On maximum-likelihood estimation in the all-or-nothing regime

Luca Corinzia, and Paolo Penna
Department of Computer Science
ETH Zürich, Switzerland
Email: {luca.corinzia,paolo.penna}@inf.ethz.ch

Wojciech Szpankowski
Department of Computer Science
Purdue University, USA
Email: szpan@purdue.edu

Joachim M. Buhmann
Department of Computer Science
ETH Zürich, Switzerland
Email: jbuhmann@inf.ethz.ch

***Abstract*—We study the problem of estimating a rank-1 additive deformation of a Gaussian tensor according to the *maximum-likelihood estimator* (MLE). The analysis is carried out in the sparse setting, where the underlying signal has a support that scales sublinearly with the total number of dimensions. We show that for Bernoulli distributed signals, the MLE undergoes an *all-or-nothing* (AoN) phase transition, already established for the minimum mean-square-error estimator (MMSE) in the same problem. The result follows from two main technical points: (i) the connection established between the MLE and the MMSE, using the first and second-moment methods in the constrained signal space, (ii) a recovery regime for the MMSE stricter than the simple error vanishing characterization given in the standard AoN, that is here proved as a general result.**

I. Introduction

A fundamental question in information theory, statistics, and machine learning is to establish the *computational* limits of estimation problems and determine the *statistical* limit as the inviolable benchmark for the same problem. The common picture that arises in many problems is given by the presence of *phase transitions* where the behaviour of the optimal estimators changes abruptly with the variation of the parameter of the problem. Typically, at least two phase transitions are present: the statistical phase transition that establishes the limit of *any* estimator, and the computational phase transition at higher signal strength that establishes the limit of *tractable* estimators, despite the two can coincide. In the sparse setting of many estimation problems, a rather different picture emerges, as the statistical and computational phase transitions are situated at different scales of the parameters space and such gap diverges in the limit of vanishing sparsity. Moreover, the statistical phase transition is characterized by the so-called *all-or-nothing* phenomenon (AoN): below a critical signal strength, the recovery of the planted signal is impossible, above the threshold is possible and with vanishing error. Although the AoN is conjectured to extend to the behaviour of any optimal estimator, the analysis has been so far focused on the minimum mean-square-error (MMSE) estimation, that is typically a *bulk* estimator and hence can be too coarse for specific applications in the sparse setting. Hence, it would be desirable to extend the analysis of this phenomenon to other estimators, like the maximum-likelihood estimator (MLE), which recently received attention for showing optimal performance in retrieving the planted signal [1] in a tensor-PCA model.

A. Contribution

In this work, we provide new results on the AoN phenomenon in the sparse estimation setting, with the following main contributions:

- In Theorem 1, we generalize the AoN phenomenon proved in [2] for the additive Gaussian noise model and the MMSE to arbitrarily *asymptotics* in the *recovery* regime. The proof follows a conditional second-moment method argument [3] and extends the proof given in [2] with a careful control of the asymptotics of the bounds. This result is of independent interest as more stringent conditions than the simple error vanishing characterization are needed in specific applications.
- As an application of the first result, we study the *maximum-likelihood estimator* in the sparse tensor-PCA problem and show in Theorem 2 that also this estimator undergoes a weak AoN transition.
- The proof of the latter results is of independent interest, as it exploits the relations between different estimators with first and second-moment methods, crucially introducing the analysis of estimators *constrained* in the signal hypothesis space. As a side result, the weak AoN phenomenon is proved for the constrained MMSE in Theorem 3.

B. Related work

The problem of high dimensional statistical estimation that we study here has received much attention recently, with considerable progress obtained in the last years in understanding planted matrix and tensor models. Early works on statistical and computational limits of estimation focused on dense problems where the signal effective dimensionality scales linearly with the problem's dimensionality. Examples include: (i) compressed sensing [4] and matrix-PCA [5], [6] where the approximated message passing (AMP) algorithms are introduced and demonstrated to match the statistical phase transition; (ii) the tensor-PCA extension [7] where the statistical and computational transitions are currently separated by a gap, considering a wide range of algorithms, i.e. spectral [8], [9], AMP [10], Sum-of-Square [11] and gradient descent [12], [13]. Many of these works focused on the mean-square-error and (high dimensional, i.e., matrix or tensorial) posterior average estimator. Nonetheless, recently more attention has been given to other estimators, e.g., in [1] where the vectorial maximum-likelihood estimator has been shown to reach optimal correlation with the planted signal in the tensor-PCA model. See [14] for a thorough review in the field.

In the sparse regime in which the hidden signal's dimensionality is sublinear to the problem's dimensionality, the AoN phenomenon emerges. This phenomenon have been shown recently to hold in a wide range of problems, i.e., for sparse linear regression [15], sparse matrix-PCA [16] and sparse tensor-PCA [2] according to the mean-square-error loss. However, to the best of our knowledge, only a few other works studied how the same AoN phenomenon extends to other estimators. Examples include [17], [18], where non-matching upper and lower bounds are provided for the transition of the vectorial-MLE in the sparse planted hypergraph problem (equivalent to sparse tensor-PCA up to a reparameterization of the dimensionality of the problem), and [19] where the AoN is proved in the sparse linear regression model for the vectorial-MLE estimator.

II. Setting

We study the estimation problem with observations given by the Gaussian additive model

$$\mathbf{Y} = \sqrt{\lambda}\mathbf{X} + \mathbf{Z} \tag{1}$$

where the signal to be estimated $\mathbf{X} \in \mathbb{R}^n$ is corrupted by Gaussian noise $\mathbf{Z}$, with the collection $\{Z_i\}_{i=1}^n \overset{iid}{\sim} \mathbb{P}_z = \mathcal{N}(0,1)$. We further assume that the prior distribution of $\mathbf{X}$, denoted by $\mathbb{P}_n$, is uniform and discrete with support $supp(\mathbb{P}_n) \subset \mathcal{S}_{n-1}$, where $\mathcal{S}_{n-1}$ is the unit sphere in $\mathbb{R}^n$. We denote by $M_n = |supp(\mathbb{P}_n)|$ the cardinality of the support of $\mathbf{X}$ and by $\mathbb{Q}_{y|x}(\mathbf{Y}|\mathbf{X}) = \mathcal{N}(\mathbf{Y}|\sqrt{\lambda}\mathbf{X}, \mathbf{1}_{n\times n})$ the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$, where $\mathbf{1}_{n\times n}$ is the identity matrix. We hence define

$$\mathbb{Q}_{\lambda,n}(\mathbf{Y}) = \mathbb{E}_{\mathbb{P}_n}[\mathbb{Q}_{y|x}(\mathbf{Y}|\mathbf{X})]$$

as the distribution over the observations $\mathbf{Y}$, highlighting the respective signal-to-noise-ratio (snr) λ and the problem dimension n for convenience.

Definition 1. *Let us denote as $\bar{\mathcal{S}}_{n-1} = \{\mathbf{X} \in \mathbb{R}^n \colon \|\mathbf{X}\| \leq 1\}$ the unit ball. For any set A we use the short notation $\min_{\hat{\mathbf{X}}(\mathbf{Y})\in A}$ for $\min_{\hat{\mathbf{X}}\colon \mathbf{Y}\to\hat{\mathbf{X}}(\mathbf{Y})\in A}$ and analogously for other operators. For a generic bounded loss function* L *we define the respective optimal estimator as*

$$\mathbf{X}_{\mathrm{L}}(\mathbf{Y}) = \underset{\hat{\mathbf{X}}(\mathbf{Y})\in\bar{\mathcal{S}}_{n-1}}{\operatorname{argmin}} \ \mathbb{E}\left[\mathrm{L}(\mathbf{X}, \hat{\mathbf{X}}(\mathbf{Y}))\right].$$

The minimum loss achieved by such estimator is the quantity

$$\mathrm{L}_n(\lambda) = \mathbb{E}\left[\mathrm{L}(\mathbf{X}, \mathbf{X}_{\mathrm{L}}(\mathbf{Y}))\right].$$

We further define the best estimator that solves the optimization problem constrained to set A with $supp(\mathbb{P}_n) \subset A \subset \bar{\mathcal{S}}_{n-1}$ as

$$\mathbf{X}_{\text{C-L}}(\mathbf{Y}) = \underset{\hat{\mathbf{X}}(\mathbf{Y})\in A}{\operatorname{argmin}} \, \mathbb{E}\left[\mathrm{L}(\mathbf{X}, \hat{\mathbf{X}}(\mathbf{Y}))\right],$$

and the respective minimum loss achieved as

$$\text{C-L}_n(\lambda) = \mathbb{E}\left[\mathrm{L}(\mathbf{X}, \mathbf{X}_{\text{C-L}}(\mathbf{Y}))\right].$$

We can easily observe that since $A \subset \bar{\mathcal{S}}_{n-1}$ then $\text{C-L}_n(\lambda) \geq \mathrm{L}_n(\lambda)$ for any loss function L.

Definition 2. *Denote by* L *a bounded loss function $\mathrm{L}\colon \bar{\mathcal{S}}_{n-1}^{\otimes 2} \to \mathbb{R}_+$ with $\mathrm{L}(\mathbf{X},\mathbf{X}) = 0$. Denote by $c > 0$ the optimal error obtained by the estimator* independent *on the observations $\mathbf{Y}$ as $c = \lim_{n\to\infty} \mathrm{L}_n(0)$. The estimation problem with observations given by Equation (1) with prior $\mathbb{P}_n$ satisfies the* all-or-nothing *phenomenon (AoN) with respect to the loss* L*, with recovery asymptotics at least $\tau_n \in o(1)$ and critical snr λ_n if*

$$\lim_{n\to\infty} \mathrm{L}_n(\beta\lambda_n) = \begin{cases} c & \beta < 1 \\ 0 & \beta > 1 \end{cases} \tag{2}$$

and moreover

$$\mathrm{L}_n(\beta\lambda_n) \in o(\tau_n)$$

for $\beta > 1$, where β is a constant independent on n.

Intuitively, in the AoN phenomenon the estimation is impossible for a normalized snr smaller then a critical value $\beta_c = 1$, as the loss converges to the error achieved by an uninformative estimator, equivalent to the loss given by $\beta = 0$, while for higher snr the estimation is almost perfect, with error smaller then a given $\tau_n \to 0$. Note the difference of this definition from the one given in [2] in the recovery regime $\beta > 1$. In latter case the simpler condition $\mathrm{L}_n(\beta\lambda_n) \to 0$ is given such that the asymptotics of the loss in the recovery regime in no further characterize. In the following we denote by τ_n a vanishing sequence such that $\tau_n \in o(1)$.

III. Generalized all-or-nothing phenomenon

In this section, we consider the mean-square-error (MSE), $\mathbb{E}[\|\mathbf{X}-\hat{\mathbf{X}}\|^2]$ where the expectation is taken with respect to $\mathbb{P}_n$ and $\mathbb{P}_z$, and $\hat{\mathbf{X}} \coloneqq \hat{\mathbf{X}}(\mathbf{Y})$ is an estimator of the signal given the observation $\mathbf{Y}$. The MSE is minimized by the posterior average $\mathbf{X}_{l_2} = \mathbb{E}[\mathbf{X}|\mathbf{Y}]$, with the average taken respect to the posterior $\mathbb{P}_n(\mathbf{X}|\mathbf{Y})$ (see [20] and Lemma 12). The minimum mean-square-error (MMSE) is then $\mathrm{MMSE}_n(\lambda) \coloneqq \mathbb{E}[\|\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}]\|^2]$. This error is the minimum square-error achievable by any estimator that has access to the observations $\mathbf{Y}$. In the following, we denote by $D(p||q)$ the KL divergence between the distribution p and q, and by $\pm o(\tau_n)$ a sequence $f_n \in o(\tau_n)$ that is respectively non-negative and non-positive. We further use the scaling $\lambda_n = 2\log M_n$. A sufficient condition for having the AoN phenomenon is given by the property of the overlap rate function defined here.

Definition 3. *For any $t \in [-1, 1]$ define the overlap rate function between two independent instances of the signal $\mathbf{X}$ and $\mathbf{X}'$ as*

$$r_n(t) = -\frac{1}{\log M_n} \log \mathbb{P}_n^{\otimes 2}[\langle \mathbf{X}, \mathbf{X}' \rangle \geq t].$$

where $\langle \mathbf{X}, \mathbf{X}' \rangle$ is the scalar product of two vectors.

Intuitively, the rate function describes the rate of the exponential decay of the overlap $\langle \mathbf{X}, \mathbf{X}' \rangle$. The following theorem shows that a simple lower bound on the overlap rate function is sufficient to establish the AoN with recovery asymptotics τ_n if the latter is not too small.

Theorem 1. *For any $\epsilon > 0$ constant, if $\lambda_n^{-1/2+\epsilon} \in o(\tau_n)$ and the overlap rate function $r_n(t)$ satisfies*

$$r_n(t) \geq \frac{2t}{1+t} - o(\tau_n),$$

then the probability $\mathbb{P}_n$ of the problem defined in Equation (1) satisfies the AoN in Definition 2 with recovery asymptotics at least τ_n according the mean-square-error.

Proof. The first part of Definition 2 related to Equation (2) follows easily noting that the assumption given here is stricter than the one given in [2]. We hence have to prove only that the stronger asymptotics holds in the recovery regime. The proof follows the steps of the proof in [2] and mainly uses the widely known I-MMSE relation that relates the MMSE to the mutual information $I(\mathbf{X}, \mathbf{Y})$ and hence to the $D(\mathbb{Q}_{\lambda,n}||\mathbb{Q}_{0,n})$. It then uses the conditional second-moment method to bound such divergence. We first have the following bound that connects the KL divergence and the MMSE, that is proved in the appendix using the I-MMSE relation.

Lemma 1. *If $\frac{1}{\lambda_n} D(\mathbb{Q}_{\lambda_n,n}||\mathbb{Q}_{0,n}) \in o(\tau_n)$ then for any $\beta > 1$ constant,*

$$\mathrm{MMSE}_n(\beta\lambda_n) \in o(\tau_n).$$

We can now bound the KL divergence $\frac{1}{\lambda_n} D(\mathbb{Q}_{\lambda_n,n}||\mathbb{Q}_{0,n})$ conditioning on a high probability event defined as follows.

Definition 4. *Let $\mathbb{Q}_{xy} = \mathbb{P}_n \otimes \mathbb{Q}_{y|x}$ the joint probability distribution of the vectors $(\mathbf{X}, \mathbf{Y})$ of problem defined in Equation (1) with snr λ_n. A series of events $\Omega_n \subset supp(\mathbb{P}_n) \otimes supp(\mathbb{Q}_{\lambda_n,n})$ occurs with high probability $1 - o(\tau_n)$ uniformly over $\mathbf{X}$ if*

$$\mathbb{Q}_{xy}[\Omega_n|\mathbf{X}] = 1 - o(\tau_n) \tag{3}$$

for any $\mathbf{X} \in supp(\mathbb{P}_n)$.

Let us define as $\tilde{\mathbb{Q}}_{\lambda,n}$ the probability distribution of $\mathbf{Y}$ condition on a high probability event Ω_n. Then the following bound holds.

Lemma 2. *If Ω_n is an event that occurs with uniform high probability $1 - o(\tau_n)$ then*

$$\frac{1}{\lambda_n} D(\mathbb{Q}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) \leq \frac{1}{\lambda_n} D(\tilde{\mathbb{Q}}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) + o(\tau_n)$$

To complete the proof, we need a claim that relates the KL given by the conditional distribution $\tilde{\mathbb{Q}}_{\lambda,n}$ of $\mathbf{Y}$ given a high probability event, to the overlap rate function of the problem. We first introduce the high probability events as follows.

Lemma 3. *Given a sequence C_n with $\frac{1}{C_n} e^{-C_n^2/2} \in o(\tau_n)$, the event*

$$\Omega_n = \{(\mathbf{X}, \mathbf{Y})\colon |\langle \mathbf{X}, \mathbf{Y} \rangle - \sqrt{\lambda_n}| \leq C_n\}$$

satisfies Definition 4.

We can hence prove the following.

Lemma 4. *For any $\epsilon > 0$ constant, if $\lambda_n^{-1/2+\epsilon} \in o(\tau_n)$, conditioning on the event $\Omega_n = \{(\mathbf{X}, \mathbf{Y})\colon |\langle \mathbf{X}, \mathbf{Y} \rangle - \sqrt{\lambda_n}| \leq \sqrt{\log \lambda_n}\}$, the following bound holds:*

$$\frac{1}{\lambda_n} D(\tilde{\mathbb{Q}}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) \leq \sup_{t \in [0,1]} \left[\frac{t}{t+1} - \frac{r_n(t)}{2}\right] + o(\tau_n).$$

We can then finally bound the KL divergence conditioning on the events Ω_n defined in Lemma 4 as

$$\begin{aligned}
\frac{1}{\lambda_n} D(\mathbb{Q}_{\lambda_n,n}||\mathbb{Q}_{0,n}) &\leq \frac{1}{\lambda_n} D(\tilde{\mathbb{Q}}_{\lambda_n,n}||\mathbb{Q}_{0,n}) + o(\tau_n) \\
&\leq \sup_{t \in [0,1]} \left[\frac{t}{t+1} - \frac{r_n(t)}{2}\right] + o(\tau_n) \\
&\in o(\tau_n)
\end{aligned}$$

where the first inequality comes from Lemma 2, Lemma 3 and the assumption that $\lambda_n^{-1/2+\epsilon} \in o(\tau_n)$, the second inequality comes from Lemma 4 and the final inclusion is due to the assumption of the theorem on the rate function $r_n(t)$. The missing proofs of the lemmas are postponed to the appendix. □

IV. Maximum-likelihood estimation

We here study the MLE, showing that a weaker AoN phenomenon extends to the behaviour of this estimator in the case of the sparse tensor-PCA model.

Definition 5. *The MLE for the generic model in Equation (1) is the estimator that maximizes the likelihood as*

$$\mathbf{X}_{\mathrm{MLE}} = \underset{\hat{\mathbf{X}}(\mathbf{Y}) \in \bar{\mathcal{S}}_{n-1}}{\operatorname{argmax}} \ \mathbb{Q}_{y|x}(\mathbf{Y}|\mathbf{X}(\mathbf{Y}))$$

The following characterization of the MLE follows easily from the definition.

Lemma 5. *The MLE minimizes the probability of error*

$$\mathrm{EP}_n(\hat{\mathbf{X}}) := \mathbb{P}_n \mathbb{Q}_{y|x}[\hat{\mathbf{X}}(\mathbf{Y}) \neq \mathbf{X}] = \mathbb{E}\left[\mathbb{1}_{\{\|\hat{\mathbf{X}}(\mathbf{Y}) - \mathbf{X}\|^2 > 0\}}\right]$$

According to the latter lemma and Definition 1, we hence characterize the MLE as the optimal estimator according to the 0-1 loss, hence we can denote $\mathbf{X}_{\text{0-1}} = \mathbf{X}_{\mathrm{MLE}}$ and by

$$\begin{aligned}
\mathrm{MEP}_n(\lambda) &= \mathbb{P}_n \mathbb{Q}_{y|x}[\mathbf{X}_{\mathrm{MLE}}(\mathbf{Y}) \neq \mathbf{X}] \\
&= \mathbb{E}\left[\mathbb{1}_{\{\|\mathbf{X}_{MLE} - \mathbf{X}\|^2 > 0\}}\right]
\end{aligned}$$

the minimum error probability obtained by such estimator. Based on the same definition, the constrained version is further defined.

A. Application to the sparse tensor-PCA problem

In the following we assume for $d \geq 2$ the following sparse tensor-PCA model with observations

$$\mathbf{Y} = \sqrt{\lambda}\mathbf{x}^{\otimes d} + \mathbf{Z}, \tag{4}$$

that corresponds to the additive Gaussian model defined in Equation (1) using $\mathbf{X} = \mathbf{x}^{\otimes d}$, with $\mathbf{x} \in \mathbb{R}^p$ and $n = p^d$, and considering the Frobenius norm for tensors in $\mathbb{R}^n$ [1]. A discrete uniform prior $\tilde{\mathbb{P}}_p$ over $\mathcal{S}_{p-1}$ induces a discrete uniform prior $\mathbb{P}_n$ over $\mathcal{S}_{n-1}$, hence the assumption of the model defined in Equation (1) are satisfied. Here and in the following we assume $\tilde{\mathbb{P}}_p$ to be a Bernoulli prior over the subset of the unit sphere with k binary entries, hence

$$\mathbf{x} \in \left\{0, \frac{1}{\sqrt{k}}\right\}^p \cap \mathcal{S}_{p-1} = \mathcal{C}_{p,k} = \mathit{supp}(\tilde{\mathbb{P}}_p)$$

The cardinality of the hypothesis space is hence $M_p = \binom{p}{k}$, and $\lambda_n = \log M_p = k \log\left(\frac{p}{k}\right)(1 + o(1))$. The prior $\tilde{\mathbb{P}}_p$ maps to the uniform prior $\mathbb{P}_n$ over the space $\mathit{supp}(\mathbb{P}_n) \subsetneq \mathcal{C}_{n,s} = \{0, s^{-1/2}\}^n \cap \mathcal{S}_{n-1}$, where $s = k^d$. Note here the difference between the $\mathit{supp}(\mathbb{P}_n)$, that is the set of tensors formed as $\mathbf{x}^{\otimes d}$ with $\mathbf{x} \in \mathcal{C}_{p,k}$, and the set $\mathcal{C}_{n,s}$, that is the set of tensors with *any* s entries equal to $s^{-1/2}$. We here study the constrained estimators C-MMSE and C-MEP on the set $\mathcal{C}_{n,s}$, as it allows an easy characterization in terms of the unconstrained one. The main theorem of this section gives a sufficient condition for an AoN phenomenon to hold for the MEP.

Theorem 2. *For the sparse Bernoulli tensor-PCA model defined in Equation* (4)*, with* $k \in o\left(\log^{\frac{1}{4d-1}} p\right)$*, the* MEP *satisfies the weak AoN transition as:*

$$\begin{aligned} \liminf_{n\to\infty} \mathrm{MEP}_n(\beta\lambda_n) &\geq \frac{1}{4} \qquad && \beta < 1 \\ \lim_{n\to\infty} \mathrm{MEP}_n(\beta\lambda_n) &= 0 && \beta > 1 \end{aligned}$$

The same transition holds for the $\text{C-MEP}_n(\beta\lambda_n)$*.*

We conjecture that the MLE undergoes a strict AoN transition, but further work is necessary to establish the full characterization of the MLE in the impossibility regime.

Proof. The main idea of the proof is to relate the MMSE to the MEP studying the constrained counterpart of both. For these latter two quantities, a simple first-moment method can be applied, as there exists a minimum non-vanishing distance between any two points in the constrained set $\mathcal{C}_{n,s}$. Hence, for any estimator $\hat{\mathbf{X}}(\mathbf{Y})$, using the Markov inequality we can derive the following bound on its error probability:

$$\begin{aligned} \mathrm{EP}_n(\hat{\mathbf{X}}) = \mathbb{E}\left[\mathbb{1}_{\{\|\hat{\mathbf{X}}-\mathbf{X}\|^2>0\}}\right] &= \mathbb{P}\left[\|\hat{\mathbf{X}} - \mathbf{X}\|^2 \geq \frac{2}{s}\right] \\ &\leq \frac{s}{2}\mathbb{E}\left[\|\hat{\mathbf{X}} - \mathbf{X}\|^2\right] \end{aligned}$$

Note that this is only possible if the estimator is constrained in $\mathcal{C}_{n,s}$. The following bounds can hence be derived with the full proof given in the appendix.

Lemma 6. *Given the problem in Equation* (4)*, the following bounds hold:*

$$\frac{s}{2}\,\text{C-MMSE}_n(\lambda) \geq \text{C-MEP}_n(\lambda) \tag{5}$$

$$\text{C-MEP}_n(\lambda) \geq \frac{1}{4}\,\text{C-MMSE}_n(\lambda)^2 \tag{6}$$

$$\mathrm{MEP}_n(\lambda) \geq \frac{1}{4}\,\mathrm{MMSE}_n(\lambda)^2 \tag{7}$$

The inequality given in Equation (5) relates now the C-MMSE to C-MEP, such that if the C-MMSE is small enough, then the MEP is small too. We can further have a bound that relates the MMSE to the C-MMSE in the same direction, so as to derive a chain of inequalities between the MMSE and the MEP. This is given by the following lemma.

Lemma 7. *For any* $\epsilon > 0$, $\mathrm{MMSE}_n(\lambda) < \epsilon$ *if and only if* $\text{C-MMSE}_n(\lambda) < 4\epsilon s$*.*

Proof. Let us define the (constrained-) MSE distance as, respectively,

$$\text{C-MSE}_n(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X}_{\text{C-}l_2}(\mathbf{Y}) - \mathbf{X}\|^2$$

and

$$\mathrm{MSE}_p(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X}_{l_2}(\mathbf{Y}) - \mathbf{X}\|^2.$$

We can write the expectation conditioning on the event A that the first distance is smaller then a given $\delta < \frac{1}{2s}$,

$$A = \{(\mathbf{X}, \mathbf{Y}) \colon \text{C-MSE}_n(\mathbf{X}, \mathbf{Y}) \leq \delta\}$$

as:

$$\begin{aligned} \text{C-MMSE}_n(\lambda) &= \mathbb{E}\left[\text{C-MSE}_n(\mathbf{X}, \mathbf{Y})\right] \\ &= \mathbb{E}\left[\text{C-MSE}_n(\mathbf{X}, \mathbf{Y})|A\right]\mathbb{P}[A] + \\ &\quad + \mathbb{E}\left[\text{C-MSE}_n(\mathbf{X}, \mathbf{Y})|A^c\right]\mathbb{P}[A^c]. \end{aligned} \tag{8}$$

We now characterize the optimal constrained l_2 estimator as the simple rounding of the top entries of the standard posterior average estimator.

Lemma 8. *The optimal estimator constrained in the hypothesis space* $\mathcal{C}_{n,s}$ *for the problem in Equation* (4) *for the MSE reads*

$$\mathbf{X}_{\text{C-}l_2} = \operatorname*{argmin}_{\hat{\mathbf{X}}(\mathbf{Y}) \in \mathcal{C}_{n,s}} \mathbb{E}[\|\hat{\mathbf{X}}(\mathbf{Y}) - \mathbf{X}\|^2] = \mathrm{Top}_s\left(\mathbb{E}[\mathbf{X}|\mathbf{Y}]\right),$$

where the $\mathrm{Top}_s(\cdot)$ *operator rounds the top* s *entries of* $\mathbf{X}$ *to* $s^{-1/2}$*, and zeros out all other entries.*

[1] Note that this problem can also encompass the planted problem in hypergraph, with observations in the *upper-triangular part* of the tensor as

$$\mathbf{Y} = (\sqrt{\lambda}\mathbf{x}^{\otimes d} + \mathbf{Z})\mathbb{1}_{\{i_1 < \cdots < i_d\}}$$

and with $n = \binom{p}{d}$. Results easily extend to the hypergraph variation seamlessly.

Now we can easily note that the following geometrical lemma:

Lemma 9. *For any integer n and s and $\mathbf{U} \in \mathcal{C}_{n,s}$, $\mathbf{V} \in \left[0, s^{-1/2}\right]^n$ and $\delta < \frac{1}{2s}$, such that $\|\mathbf{U} - \mathbf{V}\|^2 \leq \delta$,*

$$\mathrm{Top}_s(\mathbf{V}) = \mathbf{U}.$$

Combining Lemma 8 and Lemma 9 it follows that

$$\mathbb{E}\left[\text{C-MSE}_n(\mathbf{X}, \mathbf{Y}) | A\right] = 0 \tag{9}$$

as $\mathrm{Top}_s(\mathbf{X}_{\text{C-}l_2}) = \mathbf{X}_{\text{C-}l_2}$. Using the same decomposition for the MMSE with respect to the event

$$B = \{(\mathbf{X}, \mathbf{Y})\colon\ \mathrm{MSE}_n(\mathbf{X}, \mathbf{Y}) \leq \delta\}$$

we get

$$\begin{aligned}\mathrm{MMSE}_n(\beta) = \mathbb{E}[\mathrm{MSE}_n(\mathbf{X}, \mathbf{Y}) | B]\mathbb{P}[B] + \\ + \mathbb{E}[\mathrm{MSE}_n(\mathbf{X}, \mathbf{Y}) | B^c]\mathbb{P}[B^c] < \epsilon \end{aligned} \tag{10}$$

Bounding as $\mathbb{E}[\mathrm{MSE}_n(\mathbf{X}, \mathbf{Y}) | B] \geq 0$, $\mathbb{P}[B] \geq 0$ and $\mathbb{E}[\mathrm{MSE}_n(\mathbf{X}, \mathbf{Y}) | B^c] \geq \delta$ we get from Equation (10)

$$\mathbb{P}[B^c] < \frac{\epsilon}{\delta}. \tag{11}$$

From Lemma 8, we have further that for $\delta < \frac{1}{2s}$,

$$B \subset \{\text{C-MSE}_n(\mathbf{X}, \mathbf{Y}) = 0\} \subset A$$

hence that

$$\mathbb{P}[A^c] \leq \mathbb{P}[B^c]. \tag{12}$$

Plugging Equations 9, 11 and 12 into the decomposition in Equation (8) and using the fact that $\text{C-MSE}_n(\mathbf{X}, \mathbf{Y}) \leq 2$ we finally get

$$\text{C-MMSE}_n(\lambda) \leq \frac{2\epsilon}{\delta}.$$

The theorem follows from the arbitrariness of $\delta < \frac{1}{2s}$. □

Plugging in the result of Lemma 7 and Equation (5) we get the further lemma that relates the MMSE in the recovery regime to the C-MMSE and the MEP.

Lemma 10. $\mathrm{MMSE}_n(\lambda_n) \in o(1/s)$ *if and only if* $\text{C-MMSE}_n(\lambda_n) \in o(1)$. *If* $\mathrm{MMSE}_n(\lambda_n) \in o(1/s^2)$ *then* $\mathrm{MEP}_n(\lambda_n) \in o(1)$

In the same regime $\beta > 1$, we can hence use the results on the generalized AoN, Theorem 1, to have the MMSE to be $o(1/s^2)$. For such theorem to hold, we use the following lemma on the overlap rate function of the sparse tensor-PCA problem.

Lemma 11 (Proof given in Proposition 3, [2])**.** *For the Bernoulli sparse tensor-PCA problem with signal $\mathbf{X} = \mathbf{x}^{\otimes d}$, $d \geq 2$, and $\mathbf{x} \in \{0, 1/\sqrt{k}\}^p \cap \mathcal{S}_{p-1}$ the following bound on the overlap rate function of the tensors $\mathbf{X}, \mathbf{X}'$ holds for any $t \in [0,1]$:*

$$r_n(t) \geq \sqrt{t} - \frac{\mathcal{O}(1)}{\lambda_n}$$

Combining Lemma 11, Theorem 1, Lemma 7 and Lemma 6 we finally get the claim of the main theorem as

$$\sqrt{t} \geq \frac{2t}{1+t}$$

for $t \in [0,1]$ and as $k \in o\left(\log^{\frac{1}{4d-1}} p\right)$ implies

$$\lambda_n^{-1/2+\epsilon} = (k\log(p/k) + o(1))^{-1/2+\epsilon} \in o(1/s^2) = o(1/k^{2d})$$

□

Analogously, we can prove the following transition for the constrained MMSE.

Theorem 3. *For the sparse Bernoulli tensor PCA model with $k \in o\left(\log^{\frac{1}{2d-1}} p\right)$, the* C-MMSE *satisfies the AoN transition*

$$\begin{aligned} \liminf_{n\to\infty} \text{C-MMSE}_n(\beta\lambda_n) &\geq 1 \qquad \beta < 1 \\ \lim_{n\to\infty}\ \text{C-MMSE}_n(\beta\lambda_n) &= 0 \qquad \beta > 1 \end{aligned}$$

Proof. The proof follows the same steps of the proof of Theorem 2, with only two main differences. First, in the impossibility regime, we can use the stronger bound $\text{C-MMSE}_n(\lambda) \geq \mathrm{MMSE}_n(\lambda)$ in place of Equation (6) to get the first part of the theorem. In the recovery regime, we can use the weaker requirement $\mathrm{MMSE}_n(\beta\lambda_n) \in o(1/s)$ that is satisfied by the assumption of the theorem $k \in o\left(\log^{\frac{1}{2d-1}} p\right)$. □

V. Conclusion

In this paper, we analysed the maximum-likelihood estimator for the sparse tensor-PCA problem with Bernoulli prior. We established that this estimator undergoes a weak AoN transition and conjectured that this transition is equivalent to the MMSE transition. The proof follows from the connection of the MLE to the MMSE using the first and second-moment method in the constrained signal space, and hence it is of independent interest as it can lead to further results from the community.

While this paper sets a first step in understanding a wider range of optimal estimators in sparse high-dimensional inference problems, a general theory of the all-or-nothing statistical transition is still lacking. This theory could provide a wider understating of the phenomenon, including the analysis of vectorial estimators for planted matrix and tensor PCA problems, that is not here considered and is carried out in the dense setting using rigorous tools of statistical physics and replica methods (see [10]).

The same methods established recently that tractable estimators, like the approximate-message-passing algorithms, undergo the same all-or-nothing transition in the sparse matrix PCA problem [16]. The extension of these results to the sparse tensor-PCA problem and currently optimal algorithms for this problem, like averaged gradient descent [13] and sum-of-squares algorithms [11], is of crucial importance.

Acknowledgement

This work was supported in part by NSF Center on Science of Information Grants CCF-0939370 and NSF Grants CCF-1524312, CCF-2006440, CCF-2007238.

Appendix A
Postponed proofs of Section III

Proof of Lemma 1. Using, respectively, the maximum entropy bound and 1/2 Lipschitz-continuity of the function (see Lemma 14 and [2]) we can see first that

$$\begin{aligned}0 \leq \frac{1}{\lambda_n} D(\mathbb{Q}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) - \frac{1}{2}(\beta-1) \leq \\ \leq \frac{1}{\lambda_n} D(\mathbb{Q}_{\lambda_n,n}||\mathbb{Q}_{0,n}) \in o(\tau_n)\end{aligned}$$

where the first inequality comes from the maximum entropy bound, the second from Lipschitz-continuity, and the inclusion follows from the assumption of the theorem. For convenience let us denote

$$\frac{1}{\lambda_n} D(\mathbb{Q}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) = \frac{1}{2}(\beta-1) + f_n(\beta)$$

where $\lim_{n\to\infty} \frac{f_n(\beta)}{\tau_n} = 0$ for any $\beta > 1$. We can now use the I-MMSE relation (see [2], [21]), such that $\mathrm{MMSE}_n(\beta\lambda_n) = 1 - 2\frac{d}{d\beta}\frac{1}{\lambda_n} D(\mathbb{Q}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n})$.

$$\begin{aligned}\mathrm{MMSE}(\beta\lambda_n) &= 1 - 2\frac{d}{d\beta}\left(\frac{1}{2}(\beta-1) + f_n(\beta)\right) \\ &= -2\frac{d}{d\beta} f_n(\beta)\end{aligned}$$

hence

$$\begin{aligned}\lim_{n\to\infty} \frac{\mathrm{MMSE}(\beta\lambda_n)}{\tau_n} &= \lim_{n\to\infty} -2\frac{1}{\tau_n}\frac{d}{d\beta} f_n(\beta) \\ &= \lim_{n\to\infty} -2\frac{d}{d\beta}\frac{1}{\tau_n} f_n(\beta) = 0\end{aligned}$$

where the second equality follows from the linearity of differentiation and the third from the interchanging of limit and differentiation under uniform convergence. □

Proof of Lemma 2. Following the proof of Theorem 5 in [22], and defining the function $Z(\mathbf{Y}) = \frac{\mathbb{Q}_{\beta\lambda_n,n}(\mathbf{Y})}{\mathbb{Q}_{0,n}(\mathbf{Y})}$ we have that

$$\begin{aligned}D(\tilde{\mathbb{Q}}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) - D(\mathbb{Q}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) \geq \\ \geq \mathbb{E}_{\tilde{\mathbb{Q}}_{\beta\lambda_n,n}} \log Z(\mathbf{Y}) - \mathbb{E}_{\mathbb{Q}_{\beta\lambda_n,n}} \log Z(\mathbf{Y})\end{aligned} \tag{13}$$

Using the definition of the conditional pdf, we have

$$\begin{aligned}\tilde{\mathbb{Q}}_{\beta\lambda_n,n}(\mathbf{Y}) &= \mathbb{Q}_{\beta\lambda_n,n}(\mathbf{Y}|\Omega_n) = \mathbb{E}_{\mathbb{P}_n}\mathbb{Q}_{y|x}(\mathbf{Y}|\Omega_n,\mathbf{X}) \\ &= \mathbb{E}_{\mathbb{P}_n} \frac{\mathbb{Q}_{y|x}(\mathbf{Y}|\mathbf{X})\mathbb{Q}_{xy}[\Omega_n|\mathbf{X},\mathbf{Y}]}{\mathbb{Q}_{xy}[\Omega_n|\mathbf{X}]} \\ &= \mathbb{E}_{\mathbb{P}_n} \frac{\mathbb{Q}_{y|x}(\mathbf{Y}|\mathbf{X})\mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y})}{\mathbb{Q}_{xy}[\Omega_n|\mathbf{X}]},\end{aligned} \tag{14}$$

where $\mathbb{1}_A(\cdot)$ is the indicator function of set A. Plugging Equation (14) into Equation (13) we get

$$\begin{aligned}D(\tilde{\mathbb{Q}}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) - D(\mathbb{Q}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) \geq \\ \geq \mathbb{E}_{\mathbb{P}_n} \frac{\mathbb{E}_{\mathbb{Q}_{y|x}}\left[(\mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y}) - \mathbb{Q}_{xy}[\Omega_n|\mathbf{X}])\log Z(\mathbf{Y})\right]}{\mathbb{Q}_{xy}[\Omega_n|\mathbf{X}]}\end{aligned} \tag{15}$$

Using the Cauchy-Schwartz inequality we can bound the expectation over $\mathbf{Y}$ as

$$\begin{aligned}\left|\mathbb{E}_{\mathbb{Q}_{y|x}}\left[(\mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y}) - \mathbb{Q}_{xy}[\Omega_n|\mathbf{X}])\log Z(\mathbf{Y})\right]\right| \leq \\ \sqrt{\mathbb{E}_{\mathbb{Q}_{y|x}}(\mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y}) - \mathbb{Q}_{xy}[\Omega_n|\mathbf{X}])^2 \cdot \mathbb{E}_{\mathbb{Q}_{y|x}}\left[\log^2 Z(\mathbf{Y})\right]}\end{aligned} \tag{16}$$

It is easy to see that

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}_{y|x}}&\left[(\mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y}) - \mathbb{Q}_{xy}[\Omega_n|\mathbf{X}])^2\right] = \\ &= \mathbb{E}_{\mathbb{Q}_{y|x}}\Big[(\mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y}) + \mathbb{Q}_{xy}[\Omega_n|\mathbf{X}]^2 + \\ &\qquad - 2\cdot\mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y})\mathbb{Q}_{xy}[\Omega_n|\mathbf{X}]^2\Big] \\ &= \mathbb{E}_{\mathbb{Q}_{y|x}}\left[\mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y})\right] + \mathbb{Q}_{xy}[\Omega_n|\mathbf{X}]^2 + \\ &\qquad - 2\cdot\mathbb{E}_{\mathbb{Q}_{y|x}}\left[\mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y})\right]\mathbb{Q}_{xy}[\Omega_n|\mathbf{X}] = \\ &= Q_{xy}[\Omega_n|\mathbf{X}] - \mathbb{Q}_{xy}[\Omega_n|\mathbf{X}]^2\end{aligned}$$

hence recombining the latter and Equation (15) and Equation (16) we get

$$\begin{aligned}D(\tilde{\mathbb{Q}}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) - D(\mathbb{Q}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) \geq \\ \geq -\mathbb{E}_{\mathbb{P}_n}\sqrt{\frac{1-\mathbb{Q}_{xy}[\Omega_n|\mathbf{X}]}{\mathbb{Q}_{xy}[\Omega_n|\mathbf{X}]}\mathbb{E}_{\mathbb{Q}_{y|x}}\left[\log^2 Z(\mathbf{Y})\right]}.\end{aligned}$$

Using again the Cauchy-Schwartz inequality over the expectations on $\mathbb{P}_n$ we finally get

$$\begin{aligned}&D(\tilde{\mathbb{Q}}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) - D(\mathbb{Q}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) \geq \\ &\quad -\sqrt{\mathbb{E}_{\mathbb{P}_n}\left(\frac{1-\mathbb{Q}_{xy}[\Omega_n|\mathbf{X}]}{\mathbb{Q}_{xy}[\Omega_n|\mathbf{X}]}\right)^2} \cdot \sqrt{\mathbb{E}_{\mathbb{P}_n}\mathbb{E}_{\mathbb{Q}_{y|x}}\left[\log^2 Z(\mathbf{Y})\right]} \\ &\quad = -o(\tau_n)\cdot\sqrt{\mathbb{E}_{\mathbb{Q}_{\beta\lambda_n,n}}\left[\log^2 Z(\mathbf{Y})\right]}\end{aligned}$$

where the equality comes from the assumption on the event Ω_n. Using now the result from Proposition 3 in [2],

$$\sqrt{\mathbb{E}_{\mathbb{Q}_{\beta\lambda_n,n}}\left[\log^2 Z(\mathbf{Y})\right]} = \mathcal{O}(\log M_n)$$

we get the claim. □

Proof of Lemma 3. We can see easily that

$$\begin{aligned}\mathbb{Q}_{xy}[\Omega_n|\mathbf{X}] &= \mathbb{Q}_{y|z}[\Omega_n|\mathbf{X}] \\ &= \mathbb{P}_z[|\langle\mathbf{X},\mathbf{Z}\rangle| \leq C_n] \\ &= 2\phi(C_n) - 1 \\ &= 1 - \sqrt{\frac{2}{\pi}}\frac{1}{C_n}e^{-C_n^2/2}\left(1 + \mathcal{O}\left(\frac{1}{C_n}\right)\right)\end{aligned}$$

where in the second equality we used the fact that $\mathbf{Y} = \sqrt{\lambda_n}\mathbf{X} + \mathbf{Z}$ and that $\|\mathbf{X}\|^2 = 1$, the third follows from the fact that $\langle\mathbf{X},\mathbf{Z}\rangle$ is a univariate Gaussian random variable distributed as $\mathcal{N}(0,1)$, and $\phi(\cdot)$ is the cdf of the standard Gaussian, with asymptotics given, for large x, as $\phi(x) = 1 - \frac{1}{\sqrt{2\pi}x}e^{-x^2/2}\left(1 + \mathcal{O}\left(\frac{1}{x^2}\right)\right)$. We hence have the claim as

$$1 - \mathbb{Q}_{xy}[\Omega_n|\mathbf{X}] = \mathcal{O}\left(\frac{1}{C_n}e^{-C_n^2/2}\right) \in o(\tau_n).$$

□

Proof of Lemma 4. Using the Jensen inequality, it can be easily seen that for any two distributions

$$D(p||q) = \mathbb{E}_p \log \frac{p(x)}{q(x)} \leq \log \left(\mathbb{E}_p \frac{p(x)}{q(x)} \right) = \log \left(\mathbb{E}_q \left(\frac{p(x)}{q(x)} \right)^2 \right)$$

To bound the KL, we can hence study the ratio p/q. Using eq. (14) we can hence write

$$\frac{\tilde{\mathbb{Q}}_{\beta\lambda_n,n}(\mathbf{Y})}{\mathbb{Q}_{0,n}(\mathbf{Y})} = \mathbb{E}_{\mathbb{P}_n} \frac{1}{\mathbb{Q}_{xy}[\Omega_n|\mathbf{X}]} \frac{\mathbb{Q}_{\beta\lambda_n,n}(\mathbf{Y}|\mathbf{X})}{\mathbb{Q}_{\beta 0,n}(\mathbf{Y})} \mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y}).$$

Plugging in the definition of the model for $\mathbf{Y}$ we can easily see that

$$\begin{aligned}\frac{\mathbb{Q}_{\beta\lambda_n,n}(\mathbf{Y}|\mathbf{X})}{\mathbb{Q}_{0,n}(\mathbf{Y})} &= \frac{\exp(-\frac{1}{2}\|\mathbf{Y} - \sqrt{\lambda_n}\mathbf{X}\|^2)}{\exp(-\frac{1}{2}\|\mathbf{Y}\|^2)} \\ &= \exp\left(\sqrt{\lambda_n}\langle \mathbf{X}, \mathbf{Y}\rangle - \frac{\lambda_n}{2} \right)\end{aligned}$$

where in the second inequality we used the fact that $\|\mathbf{X}\|^2 = 1$. Using the latter, we can obtain

$$\left(\frac{\tilde{\mathbb{Q}}_{\beta\lambda_n,n}(\mathbf{Y}|\mathbf{X})}{\mathbb{Q}_{0,n}(\mathbf{Y})} \right)^2 = \\ = \mathbb{E}_{\mathbb{P}_n^{\otimes 2}} \frac{\exp\left(\sqrt{\lambda_n}\langle \mathbf{X}+\mathbf{X}', \mathbf{Y}\rangle - \lambda_n\right)}{\mathbb{Q}_{xy}[\Omega_n|\mathbf{X}] Q_{xy}[\Omega_n|\mathbf{X}']} \mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y}) \mathbb{1}_{\Omega_n}(\mathbf{X}',\mathbf{Y})$$

Using the fact that for $C_n = \sqrt{\log \lambda_n}$ and $\lambda_n^{-1/2+\epsilon} \in o(\tau_n)$ we can satisfy the assumption of Lemma 3, we can exchange the small-o notation and the integrals as

$$\begin{aligned}\left(\frac{\tilde{\mathbb{Q}}_{\beta\lambda_n,n}(\mathbf{Y}|\mathbf{X})}{\mathbb{Q}_{0,n}(\mathbf{Y})} \right)^2 &= \\ = (1+o(\tau_n)) &\mathbb{E}_{\mathbb{P}_n^{\otimes 2}} \left[\mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y}) \mathbb{1}_{\Omega_n}(\mathbf{X}',\mathbf{Y}) \right. \\ &\left. \times \exp\left(\sqrt{\lambda_n}\langle \mathbf{X}+\mathbf{X}', \mathbf{Y}\rangle - \lambda_n \right) \right]\end{aligned}$$

We hence have a bound for the KL that reads

$$\frac{1}{\lambda_n} D(\tilde{\mathbb{Q}}_{\beta\lambda_n,n}||\mathbb{Q}_{0,n}) \leq \frac{1}{\lambda_n} \log \left[\mathbb{E}_{\mathbb{P}_n^{\otimes 2}} m_n(\mathbf{X},\mathbf{X}') \right] + o(\tau_n)$$

where m_n is defined as

$$\begin{aligned}m_n(\mathbf{X},\mathbf{X}') \coloneqq \mathbb{E}_{\mathbb{Q}_{0,n}} &\left[\mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y}) \mathbb{1}_{\Omega_n}(\mathbf{X}',\mathbf{Y}) \right. \\ &\left. \times \exp\left(\sqrt{\lambda_n}\langle \mathbf{X}+\mathbf{X}', \mathbf{Y}\rangle - \lambda_n \right) \right]\end{aligned}$$

and we used the fact that $\frac{o(\tau_n)}{\lambda_n} \in o(\tau_n)$ and where we used Fubini's theorem to exchange the order of the integrals. Note that $\mathbb{Q}_{0,n} = \mathbb{P}_z = \mathcal{N}(0,1)$. Now it is sufficient to prove that

$$\frac{1}{\lambda_n} \log \left[\mathbb{E}_{\mathbb{P}_n^{\otimes 2}} m_n(\mathbf{X},\mathbf{X}') \right] \leq \sup_{t\in[0,1]} \left(\frac{t}{t+1} - \frac{r_n(t)}{2} \right) + o(\tau_n)$$

to get the claim. We can readily see that the function m_n depends only on the overlap $\rho = \langle \mathbf{X}, \mathbf{X}'\rangle$ due to the rotational invariance of the Gaussian pdf. Using Lemma 15, the definition of r_n, the monotonicity of the exponential function and the simple inequality

$$sup(f+g) \leq \sup f + \sup g$$

we get

$$\begin{aligned}\frac{1}{\lambda_n} \log \mathbb{E}_{\mathbb{P}_n^{\otimes 2}} m_n(\rho) \leq \frac{\log(2L_n)}{\lambda_n} + \\ + \sup_{t\in[-1,1]} \left(\left(\frac{t}{1+t} \right)_+ - \frac{r_n(t)}{2} \right) + \frac{C_n}{\lambda_n^{1/2}} + \mathcal{O}\left(\frac{1}{L_n} \right)\end{aligned}$$

We can easily observe that the supremum can be limited to the interval $t \in [0,1]$ noting that $r_n(-1) = 0$ and $r_n(t)$ is a non-negative function. The claim then follows easily from the assumption of $\lambda_n^{-1/2+\epsilon} \in o(\tau_n)$ and choosing $L_n = \lfloor \lambda_n^{1/2} \rfloor$ and $C_n = \sqrt{\log \lambda_n}$. □

Appendix B
Postponed proofs of Section IV

Proof of Lemma 5.

$$\begin{aligned}\mathbb{P}_n \mathbb{Q}_{y|x}[\hat{\mathbf{X}}(\mathbf{Y}) \neq \mathbf{X}] &= \mathbb{E}_{P_n} \mathbb{E}_{\mathbb{Q}_{y|x}} \mathbb{1}_{\{\hat{\mathbf{X}}(\mathbf{Y}) \neq \mathbf{X}\}} \\ &= 1 - \mathbb{E}_{P_n} \mathbb{E}_{\mathbb{Q}_{y|x}} \mathbb{1}_{\{\hat{\mathbf{X}}(\mathbf{Y}) = \mathbf{X}\}} \\ &= 1 - \int d\mathbf{Y} \mathbb{Q}_{y|x}(\mathbf{Y}|\hat{\mathbf{X}}(\mathbf{Y})) \mathbb{P}_n(\hat{\mathbf{X}}(\mathbf{Y}))\end{aligned}$$

hence $\operatorname*{argmin}_{\hat{\mathbf{X}}(\cdot)} \mathbb{P}_n \mathbb{Q}_{y|x}[\hat{\mathbf{X}}(\mathbf{Y}) \neq \mathbf{X}]$ satisfies for every $\mathbf{Y}$

$$\begin{aligned}\left(\operatorname*{argmin}_{\hat{\mathbf{X}}(\cdot)} \mathbb{P}_n \mathbb{Q}_{y|x}[\hat{\mathbf{X}}(\mathbf{Y}) \neq \mathbf{X}] \right)(\mathbf{Y}) &= \\ = \operatorname*{argmax}_{\hat{\mathbf{X}}(\mathbf{Y})} &\mathbb{Q}_{y|x}(\mathbf{Y}|\hat{\mathbf{X}}(\mathbf{Y})) \mathbb{P}_n(\hat{\mathbf{X}}(\mathbf{Y})),\end{aligned}$$

that, for uniform prior, corresponds to the MLE estimator. □

Proof of Lemma 6. The two bounds are, respectively, given by the first and second-moment methods. Equation (5) follows easily from from the Markov inequality as for any estimator $\hat{\mathbf{X}}(\mathbf{Y})$,

$$\begin{aligned}\text{EP}_n(\hat{\mathbf{X}}) = \mathbb{E}\left[\mathbb{1}_{\{\|\hat{\mathbf{X}}-\mathbf{X}\|^2 > 0\}} \right] &= \Pr\left[\|\hat{\mathbf{X}}-\mathbf{X}\|^2 \geq \frac{2}{s} \right] \\ &\leq \frac{s}{2} \mathbb{E}\left[\|\hat{\mathbf{X}}-\mathbf{X}\|^2 \right]\end{aligned}$$

hence

$$\begin{aligned}\text{C-MEP}_n(\lambda) &= \min_{\hat{\mathbf{X}}(\mathbf{Y}) \in \mathcal{C}_{n,s}} \mathbb{E}\left[\|\hat{\mathbf{X}}-\mathbf{X}\|_0 \right] \\ &\leq \frac{s}{2} \min_{\hat{\mathbf{X}}(\mathbf{Y}) \in \mathcal{C}_{n,s}} \mathbb{E}[\|\hat{\mathbf{X}}-\mathbf{X}\|^2] = \frac{s}{2} \text{C-MMSE}_n(\lambda).\end{aligned}$$

To prove the second bound, we use the Paley–Zygmund inequality that reads for a general *positive* random variable Z and $0 \leq \theta \leq \mathbb{E}[Z]$

$$\Pr[Z > \theta] \geq \frac{(\mathbb{E}[Z] - \theta)^2}{\mathbb{E}[Z^2]}. \tag{17}$$

Using Equation (17) for the random variable $\|\hat{\mathbf{X}}-\mathbf{X}\|^2$ we obtain

$$\text{EP}_n(\hat{\mathbf{X}}) = \Pr\left[\|\hat{\mathbf{X}}-\mathbf{X}\|^2 > 0 \right] \geq \frac{\mathbb{E}\left[\|\hat{\mathbf{X}}-\mathbf{X}\|^2 \right]^2}{\mathbb{E}\left[\|\hat{\mathbf{X}}-\mathbf{X}\|^4 \right]}$$

from which we get

$$\begin{aligned}\mathrm{MEP}_n(\lambda) &= \min_{\hat{\mathbf{X}}(\mathbf{Y})\in\bar{\mathcal{S}}_{n-1}} \Pr\left[\|\hat{\mathbf{X}}-\mathbf{X}\|^2 > 0\right] \\ &\geq \min_{\hat{\mathbf{X}}(\mathbf{Y})\in\bar{\mathcal{S}}_{n-1}} \frac{\mathbb{E}[\|\hat{\mathbf{X}}-\mathbf{X}\|^2]^2}{\mathbb{E}[\|\hat{\mathbf{X}}-\mathbf{X}\|^4]} \\ &\geq \frac{\left(\min_{\hat{\mathbf{X}}(\mathbf{Y})\in\bar{\mathcal{S}}_{n-1}} \mathbb{E}[\|\hat{\mathbf{X}}-\mathbf{X}\|^2]\right)^2}{\max_{\hat{\mathbf{X}}(\mathbf{Y})\in\bar{\mathcal{S}}_{n-1}} \mathbb{E}[\|\hat{\mathbf{X}}-\mathbf{X}\|^4]} \\ &\geq \frac{1}{4}\,\mathrm{MMSE}_p(\lambda)^2\end{aligned}$$

where in the last inequality we used the definition of the MMSE and the fact that for any two vectors

$$\boldsymbol{a},\boldsymbol{b} \in \bar{\mathcal{S}}_{n-1}, \quad \|\boldsymbol{a}-\boldsymbol{b}\|^2 \leq 2.$$

The third inequality follows analogously. □

Proof of Lemma 8.

$$\mathbb{E}[\|\hat{\mathbf{X}}-\mathbf{X}\|^2] = \mathbb{E}\left[\|\hat{\mathbf{X}}\|^2\right] + \mathbb{E}\left[\|\mathbf{X}\|^2\right] - 2\mathbb{E}\left[\sum_i^n \hat{X}_i X_i\right]. \tag{18}$$

Given the constraint on the estimator and on the fact that $\|\mathbf{X}\|^2 = \|\hat{\mathbf{X}}\|^2 = 1$, the optimization problem becomes:

$$\operatorname*{argmax}_{\hat{\mathbf{X}}(\mathbf{Y})\in\mathcal{C}_{n,s}} \mathbb{E}_{\mathbb{Q}_{\lambda,n}} \sum_{i=1}^n \hat{\mathbf{X}}_i(\mathbf{Y})\mathbb{E}[X_i|\mathbf{Y}]$$

hence for every fixed $\mathbf{Y}$ the optimal estimator reads

$$\operatorname*{argmax}_{\hat{\mathbf{X}}(\mathbf{Y})\in\mathcal{C}_{n,s}} \sum_i^n \hat{\mathbf{X}}_i(\mathbf{Y})\mathbb{E}[\mathbf{X}|\mathbf{Y}]_i.$$

The theorem follows easily from linearity and from $\mathcal{C}_{n,s}$ being a binary set for which the greedy algorithm is optimal. □

Proof of Lemma 9.

$$\begin{aligned}\delta \geq \|\mathbf{U}-\mathbf{V}\|^2 &= \sum_{i:\, U_i=s^{-1/2}} \left(s^{-1/2}-V_i\right)^2 + \sum_{i:\, U_i=0} V_i^2 \\ &\geq \left(\max_{i:\, U_i=s^{-1/2}} \left(s^{-1/2}-V_i\right)\right)^2 + \left(\max_{i:\, U_i=0} V_i\right)^2 \\ &= \left(s^{-1/2} - \min_{i:\, U_i=s^{-1/2}} V_i\right)^2 + \left(\max_{i:\, U_i=0} V_i\right)^2\end{aligned}$$

Multiplying both sides of the latter inequality by s we get a inequality of the form $(1-a)^2+b^2 \leq s\delta < \frac{1}{2}$ for $a,b\in[0,1]$. It is easy to observe using simple calculus that this implies $a > b$ and hence

$$\min_{i:\, U_i=s^{-1/2}} V_i > \max_{i:\, U_i=0} V_i.$$

Given this condition, and the definition of the $\mathrm{Top}_s(\cdot)$ operator, it follows that $\mathrm{Top}_s(\mathbf{V}) = \mathbf{U}$. Note that the strict inequality $\delta < \frac{1}{2s}$ is essential to guarantee the strict inequality above and hence that there are no ties in the selection of the top s entries. □

APPENDIX C
USEFUL LEMMAS

Lemma 12. *The posterior average $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$ is the optimal estimator (MMSE) for the l_2 loss $l_2(\hat{\mathbf{x}},\mathbf{x}) = \|\hat{\mathbf{X}}-\mathbf{X}\|^2$ and the generic model in eq. (1), formally*

$$\operatorname*{argmin}_{\hat{\mathbf{X}}(\mathbf{Y})\in\bar{\mathcal{S}}_{n-1}} \mathbb{E}[\|\hat{\mathbf{X}}(\mathbf{Y})-\mathbf{X}\|^2] = \mathbb{E}[\mathbf{X}|\mathbf{Y}]$$

Proof.

$$\frac{\partial}{\partial\hat{\mathbf{X}}}\mathbb{E}[\|\hat{\mathbf{X}}-\mathbf{X}\|^2] = \mathbb{E}_{\mathbb{Q}_{\lambda,n}} \sum_{\mathbf{X}} \mathbb{P}_n(\mathbf{X}|\mathbf{Y})\frac{\partial}{\partial\hat{\mathbf{X}}}\|\hat{\mathbf{X}}-\mathbf{X}\|^2 = \tag{19}$$

$$= \mathbb{E}_{\mathbb{Q}_{\lambda,n}} \sum_{\mathbf{X}} \mathbb{P}_n(\mathbf{X}|\mathbf{Y})2(\hat{\mathbf{X}}-\mathbf{X}) = 0 \tag{20}$$

from which it follows easily that for every $\mathbf{Y}$, the gradient is equal to zero if $\hat{\mathbf{X}}(\mathbf{Y}) = \mathbb{E}[\mathbf{x}|\mathbf{Y}]$. It can be easily shown that the Hessian of the loss is positive semidefinite and hence satisfies the property of having a global minimum. □

We here further characterize the MLE estimator as following:

Lemma 13 (equivalent to Theorem 1 in [17])**.** *For the model defined in Equation (1), the MLE estimator reads:*

$$\mathbf{X}_{\text{0-1}}(\mathbf{Y}) = \operatorname*{argmax}_{\mathbf{X}\in\bar{\mathcal{S}}_n} \sum_{i_1,\dots,i_d} Y_{i_1,\dots,i_d} X_{i_1,\dots,i_d} \tag{21}$$

Proof.

$$\begin{aligned}\log\mathbb{Q}_{y|x}(\mathbf{Y}|\mathbf{X}) &= \sum_{i_1,\dots,i_d} \log\mathbb{Q}_{y|x}(Y_{i_1,\dots,i_d}|x_{i_1}\cdot\dots\cdot x_{i_d}) \\ &= \sum_{i_1,\dots,i_d} -\frac{1}{2}\log 2\pi - \frac{1}{2}(Y_{i_1,\dots,i_d} - \beta x_{i_1}\cdot\dots\cdot x_{i_d})^2 \\ &= \sum_{i_1,\dots,i_d} -\frac{1}{2}\log 2\pi - \frac{1}{2}\beta^2 x_{i_1}^2\cdot\dots\cdot x_{i_d}^2 + \\ &\qquad - \frac{1}{2}Y_{i_1,\dots,i_d}^2 + \beta Y_{i_1,\dots,i_d} x_{i_1}\cdot\dots\cdot x_{i_d} \\ &= -\frac{n}{2}\log 2\pi - \frac{1}{2}\lambda d!\binom{k}{d} - \frac{1}{2}\sum_{i_1,\dots,i_d} Y_{i_1,\dots,i_d}^2 + \\ &\qquad + \sqrt{\lambda}\sum_{i_1,\dots,i_d} Y_{i_1,\dots,i_d} x_{i_1}\cdot\dots\cdot x_{i_d}.\end{aligned}$$

The theorem follows easily noting that only the last term depends on $\mathbf{X}$. □

Lemma 14. *Given the setting in Equation (1), for all n and $\lambda > 0$, the function $\beta \to \frac{1}{\lambda}D(\mathbb{Q}_{\beta\lambda,n}||\mathbb{Q}_{0,n})$ is nonnegative, nondecreasing, 1/2-Lipschitz and satisfies the bound*

$$\frac{1}{\lambda}D(\mathbb{Q}_{\beta\lambda,n}||\mathbb{Q}_{0,n}) \geq \frac{1}{2} - \frac{\log M_n}{\lambda}.$$

Proof. The proof is given in Lemma 2 and Lemma 3 in [2]. □

Lemma 15. *Given the setting of the problem defined in Equation* (4) *and the function*

$$\begin{aligned} m_n(\mathbf{X},\mathbf{X}') :=\; & \mathbb{E}_{\mathbb{Q}_{0,n}}\big[\mathbb{1}_{\Omega_n}(\mathbf{X},\mathbf{Y})\mathbb{1}_{\Omega_n}(\mathbf{X}',\mathbf{Y}) \\ & \times \exp\left(\sqrt{\lambda_n}\langle\mathbf{X}+\mathbf{X}',\mathbf{Y}\rangle - \lambda_n\right)\big], \end{aligned}$$

there exist a constant $C > 0$ *such that for any integer sequence* L_n *the following bound holds:*

$$\begin{aligned} \mathbb{E}_{\mathbb{P}_n^{\otimes 2}} m_n(\rho) \le 2L_n \sup_{t\in[-1,1]} \exp\Bigg(& \lambda_n\left(\frac{t}{1+t}\right)_+ + \\ & + \log\mathbb{P}[\rho \ge t] + C_n\lambda_n^{1/2} + \mathcal{O}\left(\frac{\lambda_n}{L_n}\right)\Bigg) \end{aligned}$$

Proof. The proof follows, *mutatis mutandis*, the proof of Theorem 4 in [2], with the minor change of the definition of the events Ω_n that are here defined as $\{|\langle\mathbf{X},\mathbf{Z}\rangle - \sqrt{\lambda_n}| \le C_n\}$. □

References

[1] A. Jagannath, P. Lopatto, L. Miolane *et al.*, "Statistical thresholds for tensor pca," *Annals of Applied Probability*, vol. 30, no. 4, pp. 1910–1933, 2020.

[2] J. Niles-Weed and I. Zadik, "The all-or-nothing phenomenon in sparse tensor pca," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[3] W. Szpankowski, *Average case analysis of algorithms on sequences*. John Wiley & Sons, 2011, vol. 50.

[4] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[5] J. Barbier, M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborová, "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula," *Advances in Neural Information Processing Systems*, pp. 424–432, 2016.

[6] Y. Deshpande and A. Montanari, "Information-theoretically optimal sparse pca," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2014, pp. 2197–2201.

[7] E. Richard and A. Montanari, "A statistical model for tensor pca," in *Advances in Neural Information Processing Systems*, 2014, pp. 2897–2905.

[8] A. Montanari, D. Reichman, and O. Zeitouni, "On the limitation of spectral methods: From the gaussian hidden clique problem to rank one perturbations of gaussian tensors," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1572–1579, 2016.

[9] A. Perry, A. S. Wein, and A. S. Bandeira, "Statistical limits of spiked tensor models," *Ann. Inst. H. Poincaré Probab. Statist.*, vol. 56, no. 1, pp. 230–264, 02 2020. [Online]. Available: https://doi.org/10.1214/19-AIHP960

[10] T. Lesieur, L. Miolane, M. Lelarge, F. Krzakala, and L. Zdeborová, "Statistical and computational phase transitions in spiked tensor estimation," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 511–515.

[11] S. B. Hopkins, J. Shi, and D. Steurer, "Tensor principal component analysis via sum-of-square proofs," in *Conference on Learning Theory (COLT)*, 2015, pp. 956–1006.

[12] G. B. Arous, R. Gheissari, A. Jagannath *et al.*, "Algorithmic thresholds for tensor pca," *Annals of Probability*, vol. 48, no. 4, pp. 2052–2087, 2020.

[13] G. Biroli, C. Cammarota, and F. Ricci-Tersenghi, "How to iron out rough landscapes and get optimal performances: averaged gradient descent and its application to tensor pca," *Journal of Physics A: Mathematical and Theoretical*, vol. 53, no. 17, p. 174003, 2020.

[14] L. Zdeborová and F. Krzakala, "Statistical physics of inference: Thresholds and algorithms," *Advances in Physics*, vol. 65, no. 5, pp. 453–552, 2016.

[15] G. Reeves, J. Xu, and I. Zadik, "The all-or-nothing phenomenon in sparse linear regression," in *Conference on Learning Theory*, 2019, pp. 2652–2663.

[16] J. Barbier, N. Macris, and C. Rush, "All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[17] L. Corinzia, P. Penna, L. Mondada, and J. M. Buhmann, "Exact recovery for a family of community-detection generative models," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 415–419.

[18] L. Corinzia, P. Penna, W. Szpankowski, and J. M. Buhmann, "Statistical and computational thresholds for the planted k-densest sub-hypergraph problem," *arXiv preprint arXiv:2011.11500*, 2020.

[19] D. Gamarnik and I. Zadik, "High dimensional linear regression with binary coefficients: Mean squared error and a phase transition," in *Conference on Learning Theory (COLT)*, 2017.

[20] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[21] S. Verdú *et al.*, "Generalizing the fano inequality," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1247–1251, 1994.

[22] J. Banks, C. Moore, R. Vershynin, N. Verzelen, and J. Xu, "Information-theoretic bounds and phase transitions in clustering, sparse pca, and submatrix localization," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 4872–4894, 2018.