}}\left(-\delta_N \nabla_{\mathrm{St}} \mathcal{L}(\boldsymbol{X}_{t-1};\boldsymbol{Y}^t)\right), \tag{3}$$

where $\mathcal{R}_{\boldsymbol{X}} \colon T_{\boldsymbol{X}}\mathrm{St}(N,r) \to \mathrm{St}(N,r)$ denotes a retraction map and

$$T_{\boldsymbol{X}}\mathrm{St}(N,r) = \left\{\boldsymbol{V} \in \mathbb{R}^{N\times r} \colon \boldsymbol{X}^\top \boldsymbol{V} + \boldsymbol{V}^\top \boldsymbol{X} = 0\right\}$$

denotes the tangent space at $\boldsymbol{X} \in \mathrm{St}(N,r)$. Here, we choose the polar retraction defined by

$$\mathcal{R}_{\boldsymbol{X}}(\boldsymbol{U}) = (\boldsymbol{X}+\boldsymbol{U})\left(\boldsymbol{I}_r + \boldsymbol{U}\boldsymbol{U}^\top\right)^{-1/2}.$$

Moreover, for a function $f \colon \mathrm{St}(N,r) \to \mathbb{R}$,

$$\nabla_{\mathrm{St}} f(\boldsymbol{X}) = \nabla f(\boldsymbol{X}) - \frac{1}{2}\boldsymbol{X}(\boldsymbol{X}^\top \nabla f(\boldsymbol{X}) + \nabla f(\boldsymbol{X})^\top \boldsymbol{X}).$$

Output: $\boldsymbol{X}_M$

Main result for $p \geq 3$

Main result for $p \geq 3$

Theorem (Recovery of a permutation of the spikes for $p \geq 3$)

Let $\boldsymbol{X}_0$ be uniformly distributed on $\mathrm{St}(N, r)$. Assume that $M \gg \log(N) N^{p-2}$, and consider the online SGD started from $\boldsymbol{X}_0$ with step size $\delta_N \ll \log(N)^{-1} N^{-\frac{p-1}{2}}$. Then, after M steps, there exists a permutation $\sigma_* \in S_r$ such that for all $k \in [r]$,

$$|\langle \boldsymbol{v}_{\sigma_*(k)}, (\boldsymbol{X}_M)_k \rangle| \to 1 \quad \text{in probability.}$$

Main result for $p \geq 3$

Theorem (Recovery of a permutation of the spikes for $p \geq 3$)

Let $\boldsymbol{X}_0$ be uniformly distributed on $\mathrm{St}(N, r)$. Assume that $M \gg \log(N) N^{p-2}$, and consider the online SGD started from $\boldsymbol{X}_0$ with step size $\delta_N \ll \log(N)^{-1} N^{-\frac{p-1}{2}}$. Then, after M steps, there exists a permutation $\sigma_* \in S_r$ such that for all $k \in [r]$,

$$|\langle \boldsymbol{v}_{\sigma_*(k)}, (\boldsymbol{X}_M)_k \rangle| \to 1 \quad \text{in probability.}$$

- If $\boldsymbol{X} \in \mathbb{R}^{N \times r}$ is a matrix with i.i.d. entries $\mathcal{N}(0,1)$, then $\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1/2}$ is uniformly distributed on $\mathrm{St}(N, r)$ (Chikuse 1994).

Theorem (Recovery of a permutation of the spikes for $p \geq 3$)

Let $\boldsymbol{X}_0$ be uniformly distributed on $\mathrm{St}(N, r)$. Assume that $M \gg \log(N) N^{p-2}$, and consider the online SGD started from $\boldsymbol{X}_0$ with step size $\delta_N \ll \log(N)^{-1} N^{-\frac{p-1}{2}}$. Then, after M steps, there exists a permutation $\sigma_* \in S_r$ such that for all $k \in [r]$,

$$|\langle \boldsymbol{v}_{\sigma_*(k)}, (\boldsymbol{X}_M)_k \rangle| \to 1 \quad \text{in probability.}$$

- If $\boldsymbol{X} \in \mathbb{R}^{N \times r}$ is a matrix with i.i.d. entries $\mathcal{N}(0, 1)$, then $\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1/2}$ is uniformly distributed on $\mathrm{St}(N, r)$ (Chikuse 1994).
- The permutation is determined by $\lambda_i \lambda_j \langle \boldsymbol{v}_i, (\boldsymbol{X}_0)_j \rangle^{p-2}$.

Main result for $p \geq 3$

Theorem (Recovery of a permutation of the spikes for $p \geq 3$)

Let $\boldsymbol{X}_0$ be uniformly distributed on $\mathrm{St}(N, r)$. Assume that $M \gg \log(N) N^{p-2}$, and consider the online SGD started from $\boldsymbol{X}_0$ with step size $\delta_N \ll \log(N)^{-1} N^{-\frac{p-1}{2}}$. Then, after M steps, there exists a permutation $\sigma_* \in S_r$ such that for all $k \in [r]$,

$$|\langle \boldsymbol{v}_{\sigma_*(k)}, (\boldsymbol{X}_M)_k \rangle| \to 1 \quad \text{in probability.}$$

- If $\boldsymbol{X} \in \mathbb{R}^{N \times r}$ is a matrix with i.i.d. entries $\mathcal{N}(0,1)$, then $\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1/2}$ is uniformly distributed on $\mathrm{St}(N, r)$ (Chikuse 1994).
- The permutation is determined by $\lambda_i \lambda_j \langle \boldsymbol{v}_i, (\boldsymbol{X}_0)_j \rangle^{p-2}$.
- If the SNRs $\lambda_1, \ldots, \lambda_r$ are sufficiently separated, then we have exact recovery of the spikes.

Theorem (Recovery of a permutation of the spikes for $p \geq 3$)

Let $\boldsymbol{X}_0$ be uniformly distributed on $\mathrm{St}(N, r)$. Assume that $M \gg \log(N) N^{p-2}$, and consider the online SGD started from $\boldsymbol{X}_0$ with step size $\delta_N \ll \log(N)^{-1} N^{-\frac{p-1}{2}}$. Then, after M steps, there exists a permutation $\sigma_* \in S_r$ such that for all $k \in [r]$,

$$|\langle \boldsymbol{v}_{\sigma_*(k)}, (\boldsymbol{X}_M)_k \rangle| \to 1 \quad \text{in probability.}$$

- If $\boldsymbol{X} \in \mathbb{R}^{N \times r}$ is a matrix with i.i.d. entries $\mathcal{N}(0,1)$, then $\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1/2}$ is uniformly distributed on $\mathrm{St}(N, r)$ (Chikuse 1994).
- The permutation is determined by $\lambda_i \lambda_j \langle \boldsymbol{v}_i, (\boldsymbol{X}_0)_j \rangle^{p-2}$.
- If the SNRs $\lambda_1, \ldots, \lambda_r$ are sufficiently separated, then we have exact recovery of the spikes.
- Regardless of the values of the SNRs, recovery of a permutation of the spikes is always possible, provided a sample complexity of order $\log(N) N^{p-2}$.

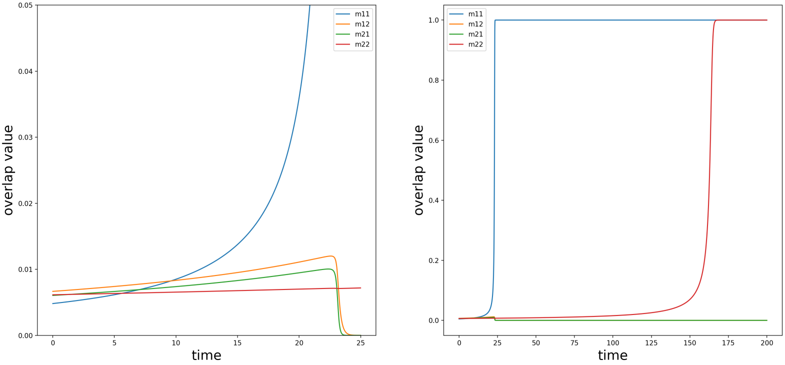

Figure 1: Evolution of the correlations $\{m_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{x}_j \rangle, 1 \leq i, j \leq 2\}$ under the population dynamics for $p = 3$, $\lambda_1 = 3$ and $\lambda_2 = 1$.

Examples

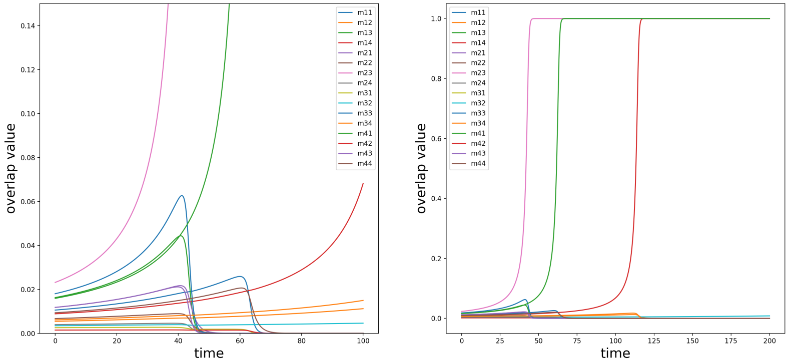

Figure 2: Evolution of the correlations $\{m_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{x}_j \rangle, 1 \leq i, j \leq 4\}$ under the population dynamics for $p = 3$, $\lambda_1 = \cdots = \lambda_4 = 1$.

Examples

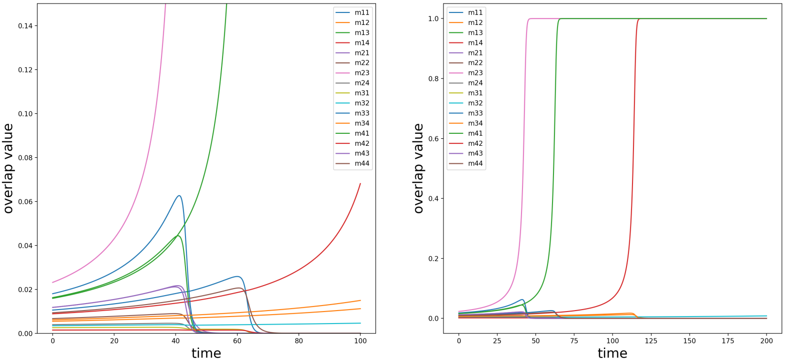

Figure 2: Evolution of the correlations $\{m_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{x}_j \rangle, 1 \leq i, j \leq 4\}$ under the population dynamics for $p = 3$, $\lambda_1 = \cdots = \lambda_4 = 1$.

Sequential elimination phenomenon: The correlations $\{\langle \boldsymbol{v}_{\sigma_*(k)}, \boldsymbol{x}_k \rangle\}_{k=1}^r$ increase one by one, sequentially eliminating those that share a row or column index.

Summary

- The number of samples required for online SGD from random initializations to recover scales as $\log(N)N^{p-2}$;
- For $p \geq 3$, recovery of a permutation of the spikes is always achievable, even when the SNRs are equal;
- The hidden vectors are recovered sequentially in a process we term sequential elimination: once a correlation exceeds a critical threshold, all correlations sharing a row or column index become sufficiently small, allowing the next correlation to grow and become macroscopic.