Learning from Low Rank Tensor Data: A Random Tensor Theory Perspective (Supplementary Material)

Mohamed El Amine Seddik[1] **Malik Tiomoko**[2] **Alexis Decurninge**[2] **Maxim Panov**[1] **Maxime Guillaud**[3]

[1]Technology Innovation Institute, PO Box: 9639, Masdar City, Abu Dhabi, UAE
[2]Huawei Technologies France, Paris, France
[3]Inria / CITI Laboratory, 6 avenue des Arts, 69621 Villeurbanne, France

Abstract

This supplementary material recalls some tensor operations (Section 1) used throughout the paper and random tensor theory tools presented in Section 2. The main proofs are then presented in Section 3. Finally, some extensions of our results to a more general data model are discussed in Section 4.

1 TENSOR OPERATIONS

We briefly recall in this section some tensor notations and operations that are used throughout the paper.

Inner product and norm: The inner product of two same-sized order k tensors $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{p_1\times\cdots\times p_k}$ is the sum of the products of their entries and is denoted as $\langle \mathbf{X}, \mathbf{Y}\rangle = \sum_{i_1,\dots,i_k} X_{i_1\cdots i_k} Y_{i_1\cdots i_k}$. In particular, the norm $\|\mathbf{X}\|$ of $\mathbf{X} \in \mathbb{R}^{p_1\times\cdots\times p_k}$ is $\|\mathbf{X}\|^2 = \langle \mathbf{X}, \mathbf{X}\rangle$.

Rank-one tensors An order k tensor $\mathbf{X} \in \mathbb{R}^{p_1\times\cdots\times p_k}$ is said to be a *rank-one* tensor if it can be written as the outer product of k vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_k$, i.e., $\mathbf{X} = \bigotimes_{j=1}^k \boldsymbol{a}_j = \boldsymbol{a}_1 \otimes \cdots \otimes \boldsymbol{a}_k$, where the outer product $\bigotimes_{i=1}^k \boldsymbol{a}_i$ is defined such that $\left(\bigotimes_{j=1}^k \boldsymbol{a}_j\right)_{i_1\dots i_k} = \prod_{j=1}^k (\boldsymbol{a}_j)_{i_j}$, i.e., each element of the rank-one tensor is the product of the elements of the corresponding vectors.

Tensor multiplication: The j-mode (matrix) product of a tensor $\mathbf{X} \in \mathbb{R}^{p_1\times\cdots\times p_k}$ with a matrix $\boldsymbol{M} \in \mathbb{R}^{m\times p_j}$ is denoted $\mathbf{X} \times_j \boldsymbol{M}$ and is a tensor of size $p_1 \times \cdots \times p_{j-1} \times m \times p_{j+1} \times \cdots \times p_k$. Element-wise, the j-mode (matrix) product is defined as $(\mathbf{X} \times_j \boldsymbol{M})_{i_1\cdots i_{j-1} k i_{j+1}\cdots i_k} = \sum_{i_j=1}^{p_j} X_{i_1\cdots i_k} M_{k i_j}$. Similarly, the j-mode (vector) product or *contraction* of an order k tensor $\mathbf{X} \in \mathbb{R}^{p_1\times\cdots\times p_k}$ with a vector $\boldsymbol{v} \in \mathbb{R}^{p_j}$ is also denoted as $\mathbf{X} \times_j \boldsymbol{v}$ and results in a tensor of order $k-1$ of dimension $p_1 \times \cdots \times p_{j-1} \times p_{j+1} \times \cdots \times p_k$. Element-wise, the j-mode contraction is defined as $(\mathbf{X} \times_j \boldsymbol{v})_{i_1\cdots i_{j-1} i_{j+1}\cdots i_k} = \sum_{i_j=1}^{p_j} X_{i_1\cdots i_k} v_{i_j}$, which basically consists in computing the inner product of each mode-j *fiber* with the vector $\boldsymbol{v}$.

Tensor Rank and the CANDECOMP/PARAFAC Decomposition (CPD): The CP decomposition [Hitchcock, 1927, Landsberg, 2012] produces a decomposition of a tensor $\mathbf{X} \in \mathbb{R}^{p_1\times\cdots\times p_k}$ into a sum of rank-one tensors, i.e., $\mathbf{X} = \sum_{i=1}^r \bigotimes_{j=1}^k \boldsymbol{a}_j^{(i)}$. The rank of $\mathbf{X}$ denoted $\operatorname{rank}(\mathbf{X})$ is defined as the smallest possible integer r for which $\mathbf{X}$ decomposes as above.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

2 RANDOM TENSOR THEORY

The random tensor theory consists of generalizing classical random matrix theory [Marčenko and Pastur, 1967, Baik et al., 2005] to random tensor models. The first line of research on this topic was proposed by Montanari and Richard [2014] who introduced the concept of tensor PCA. Afterward, many works have focused on the analysis of *symmetric* random tensors [Perry et al., 2020, Lesieur et al., 2017, Handschy, 2019, Jagannath et al., 2020, Goulart et al., 2021]. However, symmetric random tensor models have limited applications in machine learning since real data structures do not necessarily have such symmetric properties. In a very recent work by Seddik et al. [2021], a study of *asymmetric* spiked random tensors have been carried out. It considers an observed k-order tensor $\mathbf{T}$ of the form

$$\mathbf{T} = \beta \bigotimes_{j=1}^{k} \boldsymbol{u}_j + \frac{1}{\sqrt{\sum_{i=1}^{k} p_i}} \mathbf{Z} \in \mathbb{R}^{p_1 \times \cdots \times p_k}, \tag{1}$$

where $\boldsymbol{u}_j \in \mathbb{R}^{p_j}$ for $j \in [k]$ are unitary vectors, $\mathbf{Z}$ is a random tensor with i.i.d. $\mathcal{N}(0,1)$ entries and $\beta > 0$ is a parameter controlling the signal-to-noise ratio (SNR). The study has provided asymptotic evaluation of λ and $\langle \boldsymbol{u}_j, \boldsymbol{v}_j \rangle$ with $\lambda \bigotimes_{j=1}^{k} \boldsymbol{v}_j$ being the best rank-one approximation of $\mathbf{T}$ given by the maximum likelihood estimator (MLE) as

$$\underset{\lambda>0, \{\boldsymbol{v}_j \mid \|\boldsymbol{v}_j\|=1,\, j\in[k]\}}{\arg\min} \left\| \mathbf{T} - \lambda \bigotimes_{j=1}^{k} \boldsymbol{v}_j \right\|_{\mathrm{F}}^2. \tag{2}$$

This study was carried out in the high-dimensional regime, where $p_j \to \infty$ with $\frac{p_j}{\sum_{i=1}^{k} p_i} \to c_j \in [0,1]$. Precisely, Seddik et al. [2021] provided the following results which will be subsequently applied in order to assess the performance of the learning algorithms studied in the present work.

2.1 k-ORDER SPIKED RANDOM TENSORS

Theorem 2.1 (Theorem 8 in [Seddik et al., 2021])**.** *As $p_j \to \infty$ with $\frac{p_j}{\sum_{i=1}^{k} p_i} \to c_j \in [0,1]$, for $k \geq 3$, there exists β_s such that for $\beta > \beta_s$,*

$$\begin{cases} \lambda \xrightarrow{a.s.} \lambda^\infty(\beta), \\ |\langle \boldsymbol{u}_j, \boldsymbol{v}_j \rangle| \xrightarrow{a.s.} q_j(\lambda^\infty(\beta)), \end{cases}$$

where $\lambda^\infty(\beta)$ satisfies[1] *$f(\lambda^\infty(\beta), \beta) = 0$ with $f(z,\beta) = z + g(z) - \beta \prod_{j=1}^{k} q_j(z)$, $q_j(z) = \sqrt{1 - \frac{g_i^2(z)}{c_i}}$, $g_j(z) = \frac{g(z)+z}{2} - \frac{\sqrt{4c_j + (g(z)+z)^2}}{2}$ and $g(z)$ being the unique solution to $g(z) = \sum_{j=1}^{k} g_j(z)$.*

In essence, for an SNR β large enough, Theorem 2.1 predicts a non-zero correlation between the signal components (i.e., the $\boldsymbol{u}_j$'s) and their estimated counterparts (i.e., the $\boldsymbol{v}_j$'s) by the MLE. We refer the reader to [Seddik et al., 2021] for a more detailed discussion.

2.2 CUBIC SPIKED RANDOM TENSORS

In the case of cubic tensors, i.e., $k = 3$ and all the tensor dimensions are equal ($p_1 = p_2 = p_3$), λ^∞ and $q_j(\lambda^\infty)$ in Theorem 2.1 have closed form expressions in terms of β.

Corollary 2.2 (Corollary 3 in [Seddik et al., 2021])**.** *As $p_j \to \infty$, for $\beta > \frac{2\sqrt{3}}{3}$,*

$$\begin{cases} \lambda \xrightarrow{a.s.} \lambda^\infty(\beta) = \sqrt{\frac{\beta^2}{2} + 2 + \frac{\sqrt{3}\sqrt{(3\beta^2-4)^3}}{18\beta}}, \\ |\langle \boldsymbol{u}_j, \boldsymbol{v}_j \rangle| \xrightarrow{a.s.} \bar{q}(\beta), \end{cases}$$

with $\bar{q}(\beta) = \dfrac{\sqrt{9\beta^2 - 12 + \frac{\sqrt{3}\sqrt{(3\beta^2-4)^3}}{\beta}} + \sqrt{9\beta^2 + 36 + \frac{\sqrt{3}\sqrt{(3\beta^2-4)^3}}{\beta}}}{6\sqrt{2}\beta}$.

[1] We will sometimes omit the dependence on β for simplicity.

2.3 SPIKED RANDOM MATRICES

For $k=2$, the model in (1) becomes a so-called *spiked random matrix* which has been extensively studied using random matrix theory [Baik et al., 2005, Benaych-Georges and Nadakuditi, 2011, Capitaine et al., 2009, Péché, 2006, Ben Arous et al., 2021]. Theorem 2.1 covers also such models by not letting all tensor dimensions go to infinity which yields the following corollary.

Corollary 2.3 (Corollary 5 in [Seddik et al., 2021])**.** *As $p_1, p_2 \to \infty$ with $\frac{p_1}{p_1+p_2} \to c \in [0,1]$, for $\beta > \sqrt[4]{c(1-c)}$,*

$$\begin{cases} \lambda \xrightarrow{a.s.} \lambda^\infty(\beta) = \sqrt{\beta^2 + 1 + \frac{c(1-c)}{\beta^2}}, \\ |\langle \boldsymbol{u}_1, \boldsymbol{v}_1 \rangle| \xrightarrow{a.s.} \frac{1}{\kappa(\beta, c)}, \quad |\langle \boldsymbol{u}_2, \boldsymbol{v}_2 \rangle| \xrightarrow{a.s.} \frac{1}{\kappa(\beta, 1-c)}, \end{cases}$$

where $\kappa(\beta, c) = \beta \sqrt{\frac{\beta^2(\beta^2+1) - c(c-1)}{(\beta^4 + c(c-1))(\beta^2+1-c)}}$.

3 MAIN PROOFS

3.1 POOF OF THEOREM 3.1

Recall $\boldsymbol{w} = \mathrm{vec}(\mathbf{W})$, $\boldsymbol{X} = \mathrm{Mat}(\mathbf{X})$, $p = \sum_{j=1}^k p_j$ and $P = \prod_{j=1}^k p_j$, hence $\boldsymbol{w} = \frac{1}{\sqrt{np}} \boldsymbol{X}\boldsymbol{y}$. Denoting $\tilde{\boldsymbol{x}}_i = \mathrm{Mat}(\tilde{\mathbf{X}}_i)$ for some $\tilde{\mathbf{X}}_i \in \mathcal{C}_a$ with $a \in \{1,2\}$ independent of the training data $\mathbf{X}$, the decision function write as $f_{\mathrm{R}}(\tilde{\boldsymbol{x}}_i) = \boldsymbol{w}^\top \tilde{\boldsymbol{x}}_i = \sum_{j=1}^d w_j \tilde{x}_{ij}$. Thus, by Lyapunov's central limit theorem [Billingsley, 2008], the decision function has a Gaussian distribution for large n, we, therefore, need to compute its expectation and variance.

Computation of $\mathbb{E}[f_{\mathbf{R}}(\tilde{\boldsymbol{x}}_i)]$: Let $\boldsymbol{\mu} = \mathrm{vec}(\mathbf{M})$, then $\tilde{\boldsymbol{x}}_i = (-1)^a \boldsymbol{\mu} + \boldsymbol{z}_i$ with $\boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_P)$ and

$$\mathbb{E}[f_{\mathrm{R}}(\tilde{\boldsymbol{x}}_i)] = \frac{1}{\sqrt{np}} \mathbb{E}\left[\boldsymbol{y}^\top \boldsymbol{X}^\top \tilde{\boldsymbol{x}}_i\right] = \frac{1}{\sqrt{np}} \boldsymbol{y}^\top \boldsymbol{y} \boldsymbol{\mu}^\top (-1)^a \boldsymbol{\mu} = (-1)^a \sqrt{\frac{n}{p}} \|\boldsymbol{\mu}\|^2 = (-1)^a \sqrt{\frac{n}{p}} \|\mathbf{M}\|^2.$$

Computation of $\mathbb{E}[f(\boldsymbol{x}_i)^2]$:

$$\mathbb{E}\left[f(\boldsymbol{x}_i)^2\right] = \mathbb{E}\left[\frac{1}{np} \boldsymbol{y}^\top \boldsymbol{X}^\top \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top \boldsymbol{X}\boldsymbol{y}\right] = \mathbb{E}\left[\frac{1}{np} \boldsymbol{y}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{y}\right] + \mathbb{E}\left[\frac{1}{np} \boldsymbol{y}^\top \boldsymbol{X}^\top \boldsymbol{\mu}\boldsymbol{\mu}^\top \boldsymbol{X}\boldsymbol{y}\right] = E_1 + E_2.$$

Since $\boldsymbol{X} = \boldsymbol{\mu}\boldsymbol{y}^\top + \boldsymbol{Z}$ with $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n] = \mathrm{Mat}(\mathbf{Z}) \in \mathbb{R}^{d \times n}$, we have

$$\begin{aligned} E_1 &= \frac{1}{np} \|\boldsymbol{\mu}\|^2 \|\boldsymbol{y}\|^4 + \frac{1}{np} \mathbb{E}\left[\boldsymbol{y}^\top \boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{y}\right] = \frac{n}{p} \|\mathbf{M}\|^2 + \frac{P}{p}, \\ E_2 &= \frac{1}{np} \boldsymbol{y}^\top \boldsymbol{y} \boldsymbol{\mu}^\top \boldsymbol{\mu}\boldsymbol{\mu}^\top \boldsymbol{\mu} \boldsymbol{y}^\top \boldsymbol{y} + \frac{1}{np} \mathbb{E}\left[\boldsymbol{y}^\top \boldsymbol{Z}^\top \boldsymbol{\mu}\boldsymbol{\mu}^\top \boldsymbol{Z}\boldsymbol{y}\right] = \frac{1}{np} \|\boldsymbol{y}\|^4 \|\boldsymbol{\mu}\|^4 + \frac{1}{np} \mathrm{tr}\left(\mathbb{E}\left[\boldsymbol{Z}\boldsymbol{y}\boldsymbol{y}^\top \boldsymbol{Z}^\top\right] \boldsymbol{\mu}\boldsymbol{\mu}^\top\right), \end{aligned}$$

where $\mathbb{E}\left[\boldsymbol{Z}\boldsymbol{y}\boldsymbol{y}^\top \boldsymbol{Z}^\top\right] = \mathbb{E}\left[\left(\sum_{i=1}^n y_i \boldsymbol{z}_i\right)\left(\sum_{i=1}^n y_i \boldsymbol{z}_i^\top\right)\right] = \sum_{i=1}^n y_i^2\, \mathbb{E}\left[\boldsymbol{z}_i \boldsymbol{z}_i^\top\right] = n\boldsymbol{I}_P$. Therefore,

$$\mathbb{E}\left[f_{\mathrm{R}}(\tilde{\boldsymbol{x}}_i)^2\right] = \frac{n}{p} \|\mathbf{M}\|^2 + \frac{P}{p} + \frac{n}{p} \|\mathbf{M}\|^4 + \frac{1}{p} \|\mathbf{M}\|^2,$$

and the term $\frac{1}{p}\|\mathbf{M}\|^2$ vanishes for large values of p under Assumption 2.2. In particular, the variance of $f(\boldsymbol{x}_i)$ is given by $\mathbb{E}\left[f_{\mathrm{R}}(\tilde{\boldsymbol{x}}_i)^2\right] - \mathbb{E}\left[f_{\mathrm{R}}(\tilde{\boldsymbol{x}}_i)\right]^2 = \frac{n}{p}\|\mathbf{M}\|^2 + \frac{P}{p}$ for large values of p.

3.2 POOF OF THEOREM 3.3

Denote $\mathbf{M} = \gamma \bigotimes_{j=1}^k \boldsymbol{u}_j$ where $\boldsymbol{u}_j = \frac{\boldsymbol{\mu}_j}{\|\boldsymbol{\mu}_j\|}$, as such $\|\mathbf{M}\| = \gamma$. Therefore, from the definition of the weight tensor and further denoting $\beta = \|\mathbf{M}\| \sqrt{\frac{n}{p}}$, $\mathbf{W}$ expresses as

$$\mathbf{W} = \beta \bigotimes_{j=1}^k \boldsymbol{u}_j + \frac{1}{\sqrt{p}} \tilde{\mathbf{Z}}. \tag{3}$$

The best rank-one approximation $\lambda \bigotimes_{j=1}^k \boldsymbol{v}_j$ (with the $\boldsymbol{v}_j$'s being unitary vectors) of $\mathbf{W}$ is given by the MLE as

$$\underset{\lambda>0, \{\boldsymbol{v}_j \mid \|\boldsymbol{v}_j\|=1,\, j\in[k]\}}{\arg\min} \left\| \mathbf{W} - \lambda \bigotimes_{j=1}^k \boldsymbol{v}_j \right\|_{\mathrm{F}}^2 .$$

As in Section 3.1, for a new test datum $\tilde{\mathbf{X}}_i = (-1)^a \mathbf{M} + \tilde{\mathbf{Z}}_i$, the decision function $f_{\mathrm{TR}}(\tilde{\mathbf{X}}_i)$ is a Gaussian random variable, the mean of which expresses as follows.

$$\mathbb{E}\left[f_{\mathrm{TR}}(\tilde{\mathbf{X}}_i)\right] = \mathbb{E}\left[\left\langle \lambda \bigotimes_{j=1}^k \boldsymbol{v}_j, \tilde{\mathbf{X}}_i \right\rangle\right] = \mathbb{E}\left[(-1)^a \|\mathbf{M}\| \lambda \prod_{j=1}^k \langle \boldsymbol{u}_j, \boldsymbol{v}_j \rangle\right] \to (-1)^a \|\mathbf{M}\| \lambda^\infty(\beta) \prod_{j=1}^k q_j(\lambda^\infty(\beta)),$$

by Theorem 2.1. Moreover, the variance of $f_{\mathrm{TR}}(\tilde{\mathbf{X}}_i)$ expresses as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[f_{\mathrm{TR}}(\tilde{\mathbf{X}}_i)\right] &= \mathbb{E}\left[\left\langle \lambda \bigotimes_{j=1}^k \boldsymbol{v}_j, \tilde{\mathbf{Z}}_i \right\rangle^2\right] = \mathbb{E}\left[\lambda^2 \left(\sum_{i_1,\dots,i_k} \prod_{j=1}^k (\boldsymbol{v}_j)_{i_j} (\tilde{\mathbf{Z}}_i)_{i_1,\dots,i_k} \right)^2\right] \\
&= \mathbb{E}\left[\lambda^2 \sum_{i_1,\dots,i_k, i'_1,\dots,i'_k} \prod_{j=1}^k (\boldsymbol{v}_j)_{i_j} (\tilde{\mathbf{Z}}_i)_{i_1,\dots,i_k} \prod_{j=1}^k (\boldsymbol{v}_j)_{i'_j} (\tilde{\mathbf{Z}}_i)_{i'_1,\dots,i'_k}\right] \\
&= \mathbb{E}\left[\lambda^2 \sum_{i_1,\dots,i_k} \prod_{j=1}^k (\boldsymbol{v}_j)_{i_j}^2 (\tilde{\mathbf{Z}}_i)_{i_1,\dots,i_k}^2\right] = \mathbb{E}\left[\lambda^2 \sum_{i_1,\dots,i_k} \prod_{j=1}^k (\boldsymbol{v}_j)_{i_j}^2 \, \mathbb{E}[(\tilde{\mathbf{Z}}_i)_{i_1,\dots,i_k}^2 \mid \mathbf{Z}]\right] = \mathbb{E}[\lambda^2] \to \lambda^\infty(\beta)^2,
\end{aligned}$$

since $\mathbb{E}[(\tilde{\mathbf{Z}}_i)_{i_1,\dots,i_k}^2 \mid \mathbf{Z}] = 1$ and $\sum_{i_1,\dots,i_k} \prod_{j=1}^k (\boldsymbol{v}_j)_{i_j}^2 = \prod_{j=1}^k \|\boldsymbol{v}_j\|^2 = 1$.

3.3 POOF OF THEOREM 3.5

The equivalent random matrix model writes as

$$\tilde{\boldsymbol{X}} = \sqrt{\frac{n}{d+n}} \operatorname{vec}(\mathbf{M}) \bar{\boldsymbol{y}}^\top + \frac{1}{\sqrt{d+n}} \operatorname{Mat}(\mathbf{Z}) \in \mathbb{R}^{d\times n},$$

where $\bar{\boldsymbol{y}} = \boldsymbol{y}/\sqrt{n}$ and the normalization by $\sqrt{P+n}$ is considered for convenience. Let $\hat{\boldsymbol{y}}$ be the right singular vector of $\tilde{\boldsymbol{X}}$ corresponding to its largest singular value. Then evoking Corollary 2.3, the asymptotic alignment under Assumption 2.2 is given as

$$|\langle \hat{\boldsymbol{y}}, \bar{\boldsymbol{y}} \rangle| \xrightarrow{\text{a.s.}} \alpha = \kappa\left(\|\mathbf{M}\| \sqrt{\frac{n}{P+n}}, \frac{n}{P+n} \right)^{-1}.$$

Moreover, $\hat{\boldsymbol{y}}$ decomposes as

$$\hat{\boldsymbol{y}} = \alpha \bar{\boldsymbol{y}} + \sigma \boldsymbol{w},$$

where $\boldsymbol{w} \in \mathbb{R}^n$ is a random vector, orthogonal to $\bar{\boldsymbol{y}}$ and of unit norm. Since $\hat{\boldsymbol{y}}$ is of unit norm, σ satisfies $1 = \alpha^2 + \sigma^2$, as such $\sigma = \sqrt{1-\alpha^2}$. Finally, the Gaussianity of the entries of $\hat{\boldsymbol{y}}$ is obtained thanks to similar arguments as in [Couillet and Benaych-Georges, 2016].

3.4 POOF OF THEOREM 3.6

The equivalent random tensor model writes as

$$\tilde{\mathbf{X}} = \sqrt{\frac{n}{p+n}} \mathbf{M} \otimes \bar{\boldsymbol{y}} + \frac{1}{\sqrt{p+n}} \mathbf{Z} \in \mathbb{R}^{p_1\times\cdots\times p_k\times n},$$

where $\bar{\boldsymbol{y}} = \boldsymbol{y}/\sqrt{n}$. As such $\tilde{\mathbf{X}}$ is a spiked random tensor of order $k+1$. As in Section 3.3, we need to express the asymptotic alignment between $\hat{\boldsymbol{y}}$ and $\bar{\boldsymbol{y}}$ with $\hat{\boldsymbol{y}}$ being the $(k+1)$-th mode component of the rank-one tensor approximation of $\tilde{\mathbf{X}}$, which is straightforwardly obtained thanks to Theorem 2.1, applied to a $(k+1)$-th order tensor of dimensions $p_1 \times \cdots \times p_k \times n$, yielding

$$|\langle \hat{\boldsymbol{y}}, \bar{\boldsymbol{y}} \rangle| \xrightarrow{\text{a.s.}} \alpha = q_{k+1}\left(\lambda^\infty\left(\|\mathbf{M}\|\sqrt{\frac{n}{p+n}}\right)\right),$$

where $q_{k+1}(\cdot)$ and $\lambda^\infty(\cdot)$ are defined in Theorem 2.1.

4 LOW-RANK DATA MODEL WITH ORTHOGONAL COMPONENTS

Our results generalize to a more complex model of the following form. Suppose that the $\mathbf{X}_i$'s are distributed in two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ (of cardinality n_1 and n_2 respectively), such that for $\mathbf{X}_i \in \mathcal{C}_a$ with $a \in 1, 2$,

$$\mathbf{X}_i = \sum_{\ell=1}^{r_a} \bigotimes_{j=1}^{k} \boldsymbol{\mu}_{j,\ell}^{(a)} + \mathbf{Z}_i \in \mathbb{R}^{p_1 \times \cdots \times p_k}, \tag{4}$$

where $\mathbf{Z}_i$ is a random tensor with i.i.d. standard Gaussian entries, $\boldsymbol{\mu}_{j,\ell}^{(a)} \in \mathbb{R}^{p_j}$ are independent from $\mathbf{Z}_i$ such that $\langle \boldsymbol{\mu}_{j,\ell_1}^{(a)}, \boldsymbol{\mu}_{j,\ell_2}^{(a)} \rangle = \delta_{\ell_1 \ell_2}$. That is, the data tensors $\mathbf{X}_i$ have a rank-r_a (with r_a being small) structure with orthogonal components.

4.1 SUPERVISED SETTING

Let us denote $\mathbf{M}_a = \sum_{\ell=1}^{r_a} \bigotimes_{j=1}^{k} \boldsymbol{\mu}_{j,\ell}^{(a)}$. In a supervised setting, it is convenient to center the data by subtracting[2] $\frac{1}{2}(\mathbf{M}_1 + \mathbf{M}_2)$ from each data sample which yields tensors of the form

$$\mathbf{X}_i = (-1)^a (\mathbf{M}_1 - \mathbf{M}_2) + \mathbf{Z}_i, \tag{5}$$

where $\mathbf{M}_1 - \mathbf{M}_2$ is clearly a low-rank tensor (of rank $r_1 + r_2$) with orthogonal components. Stacking all the data samples $\mathbf{X}_i$ in a data tensor $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_k \times n}$, the ∞-Ridge classifier has weights tensor of the form

$$\mathbf{W} = \frac{1}{\sqrt{np}} \mathbf{X} \times_{k+1} \boldsymbol{y} = \sqrt{\frac{n}{p}} \mathbf{M} + \frac{1}{\sqrt{p}} \tilde{\mathbf{Z}}, \tag{6}$$

where $\tilde{\mathbf{Z}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i \mathbf{Z}_i$ and $\mathbf{M} = \mathbf{M}_1 - \mathbf{M}_2 = \sum_{\ell=1}^{r_1+r_2} \bigotimes_{j=1}^{k} \boldsymbol{\mu}_{j,\ell}$ is a rank-$(r_1 + r_2)$ tensor. Therefore, the Tensor-Ridge classifier for this case relies on a low-rank approximation of $\mathbf{W}$ of rank $r_1 + r_2$ which might be performed using tensor power iteration with deflation procedure. We, therefore, have the following theorem characterizing the performance of the Tensor-Ridge classifier in this case.

Theorem 4.1 (Performance of the Tensor-Ridge classifier for data model in (5))**.** *Under Assumption 2.2, for $\tilde{\mathbf{X}}_i \in \mathcal{C}_a$ with $a \in \{1, 2\}$ independent from the training set $\mathbf{X}$,*

$$\frac{1}{\sqrt{\sum_{\ell=1}^{r_1+r_2} \sigma_\ell^2}} \left(f_{TR}(\tilde{\mathbf{X}}_i) - m_a\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1),$$

where $m_a = (-1)^a \sum_{\ell=1}^{r_1+r_2} \sigma_\ell \mu_\ell \prod_{j=1}^{k} q_j(\sigma_\ell, \mu_\ell \sqrt{\frac{n}{p}})$ where $\mu_\ell = \| \bigotimes_{j=1}^{k} \boldsymbol{\mu}_{j,\ell} \|$ and σ_ℓ satisfies $f(\sigma_\ell, \mu_\ell \sqrt{\frac{n}{p}}) = 0$. q_j and f are defined in Theorem 2.1. Furthermore, the misclassification error verifies with probability one $\mathbb{P}\left((-1)^a g_{CP}(\tilde{\mathbf{X}}_i) < 0 \mid \tilde{\mathbf{X}}_i \in \mathcal{C}_a\right) - Q\left(\frac{|m_a|}{\sqrt{\sum_{\ell=1}^{r_1+r_2} \sigma_\ell^2}}\right) \to 0.$

Proof. The proof strategy is the same as for theorem 3.3. ∎

[2] In real scenarios one would first estimate the $\mathbf{M}_a$'s with their empirical estimates through tensor decomposition.

4.2 UNSUPERVISED SETTING

The generalization to the unsupervised setting is more challenging since the data tensor $\mathbf{X}$ for the model in (4) does not follow a CP decomposition but rather a block-term decomposition [Rontogiannis et al., 2021] which is more challenging to analyze theoretically and is therefore left for a future investigation.

References

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

Gérard Ben Arous, Daniel Zhengyu Huang, and Jiaoyang Huang. Long random matrices and tensor unfolding. *arXiv preprint arXiv:2110.10210*, 2021.

Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

Mireille Capitaine, Catherine Donati-Martin, and Delphine Féral. The largest eigenvalues of finite rank deformation of large wigner matrices: convergence and nonuniversality of the fluctuations. *The Annals of Probability*, 37(1):1–47, 2009.

Romain Couillet and Florent Benaych-Georges. Kernel spectral clustering of large dimensional data. *Electronic journal of statistics*, 10(1):1393–1454, 2016.

José Henrique Goulart, Romain Couillet, and Pierre Comon. A random matrix perspective on random tensors. *stat*, 1050:2, 2021.

Madeline Curtis Handschy. *Phase Transition in Random Tensors with Multiple Spikes*. PhD thesis, University of Minnesota, 2019.

Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6 (1-4):164–189, 1927.

Aukosh Jagannath, Patrick Lopatto, and Leo Miolane. Statistical thresholds for tensor PCA. *The Annals of Applied Probability*, 30(4):1910–1933, 2020.

Joseph M Landsberg. Tensors: geometry and applications. *Representation theory*, 381(402):3, 2012.

Thibault Lesieur, Léo Miolane, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová. Statistical and computational phase transitions in spiked tensor estimation. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, pages 511–515, 2017.

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

Andrea Montanari and Emile Richard. A statistical model for tensor PCA. *arXiv preprint arXiv:1411.1076*, 2014.

Sandrine Péché. The largest eigenvalue of small rank perturbations of hermitian random matrices. *Probability Theory and Related Fields*, 134(1):127–173, 2006.

Amelia Perry, Alexander S Wein, and Afonso S Bandeira. Statistical limits of spiked tensor models. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 56, pages 230–264. Institut Henri Poincaré, 2020.

Athanasios A Rontogiannis, Eleftherios Kofidis, and Paris V Giampouras. Block-term tensor decomposition: Model selection and computation. *IEEE Journal of Selected Topics in Signal Processing*, 15(3):464–475, 2021.

Mohamed El Amine Seddik, Maxime Guillaud, and Romain Couillet. When random tensors meet random matrices. *arXiv preprint arXiv:2112.12348*, 2021.