Learning from Low-Rank Tensor Data: a Random Tensor Theory Perspective

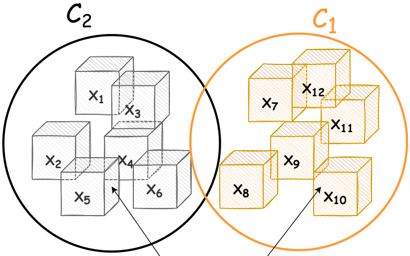
Uncertainty in Artificial Intelligence

MEA. Seddik[1], M. Tiomoko[2], A. Decurninge[2], M. Panov[1], M. Guillaud[3]

[1] Technology Innovation Institute, UAE
[2] Huawei Technologies, France
[3] Inria, France

2nd August 2023

Statistical Data Model



We consider n data points: $(\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3)_{ijk} = x_{1i}x_{2j}x_{3k}$

$$\mathbf{X}_i \in \mathcal{C}_a \quad \Leftrightarrow \quad \mathbf{X}_i = (-1)^a \boldsymbol{\mu}_1 \otimes \cdots \otimes \boldsymbol{\mu}_k + \mathbf{Z}_i \in \mathbb{R}^{p_1 \times \cdots \times p_k}$$

where $[\mathbf{Z}_i]_{i_1 \dots i_k} \sim \mathcal{N}(0, 1)$ i.i.d. and denote $\mathbf{M} = \boldsymbol{\mu}_1 \otimes \cdots \otimes \boldsymbol{\mu}_k$.

- Generalizes the classical model ($k = 1$), i.e. $\boldsymbol{x}_i = (-1)^a \boldsymbol{\mu}_1 + \boldsymbol{z}_i$.
- Even for $k \geq 2$, the standard approach consists in **flattening** the data.
- What is the **optimal** classifier? Theoretical misclassification?

Supervised Learning

Given $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n] \in \mathbb{R}^{p_1 \times \cdots p_k \times n}$ and $\boldsymbol{y} = [y_1, \ldots, y_n] \in \{-1, 1\}^n$
Denote $\boldsymbol{X} = \mathbf{X}_{(k+1)} \in \mathbb{R}^{n \times P}$ with $P = \prod_{i=1}^k p_i$.

We study the *Ridge* classifier

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \gamma \|\boldsymbol{w}\|^2 \quad \Leftrightarrow \quad \boldsymbol{w}^* = \left(\boldsymbol{X}^\top \boldsymbol{X} + \gamma \boldsymbol{I}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

For some $\gamma \gg \|\boldsymbol{X}^\top \boldsymbol{X}\|$ (optimal for the above data model)

$$\boldsymbol{w} = \frac{1}{\sqrt{np}} \boldsymbol{X}^\top \boldsymbol{y}$$

where $p = \sum_{i=1}^k p_i$. In tensor notations, the decision function is

$$f_{\mathrm{R}}(\tilde{\mathbf{X}}_i) = \langle \mathbf{W}, \tilde{\mathbf{X}}_i \rangle \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} 0 \qquad \mathbf{W} \equiv \frac{1}{\sqrt{np}} \mathbf{X} \times_{k+1} \boldsymbol{y}$$

with $\tilde{\mathbf{X}}_i$ a test datum independent of $\mathbf{X}$.

Assumption. $p_i = \mathcal{O}(n)$ and $\|\mathbf{M}\| = \mathcal{O}(1)$.

Data Flattening

Theorem. For $\tilde{\mathbf{X}}_i$ independent of $\mathbf{X}$

$$\frac{1}{\sigma}\left(f_{\mathsf{R}}(\tilde{\mathbf{X}}_i) - m_a\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \quad \Rightarrow \quad \mathcal{E} = Q\left(\frac{|m_a|}{\sigma}\right)$$

where $m_a = (-1)^a \|\mathbf{M}\|^2 \sqrt{\frac{n}{p}}$ and $\sigma = \sqrt{\frac{n}{p}\|\mathbf{M}\|^2 + \frac{P}{p}}$.

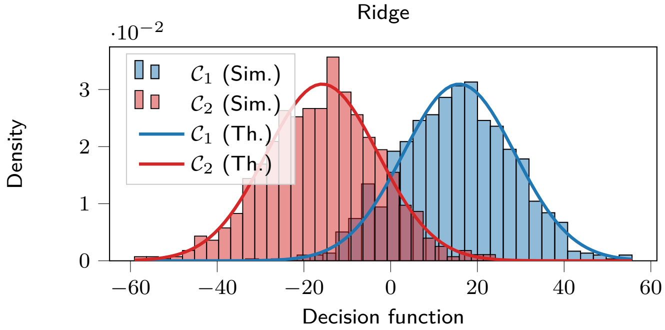


Figure: $n = 200$, shape $(15, 30, 20)$ and $\|\mathbf{M}\| = 3$.

Tensor-based Classification

Given the data model, we have

$$\mathbf{W} = \sqrt{\frac{n}{p}} \bigotimes_{i=1}^{k} \boldsymbol{\mu}_i + \frac{1}{\sqrt{p}} \mathbf{Z}$$

with $\mathbf{Z} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i \mathbf{Z}_i$ (Universality with CLT!).

Tensor-Ridge classifier is defined as

$$f_{\mathsf{TR}}(\tilde{\mathbf{X}}_i) = \left\langle \lambda^* \bigotimes_{i=1}^{k} \boldsymbol{u}_i^*, \tilde{\mathbf{X}}_i \right\rangle \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\lessgtr}} 0$$

where (best rank-one approximation of $\mathbf{W}$)

$$\left(\lambda^*, \{\boldsymbol{u}_i^*\}_{i=1}^k\right) = \underset{\lambda \in \mathbb{R}^+, \boldsymbol{u}_i \in \mathbb{S}^{p_i - 1}}{\arg\min} \left\| \mathbf{W} - \lambda \bigotimes_{i=1}^{k} \boldsymbol{u}_i \right\|_{\mathsf{F}}^2$$

Remark. The above MLE is NP-hard but feasible if $\|\mathbf{M}\| \geq \mathcal{O}(P^{1/4}/p^{1/2})$.

Random Tensor Theory

- Spiked Tensor Model:

$$\mathbf{T} = \beta \bigotimes_{i=1}^{k} \boldsymbol{x}_i + \frac{1}{\sqrt{p}} \mathbf{Z} \quad \rightarrow \quad (\lambda^*, \boldsymbol{u}_i^*) = \operatorname*{arg\,min}_{\lambda>0, \|\boldsymbol{u}_i\|=1} \left\| \mathbf{T} - \lambda \bigotimes_{i=1}^{k} \boldsymbol{u}_i \right\|$$

Theorem (Seddik *et al.* 2023)[a]. As $p_i \to \infty$ with $\frac{p_i}{\sum_{j=1}^k p_j} \to c_i \in (0, \infty)$, there exists $\beta_s > 0$ s.t. for all $\beta > \beta_s$:

$$\lambda^* \xrightarrow{\text{a.s.}} \bar{\lambda}, \quad |\langle \boldsymbol{x}_i, \boldsymbol{u}_i^* \rangle| \xrightarrow{\text{a.s.}} q_i(\bar{\lambda})$$

where $\bar{\lambda}$ satisfies $f(\bar{\lambda}, \beta) = 0$ with

$$\begin{cases} f(z, \beta) = z + g(z) - \beta \prod_{i=1}^k q_i(z), \quad q_i(z) = \sqrt{1 - \frac{g_i^2(z)}{c_i}} \\ g(z) = \sum_{i=1}^k g_i(z), \quad g_i^2(z) - (g(z) + z) g_i(z) - c_i = 0 \end{cases}$$

[a] MEA.Seddik, R.Couillet, M.Guillaud, "**When Random Tensors meet Random Matrices**", Annals of Applied Probability, 2023 (arXiv:2112.12348).

Tensor-based Classification

Theorem. For $\tilde{\mathbf{X}}_i$ independent of $\mathbf{X}$

$$\frac{1}{\sigma}\left(f_{\mathsf{TR}}(\tilde{\mathbf{X}}_i) - m_a\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \quad \Rightarrow \quad \mathcal{E} = Q\left(\frac{|m_a|}{\sigma}\right)$$

where $m_a = (-1)^a \sigma \|\mathbf{M}\| \prod_{j=1}^k q_j(\sigma)$ and $f\left(\sigma, \|\mathbf{M}\|\sqrt{\frac{n}{p}}\right) = 0$.

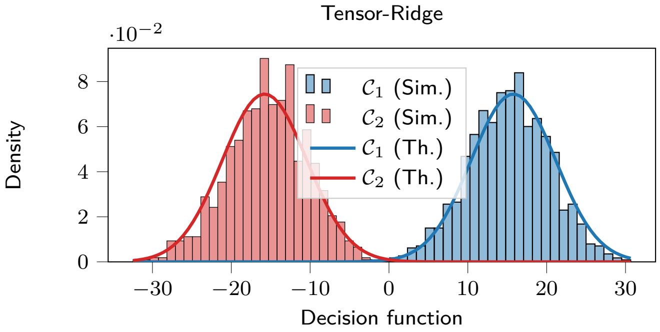
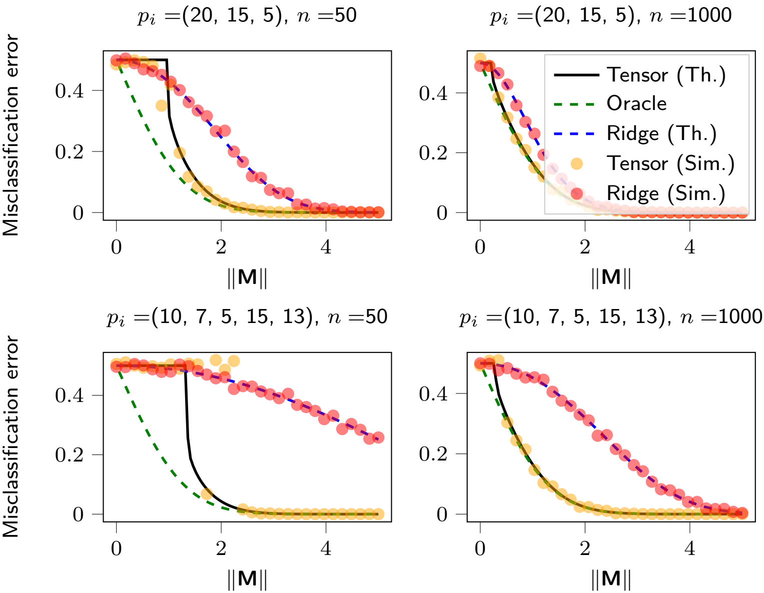
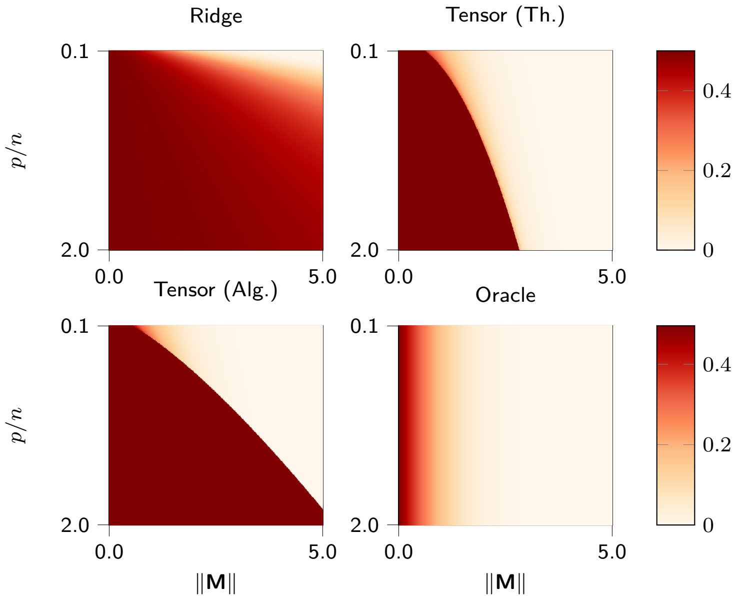


Figure: $n = 200$, shape $(15, 30, 20)$ and $\|\mathbf{M}\| = 3$.

Misclassification Errors



Phase Diagram



Linear and Tensor-based Clustering

- Linear clustering: compute the **left singular vector** of

$$\boldsymbol{X} = \mathbf{X}_{(k+1)} = \boldsymbol{y} \otimes \texttt{flatten}(\mathbf{M}) + \boldsymbol{Z} \in \mathbb{R}^{n \times P} \quad \Rightarrow \quad \hat{\boldsymbol{y}}^{\ell}$$

- Tensor-based clustering: compute the **best rank-one approximation** of

$$\mathbf{X} = \mathbf{M} \otimes \boldsymbol{y} + \mathbf{Z} \in \mathbb{R}^{p_1 \times \cdots \times p_k \times n} \quad \Rightarrow \quad \hat{\boldsymbol{y}}^{\mathcal{T}}$$

Theorem (Linear Clustering). The estimated class for $\mathbf{X}_i$ is given by $\operatorname{sign}(\hat{y}_i^{\ell})$

$$\frac{1}{\sigma_\ell}\left(\sqrt{n}\hat{y}_i^{\ell} - \alpha y_i\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \quad \Rightarrow \quad \mathcal{E} = Q\left(\frac{\alpha_\ell}{\sigma_\ell}\right)$$

where $\alpha_\ell = \kappa\left(\|\mathbf{M}\|\sqrt{\frac{n}{P+n}}, \frac{n}{P+n}\right)^{-1}$ and $\sigma_\ell = \sqrt{1-\alpha_\ell^2}$.

Theorem (Tensor Clustering). The estimated class for $\mathbf{X}_i$ is given by $\operatorname{sign}(\hat{y}_i^{\mathcal{T}})$

$$\frac{1}{\sigma_{\mathcal{T}}}\left(\sqrt{n}\hat{y}_i^{\mathcal{T}} - \alpha y_i\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \quad \Rightarrow \quad \mathcal{E} = Q\left(\frac{\alpha_{\mathcal{T}}}{\sigma_{\mathcal{T}}}\right)$$

where $\alpha_{\mathcal{T}} = q_{k+1}(\lambda^*)$, $\sigma_{\mathcal{T}} = \sqrt{1-\alpha_{\mathcal{T}}^2}$ and $f\left(\lambda^*, \|\mathbf{M}\|\sqrt{\frac{n}{p+n}}\right) = 0$.

Linear and Tensor-based Clustering

- Linear clustering: compute the **left singular vector** of

$$\boldsymbol{X} = \mathbf{X}_{(k+1)} = \boldsymbol{y} \otimes \texttt{flatten}(\mathbf{M}) + \boldsymbol{Z} \in \mathbb{R}^{n \times P} \quad \Rightarrow \quad \hat{\boldsymbol{y}}^{\ell}$$

- Tensor-based clustering: compute a **best rank-one approximation** of

$$\mathbf{X} = \mathbf{M} \otimes \boldsymbol{y} + \mathbf{Z} \in \mathbb{R}^{p_1 \times \cdots \times p_k \times n} \quad \Rightarrow \quad \hat{\boldsymbol{y}}^{\mathcal{T}}$$

Linear (error$= 6.3\%$) Tensor (error$= 0.1\%$)

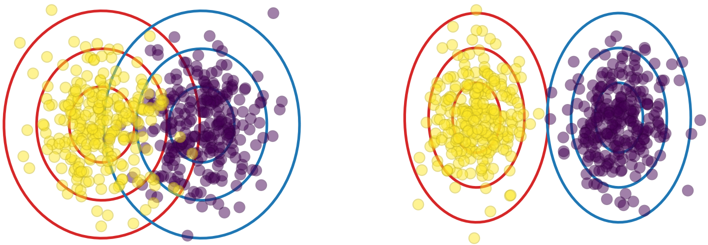
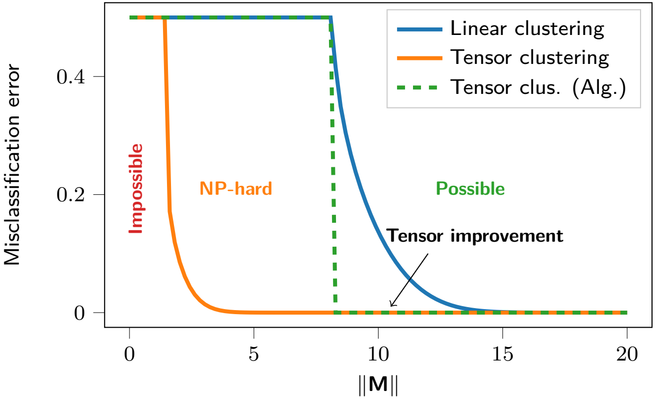


Figure: $n = 200$, shape $(15, 30, 20)$ and $\|\mathbf{M}\| = 3$.

Theoretical Performances



- Possible clustering if $\|\mathbf{M}\| \geq \mathcal{O}\left((P \times n)^{1/4}/(p+n)^{1/2}\right)$.
- **Optimal** clustering with the tensor approach.
- What about the **NP-hard** region?

Thank you for your attention!