Passed & Spurious: Descent Algorithms and Local Minima in Spiked Matrix-Tensor Models

Stefano Sarao Mannelli [1] **Florent Krzakala** [2] **Pierfrancesco Urbani** [1] **Lenka Zdeborová** [1]

Abstract

In this work we analyse quantitatively the interplay between the loss landscape and performance of descent algorithms in a prototypical inference problem, the spiked matrix-tensor model. We study a loss function that is the negative log-likelihood of the model. We analyse the number of local minima at a fixed distance from the signal/spike with the Kac-Rice formula, and locate trivialization of the landscape at large signal-to-noise ratios. We evaluate in a closed form the performance of a gradient flow algorithm using integro-differential PDEs as developed in physics of disordered systems for the Langevin dynamics. We analyze the performance of an approximate message passing algorithm estimating the maximum likelihood configuration via its state evolution. We conclude by comparing the above results: while we observe a drastic slow down of the gradient flow dynamics even in the region where the landscape is trivial, both the analyzed algorithms are shown to perform well even in the part of the region of parameters where spurious local minima are present.

1. Introduction

A central question in computational sciences is the algorithmic feasibility of optimization in high-dimensional non-convex landscapes. This question is particularly important in learning and inference problems where the value of the optimized function is not the ultimate criterium for quality of the result, instead the generalization error or the closeness to a ground-truth signal is more relevant.

[1]Institut de physique théorique, Université Paris Saclay, CNRS, CEA, 91191 Gif-sur-Yvette, France [2]Laboratoire de Physique de l'École Normale Supérieure, CNRS & Université Pierre & Marie Curie & PSL Université, 75005 Paris, France. Correspondence to: Stefano Sarao Mannelli <stefano.sarao@ipht.fr>, Lenka Zdeborová <lenka.zdeborova@ipht.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Recent years brought a popular line of research into this question where various works show for a variety of systems that spurious local minima are not present in certain regimes of parameters and conclude that consequently optimization algorithms shall succeed, without the aim of being exhaustive these include (Kawaguchi, 2016; Soudry & Carmon, 2016; Ge et al., 2016; Freeman & Bruna, 2016; Bhojanapalli et al., 2016; Park et al., 2017; Du et al., 2017; Ge et al., 2017; Du et al., 2017; Lu & Kawaguchi, 2017). The *spuriosity* of a minima is in some works defined by their distance from the global minimum, in other works as local minimizers that lead to bad generalization or bad accuracy in reconstruction of the ground truth signal. These two notions are not always equivalent, and certainly the later is more relevant and will be used in the present work.

Many of the existing works stop at the statement that absence of spurious local minimizers leads to algorithmic feasibility and the presence of such spurious local minima leads to algorithmic difficulty, at least as far as gradient-descent-based algorithms are concerned. At the same time, even gradient-descent-based algorithms may be able to perform well even when spurious local minima are present. This is because the basins of attraction of the spurious minimas may be small and the dynamics might be able to avoid them. In the other direction, even if spurious local minima are absent, algorithms might take long time to find a minimizer for entropic reasons that in high-dimensional problems may play a crucial role.

Main Results: We study the spiked matrix-tensor model, introduced and motivated in (Sarao Mannelli et al., 2018a). We view this model as a prototypical solvable example of a high-dimensional non-convex optimization problem, and anticipate that the results observed here will have a broader relevance. Our main contributions are:

- Using the Kac-Rice formula (Fyodorov, 2004; Ben Arous et al., 2017) we rigorously derive the expected number of local minimizers of the associated likelihood at a given correlation with the ground truth signal.
- We extend the recently introduced *Langevin-state-evolution* (Sarao Mannelli et al., 2018a) to the gradient

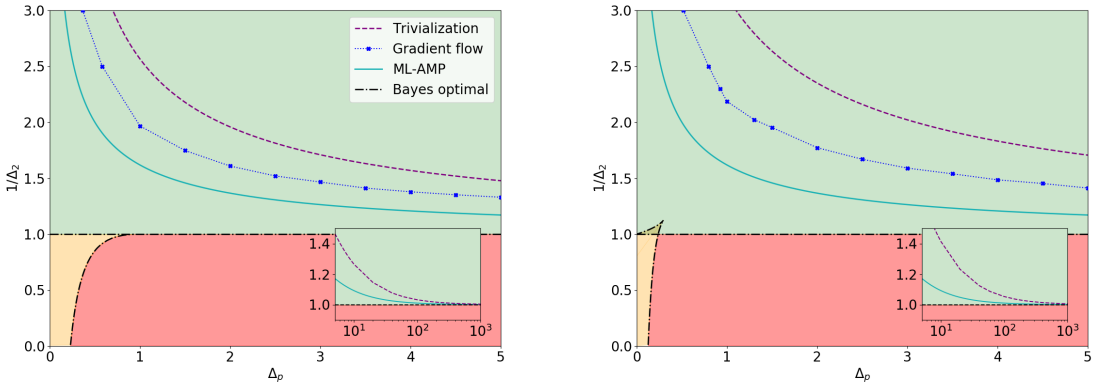


Figure 1. The figure summarizes the main results of this paper for the spiked matrix-tensor model with $p = 3$ (left) and $p = 4$ (right). As a function of the tensor-noise parameter Δ_p on the x-axes, we plot the values of $1/\Delta_2$ above which the following happens (from above): Above $\Delta_2^{\rm triv}$ (the dashed purple line) the landscape of the problem becomes trivial in the sense that all spurious local minima disappear. Above $\Delta_2^{\rm GF}$ (the dotted blue line) and $\Delta_2^{\rm ML-AMP}$ (the full cyan line), Eq. (32), the gradient flow and the ML-AMP algorithm, respectively, converge close to the ground truth signal in time linear in the input size. While the results for Kac-Rice and ML-AMP are given in a closed form, the ones for GF are obtained by extrapolating a convergence time obtained by numerical solution of integro-differential equations that describe large size behaviour of the GF. We note that all the three lines $\Delta_2^{\rm triv}$, $\Delta_2^{\rm GF}$, and $\Delta_2^{\rm ML-AMP}$ converge to 1 as $\Delta_p \to \infty$, consistently with the spiked matrix model. These three lines, related to minimization of the landscape, and their mutual positions, are the main result of this paper. The colors in the background, separated by the black dashed-dotted lines, show for comparison the phase diagram for the Bayes-optimal inference, related to the ability to approximate the marginals of the corresponding posterior probability distribution, and are taken from (Sarao Mannelli et al., 2018a). In the red region obtaining a positive correlation with the signal in information-theoretically impossible. In the green region it is possible to obtain optimal correlation with the signal using the Bayes-optimal AMP (BO-AMP). And in the orange the region the BO-AMP is not able to reach the Bayes-optimal performance. The insets show that in the limit $\Delta_p \to \infty$ all the described thresholds converge to the well-known BBP phase transition at $\Delta_2^{\rm BPP} = 1$ (Baik et al., 2005).

flow (GF) algorithm, obtaining a closed-form formula for the obtained accuracy in the limit of large system sizes. This formula is conjectured exact, and could likely be established by extending (Ben Arous et al., 2006).

- We derive and analyze the state evolution that rigorously describes the performance of the maximum-likelihood version of the approximate message passing algorithm (ML-AMP) for the present model.

We show that the above two algorithms (GF and ML-AMP) achieve the same error in the regime where they succeed. That same value of the error is also deduced from the position of all the minima strongly correlated with the signal as obtained from the Kac-Rice approach (precise statement below). We quantify the region of parameters in which the two above algorithms succeed and show that the ML-AMP is strictly better than GF. Remarkably, we show that the algorithmic performance is not driven by the absence of spurious local minima. These results are summarized in Fig. 1 and show that, in order to obtain a complete picture for settings beyond the present model, the precise interplay between absence of spurious local minima and algorithmic performance remains to be further investigated.

2. Problem Definition

In this paper we consider the spiked matrix-tensor model as studied in (Sarao Mannelli et al., 2018a). This is a statistical inference problem where the ground truth signal $x^* \in \mathbb{R}^N$ is sampled uniformly on the $N-1$-dimensional sphere, $\mathbb{S}^{N-1}(\sqrt{N})$. We then obtain two types of observations about the signal, a symmetric matrix Y, and an order p symmetric tensor T, that given the signal x^* are obtained as

$$Y_{ij} = \frac{x_i^* x_j^*}{\sqrt{N}} + \xi_{ij}, \tag{1}$$

$$T_{i_1,\dots,i_p} = \frac{\sqrt{(p-1)!}}{N^{(p-1)/2}} x_{i_1}^* \dots x_{i_p}^* + \xi_{i_1,\dots,i_p} \tag{2}$$

for $1 \le i < j \le N$ and $1 \le i_1 < \dots < i_p \le N$, using the symmetries to obtain the other non-diagonal components. Here ξ_{ij} and $\xi_{i_1,\dots,i_p}$ are for each $i < j$ and each $i_1 < \dots < i_p$ independent Gaussian random numbers of zero mean and variance Δ_2 and Δ_p, respectively.

The goal in this spiked matrix-tensor inference problem is to estimate the signal x^* from the knowledge of the matrix

Y and tensor T. If only the matrix was present, this model reduces to well known model of low-rank perturbation of a random symmetric matrix, closely related to the spiked covariance model (Johnstone, 2001). If on the contrary only the tensor is observed then the above model reduces to the spiked tensor model as introduced in (Richard & Montanari, 2014) and studied in a range of subsequent papers.

In this paper we study the matrix-tensor model where the two observations are combined. Our motivation is similar to the one exposed in (Sarao Mannelli et al., 2018a), that is, we aim to access a regime in which it is algorithmically tractable to obtain good performance with corresponding message passing algorithms yet it is challenging (e.g. leading to non-convex optimization) with sampling or gradient descent based algorithms, this happens when both $\Delta_2 = \Theta(1)$ and $\Delta_p = \Theta(1)$, while $N \to \infty$ (Sarao Mannelli et al., 2018a).

In this paper we focus on algorithms that aim to find the maximum likelihood estimator. The negative log-likelihood (Hamiltonian in physics, or loss function in machine learning) of the spiked matrix-tensor reads

$$\mathcal{L} = \sum_{i<j} \frac{1}{2\Delta_2} \left(Y_{ij} - \frac{x_i x_j}{\sqrt{N}} \right)^2 + \sum_{i_1<\cdots<i_p} \frac{1}{2\Delta_p} \left(T_{i_1 \dots i_p} - \frac{\sqrt{(p-1)!}}{N^{(p-1)/2}} x_{i_1} \dots x_{i_p} \right)^2 , \tag{3}$$

where $x \in \mathbb{S}^{N-1}(\sqrt{N})$ is constrained to the sphere.

In a high-dimensional, $N \to \infty$, noisy regime the maximum-likelihood estimator is not always optimal as it provides in general larger error than the Bayes-optimal estimator computing the marginals of the posterior, studied in (Sarao Mannelli et al., 2018a). At the same time the log-likelihood (3) can be seen as a loss function, that is non-convex and high-dimensional. The tractability and properties of such minimization problems are the most questioned in machine learning these days, and are worth detailed investigation in the present model.

3. Landscape Characterization

The first goal of this paper is to characterize the structure of local minima of the loss function (equivalently local maxima of the log-likelihood) eq. (3) as a function of the noise parameters Δ_2 and Δ_p. We compute the average number of local minimizers x having a given correlation with the ground truth signal $m = \lim_{N\to\infty} x \cdot x^*/N$. This leads to a so-called *complexity* function $\Sigma(m)$ defined as the logarithm of the expected number of local minima at correlation m with the ground truth.

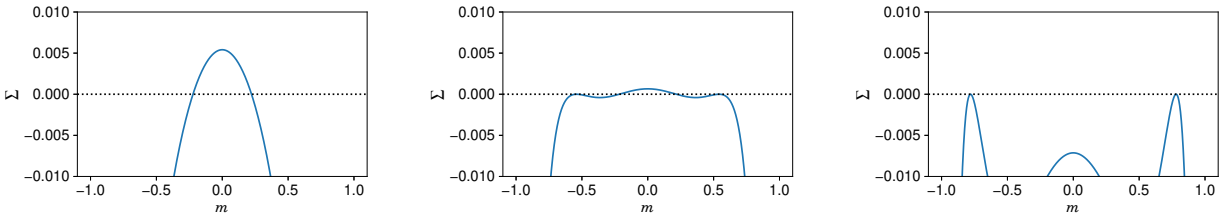
A typical example of this function, resulting from our analysis, is depicted in Fig. 2 for $p = 4$, $\Delta_p = 4.0$, and several values of Δ_2. We see from the figure that at large Δ_2 local minima appear only in a narrow range of values of m close to zero, as Δ_2 decreases the support of $\Sigma(m) \geq 0$ widens. At yet smaller values of Δ_2 the support $\Sigma(m) \geq 0$ becomes disconnected so that it is supported on an interval of value close to $m = 0$ and on two (one negative, one positive) isolated points. For yet smaller Δ_2 the complexity for values of m close to zero becomes negative, signalling what we call a *trivialization of the landscape*, where all remaining local minima are (in the leading order in N) as correlated with the ground truth as the global minima. The support of $\Sigma(m) \geq 0$ in the trivialized regions consists of two separated points. We call the value of Δ_2 at which the trivialization happens Δ_2^{triv}. In the phase diagram of Fig. 1 the trivialization of the energy landscape happens above the purple dashed line.

We use the Kac-Rice formula to determine the complexity $\Sigma(m)$ (Adler & Taylor, 2009; Fyodorov, 2015). Given an arbitrary continuous function, the Kac counting formula allows to compute the number of points where the function crosses a given value. The number of minima can be characterized using Kac's formula on the gradient of the loss (3), counting how many time the gradient crosses the zero value, under the condition of having a positive definite Hessian in order to count only local minima and not saddles. Since the spiked matrix-tensor model is characterized by a random landscape, due to the noise ξ_{ij} and $\xi_{i_1,\dots,i_p}$, we will consider the expected number of minima obtaining the Kac-Rice formula (Adler & Taylor, 2009; Fyodorov, 2015).

For mathematical convenience we will consider the rescaled configurations $\sigma = x/\sqrt{N} \in \mathbb{S}^{N-1}(1)$, and rescaled signal $\sigma^* = x^*/\sqrt{N}$. Call ϕ_{G,F_2,F_p} the joint probability density of the gradient G of the loss, and of the F_2 and F_p the contributions of the matrix and tensor to the loss, respectively. Given the value of the two contributions to the loss $F_2 = \epsilon_2 N$ and $F_p = \epsilon_p N$, and the correlation between the configuration and ground truth $m \in [-1, +1]$ that we impose using a Dirac's delta, the averaged number of minimizers is

$$\begin{aligned} \mathcal{N}(m, \epsilon_2, \epsilon_p; \Delta_2, \Delta_p) &= e^{\tilde{\Sigma}_{\Delta_2,\Delta_p}(m,\epsilon_2,\epsilon_p)} \\ &= \int_{\mathbb{S}^{N-1}} \mathbb{E}[\det H | G = 0, F_2 = N\epsilon_2, F_p = N\epsilon_p, H \succ 0] \\ &\times \phi_{G,F_2,F_p}(\sigma, 0, \epsilon_2, \epsilon_p)\, \delta(m - \sigma \cdot \sigma^*)\, \mathrm{d}\sigma \,. \end{aligned} \tag{4}$$

Rewrite the loss Eq. (3) neglecting terms that are constant with respect to the configuration and thus do not contribute



(a) $\Delta_2 = 2$ (b) $\Delta_2 = 2/3$ (c) $\Delta_2 = 2/5$

Figure 2. The complexity $\Sigma(m)$, Eq. (19), is shown for different values of parameter Δ_2 at fixed $\Delta_p = 4.0$ in the case $p = 4$. As Δ_2 is decreased (the signal to noise ratio increases) the complexity allows to identify three main scenarios in the topology of the loss landscape. In the first case (a) only a wide band of non-negative complexity around the point of zero correlation is present, in the second case (b) minima with non-trivial correlation with the signal appear but the band around $m = 0$ is still present, finally (c) the signal dominates over the noise and only minima with non-trivial correlation are present. The transition from case (b) to case (c), i.e. when the support of $\Sigma(m) \geq 0$ becomes two discontinuous points, as the bulk close to $m = 0$ becomes negative, is called the landscape trivialization. The Δ_2 at which this occurs is denoted Δ_2^{triv} and depicted in dashed purple in Fig. 1.

to the complexity

$$\begin{aligned}\hat{\mathcal{L}} = &\frac{\sqrt{N(p-1)!}}{\Delta_p} \sum_{i_1<\cdots<i_p} \xi_{i_1\dots i_p}\sigma_{i_1}\dots\sigma_{i_p} \\ &+ \frac{N(p-1)!}{\Delta_p} \sum_{i_1<\cdots<i_p} \sigma^*_{i_1}\sigma_{i_1}\dots\sigma^*_{i_p}\sigma_{i_p} \\ &+ \frac{\sqrt{N}}{\Delta_2}\sum_{i<j}\xi_{ij}\sigma_i\sigma_j + \frac{N}{\Delta_2}\sum_{i<j}\sigma^*_i\sigma_i\sigma^*_j\sigma_j\,.\end{aligned} \tag{5}$$

In the following we will use small letters f_2, f_p, g, h to characterize losses, gradient and Hessian constrained on the sphere and capital letters for the same quantities unconstrained. Define $\mathbb{I}_d$ the d-dimensional identity matrix. The following lemma characterizes ϕ_{G,F_2,F_p}.

Lemma 1. *Given the loss function Eq. (5) and a configuration x such that the correlation and the signal is m, then there exists a reference frame such that the joint probability distribution of $f_2, f_p \in \mathbb{R}$, $g \in \mathbb{R}^{N-1}$ and $h \in \mathbb{R}^{(N-1)\times(N-1)}$ is given by*

$$\frac{f_k}{N} \sim \frac{1}{k\Delta_k}m^k + \frac{1}{\sqrt{k\Delta_k}}\frac{1}{\sqrt{N}}Z_k\,; \tag{6}$$

$$\begin{aligned}\frac{g}{N} \sim &\left(\frac{1}{\Delta_p}m^{p-1} + \frac{1}{\Delta_2}m\right)\sqrt{1-m^2}\mathbb{e}_1 \\ &- \sqrt{\frac{1}{\Delta_p} + \frac{1}{\Delta_2}}\frac{1}{\sqrt{N}}\tilde{\boldsymbol{Z}}\,;\end{aligned} \tag{7}$$

$$\begin{aligned}\frac{h}{N} \sim &\left(\frac{p-1}{\Delta_p}m^{p-2} + \frac{1}{\Delta_2}\right)(1-m^2)\mathbb{e}_1\mathbb{e}_1^T \\ &+ \sqrt{\frac{p-1}{\Delta_p} + \frac{1}{\Delta_2}}\sqrt{\frac{N-1}{N}}\mathbb{W} - (pf_p + 2f_2)\mathbb{I}_{N-1}\,;\end{aligned} \tag{8}$$

with Z_k standard Gaussians and $k \in \{2, p\}$, $\tilde{\boldsymbol{Z}} \sim \mathcal{N}(0, \mathbb{I}_{N-1})$ a standard multivariate Gaussian and $\mathbb{W} \sim$ GOE$(N-1)$ a random matrix from the Gaussian orthogonal ensemble.

Proof sketch. Starting from Eq. (5), split the contributions of the matrix and tensor in F_2 and F_p, two Gaussian variables and impose the spherical constrain with a Lagrange multiplier μ.

$$f_2(\sigma) + f_p(\sigma) = F_2(\sigma) + F_p(\sigma) - \frac{\mu}{2}\Big(\sum_i \sigma_i^2 - 1\Big)\,, \tag{9}$$

$$g_i(\sigma) = G_i(\sigma) - \mu\sigma_i\,, \tag{10}$$

$$h_{ij}(\sigma) = H_{ij}(\sigma) - \mu\,. \tag{11}$$

The expression for μ in a critical point can be derived as follows. Given $g_i(\sigma) \equiv 0$, multiply Eq. (10) by σ_i, sum over the indices and obtain: $\mu = \sum_i G_i(\sigma)\sigma_i = 2f_2(\sigma) + pf_p(\sigma)$. We now restrict our study to the unconstrained random variables and substitute μ. Since the quantities f_2, f_p, g, h, μ are linear functionals of Gaussians they will be distributed as Gaussian random variables and therefore can be characterized by computing expected values and covariances. Starting from the losses coming from the matrix and the tensor in Eq. (5), $F_2(\sigma)$ and $F_p(\sigma)$, respectively, consider the moments with respect to the realization of the noise, $\xi_{i_1\dots i_p}$, ξ_{ij}. For $k \in \{2, p\}$ the first moment leads to

$$\mathbb{E}[F_k(\sigma)] = \frac{N}{k\Delta_k}(\sigma\cdot\sigma^*)^k + O(1)\,. \tag{12}$$

Let's consider the second moment but having two different configurations σ and τ,

$$\mathbb{E}\left[F_k(\sigma)F_k(\tau)\right] = \frac{N}{k\Delta_k}(\sigma\cdot\tau)^k + O(1)\,. \tag{13}$$

Using standard results for derivatives of Gaussians (see e.g. (Adler & Taylor, 2009) Eq. 5.5.4) we can obtain means and covariances of the random variables taking derivatives with respect to σ and τ. Then set $\tau = \sigma$, imposing the spherical constrain and using $\sigma \cdot \sigma^* = m$.

The last step is the definition of a convenient reference frame $\{\mathbb{e}_j\}_{j=1,\dots,N}$. Align the configuration along the last coordinate $\mathbb{e}_N = \sigma$ and the signal with a combination of the first and last coordinates $\sigma^* = \sqrt{1-m^2}\mathbb{e}_1 + m\mathbb{e}_N$. Finally, project on the sphere by discarding the last coordinate. □

We can now rewrite the determinant of the conditioned Hessian by grouping the multiplicative factor in front of the GOE in Eq. (8)

$$\det h = \left(\frac{p-1}{\Delta_p} + \frac{1}{\Delta_2}\right)^{\frac{N-1}{2}} \left(\frac{N}{N-1}\right)^{-\frac{N-1}{2}} \times \det\left[\mathbb{W} - t_N \mathbb{I}_{N-1} + \theta_N \mathbb{e}_1 \mathbb{e}_1^T\right] \tag{14}$$

with t_N and θ_N given by

$$t_N \to t = 2\frac{p\epsilon_p + 2\epsilon_2}{\sqrt{\frac{p-1}{\Delta_p} + \frac{1}{\Delta_2}}}, \tag{15}$$

$$\theta_N \to \theta = \frac{\frac{p-1}{\Delta_p} m^{p-2} + \frac{1}{\Delta_2}}{\sqrt{\frac{p-1}{\Delta_p} + \frac{1}{\Delta_2}}}(1-m^2) \tag{16}$$

in the large N-limit. Therefore the Hessian behaves like a GOE shifted by t with a rank one perturbation of strength θ. This exact same problem has already been studied in (Ben Arous et al., 2017) and we can thus deduce the expression for the complexity as

$$\begin{aligned}\tilde{\Sigma}_{\Delta_2,\Delta_p}(m,\epsilon_2,\epsilon_p) = {} & \frac{1}{2}\log\frac{\frac{p-1}{\Delta_p}+\frac{1}{\Delta_2}}{\frac{1}{\Delta_p}+\frac{1}{\Delta_2}} + \frac{1}{2}\log(1-m^2) \\ & - \frac{1}{2}\frac{\left(\frac{m^{p-1}}{\Delta_p}+\frac{m}{\Delta_2}\right)^2}{\frac{1}{\Delta_p}+\frac{1}{\Delta_2}}(1-m^2) - 2p\Delta_p\left(\epsilon_p - \frac{m^p}{2p\Delta_p}\right)^2 \\ & - 4\Delta_2\left(\epsilon_2 - \frac{m^2}{4\Delta_2}\right)^2 + \Phi(t) - L(\theta,t),\end{aligned} \tag{17}$$

with

$$\Phi(t) = \frac{t^2}{4} + \mathbb{1}_{|t|>2}\left[\log\left(\sqrt{\frac{t^2}{4}-1} + \frac{|t|}{2}\right) - \frac{|t|}{4}\sqrt{t^2-4}\right]$$

$$L(\theta,t) = \begin{cases} \frac{1}{4}\int_{\theta+\frac{1}{\theta}}^{t}\sqrt{y^2-4}dy - \frac{\theta}{2}\left(t - \left(\theta+\frac{1}{\theta}\right)\right) \\ \quad + \frac{t^2 - \left(\theta+\frac{1}{\theta}\right)^2}{8} \quad \theta > 1,\ 2 \le t < \frac{\theta^2+1}{\theta} \\ \infty \qquad t < 2 \\ 0 \qquad \text{otherwise.} \end{cases}$$

We note at this point that for the case of the pure spiked tensor model $\Delta_2 \to \infty$ the above expression reduces exactly to the complexity derived in (Ben Arous et al., 2017). The following theorem states that to the leading order Eq. (17) represents the complexity of our problem.

Theorem 1. *Given Δ_2 and Δ_p, for any $(\epsilon_2, \epsilon_p) \in \mathbb{R}^2$ and $m \in [-1,+1]$ it holds*

$$\lim_{N\to\infty}\frac{1}{N}\log \mathbb{E}\mathcal{N}(m,\epsilon_2,\epsilon_p;\Delta_2,\Delta_p) = = \tilde{\Sigma}_{\Delta_2,\Delta_p}(m,\epsilon_p,\epsilon_2) \tag{18}$$

Proof sketch. The proof comes immediately from (Ben Arous et al., 2017) Thm. 2, see also Sec. 4.1. □

The quantity that we are interested in is the projection of Eq. (17) to the maximizing values of ϵ_2 and ϵ_p:

$$\Sigma(m) = \max_{\epsilon_2,\,\epsilon_p} \tilde{\Sigma}_{\Delta_2,\Delta_p}(m,\epsilon_2,\epsilon_p). \tag{19}$$

Eq. (19) allows to understand if at a given correlation with the signal, there are regions with an exponential expected number of minima, see Fig. 2. Thus it allows to locate parameters where the landscapes is trivial.

We computed the expected number of minima, i.e. the so-called annealed average. The annealed average might be dominated by rare samples, and in general provides only an upper bound for typical samples. The quenched complexity, i.e. the average of the logarithm of the number of minima, is more involved. The quenched calculation was done in the case of a the spiked tensor model (Ros et al., 2019). It is interesting to notice that in (Ros et al., 2019) the authors found that the annealed complexity does not differ from the quenched complexity for $m = 0$. This combined with analogous preliminary results for the spiked matrix-tensor model, suggest that considering the quenched complexity would not change the conclusions of this paper presented in the phase diagrams Fig. 1.

4. Gradient Flow Analysis

In this section we analyze the performance of the gradient flow descent in the loss function (3)

$$\frac{d}{dt}x_i(t) = -\mu(t)x_i(t) - \frac{\delta\mathcal{L}}{\delta x_i}(t)\,, \tag{20}$$

where the Lagrange parameter $\mu(t)$ is set in a way to ensure the spherical constraint $x \in \mathbb{S}^{N-1}(\sqrt{N})$. Our aim is to understand the final correlation between the ground truth signal and the configuration reached by the gradient flow in large but finite time, while $N \to \infty$.

The gradient flow (20) can be seen as a zero-temperature limit of the Langevin algorithm where

$$\frac{d}{dt}x_i(t) = -\mu(t)x_i(t) - \frac{\delta \mathcal{L}}{\delta x_i}(t) - \eta_i(t)\,, \qquad (21)$$

with $\eta_i(t)$ being the Langevin noise with zero mean and covariance $\langle \eta_i(t)\eta_j(t')\rangle = 2T\delta_{ij}\delta(t-t')$, where T has the physical meaning of temperature, the notation $\langle \dots \rangle$ stands for the average over the noises ξ_{ij} and $\xi_{i_1,\dots,i_p}$. As we take the limit $T \to 0$, the noise becomes peaked around zero, effectively recovering the gradient flow.

The performance of the Langevin algorithm was characterized recently in (Sarao Mannelli et al., 2018a) using equations developed in physics of disordered systems (Crisanti et al., 1993; Cugliandolo & Kurchan, 1993). In (Sarao Mannelli et al., 2018a) this characterization was given for an arbitrary temperature T and compared to the landscape of the Bayes-optimal estimator (Antenucci et al., 2018). Here we hence summarize and use the results of (Sarao Mannelli et al., 2018a) corresponding to the limit $T \to 0$.

The Langevin dynamics with generic temperature is in the large size limit, $N \to \infty$, characterized by a set of PDEs for the self-correlation $C(t,t') = \lim_{N\to\infty}\left\langle \frac{1}{N}\sum x_i(t)x_i(t')\right\rangle$, the response function $R(t,t') = \lim_{N\to\infty}\left\langle \frac{1}{N}\sum \frac{\delta x_i(t)}{\delta \eta_i(t')}\right\rangle$, and the correlation with the signal $m(t) = \lim_{N\to\infty}\left\langle \frac{1}{N}\sum x_i(t)x_i^*\right\rangle$. Ref. (Sarao Mannelli et al., 2018a) established that as the gradient flow evolves these quantities satisfy eqs. (74)-(76) in that paper. Taking the zero-temperature limit in those equations we obtain

$$\begin{aligned}\frac{\partial}{\partial t}C(t,t') &= -\tilde{\mu}(t)C(t,t') + Q'(m(t))m(t') \\ &+ \int_0^t dt'' R(t,t'')Q''(C(t,t''))C(t',t'') \\ &+ \int_0^{t'} dt'' R(t',t'')Q'(C(t,t''))\,,\end{aligned} \qquad (22)$$

$$\begin{aligned}\frac{\partial}{\partial t}R(t,t') &= -\tilde{\mu}(t)R(t,t') \\ &+ \int_{t'}^t dt'' R(t,t'')Q''(C(t,t''))R(t'',t')\,,\end{aligned} \qquad (23)$$

$$\begin{aligned}\frac{\partial}{\partial t}m(t) &= -\tilde{\mu}(t)m(t) + Q'(m(t)) \\ &+ \int_0^t dt'' R(t,t'')m(t'')Q''(C(t,t''))\,,\end{aligned} \qquad (24)$$

with $Q(f) = f^p/(p\Delta_p) + f^2/(2\Delta_2)$ and $\tilde{\mu}(t) = \lim_{T\to 0} T\mu(t)$ the rescaled spherical constraint. Boundary conditions for the equations are $C(t,t) = 1\ \forall t$, $R(t,t') = 0$ for all $t < t'$ and $\lim_{t'\to t^-} R(t,t') = 1\ \forall t$. An additional equation for $\tilde{\mu}(t)$ is obtained by fixing $C(t,t) = 1$ in Eq. (22). In the context of disordered systems those equations have been established rigorously for a related case of the matrix-tensor model without the spike (Ben Arous et al., 2006).

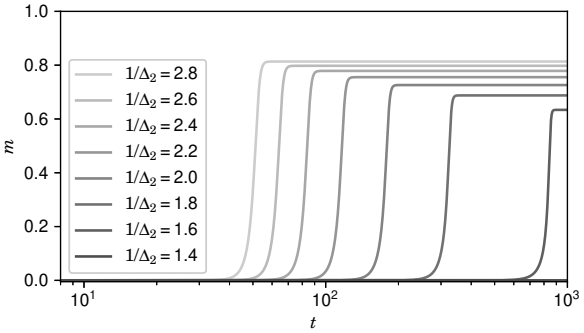


Figure 3. Eq. (24) characterizes the evolution of the correlation of the gradient flow with the ground truth signal, evaluated for several values of Δ_2, at $\Delta_p = 4.0$ starting from $m(0) = 10^{-10}$. The dynamics displays a fast increase of the convergence time as Δ_2 increases. At large times, the plateau we observe has the same value of correlation m as the minima best correlated with the signal, as predicted via Kac-Rice approach.

Eqs. (22-24) are integrated numerically showing the large-size-limit performance of the gradient flow algorithm. Example of this evolution is given in Fig. 3 for $p = 3$, $\Delta_p = 4$. The code is available online (Sarao Mannelli et al., 2018b) and linked to this paper. For consistency we confirm numerically that at large times the gradient flow reaches values of the correlation that correspond exactly to the value of the correlation of the minima correlated to the signal as obtained in the Kac-Rice approach.

As the variance Δ_2 increases the time it takes to the gradient flow to acquire good correlation with the signal increases. We define the *convergence* time t_c as the time it takes to reach 1/2 of the final plateau. The dependence of t_c on Δ_2 is consistent with a power law divergence at $\Delta_2^{\rm GF}$. This is illustrated in Fig. 4 where we plot the convergence time as a function of Δ_2 and show the power-law fit in the inset. The points $\Delta_2^{\rm GF}$ are collected and plotted in Fig. 1, dotted blue line.

From Fig. 4 we see that the gradient flow algorithm undergoes a considerable slow-down even in the region where the landscape is trivial, i.e. does not have spurious local minimizers. At the same time divergence of the convergence

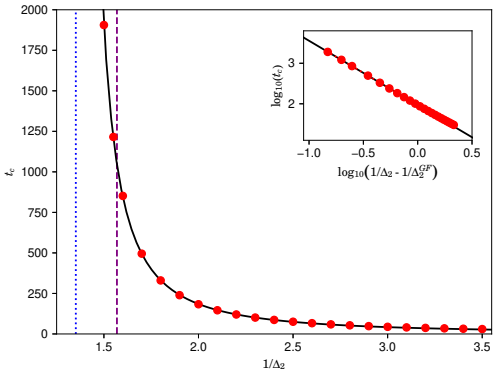


Figure 4. The convergence time the gradient flow takes to find a configuration well correlated with the signal for $\Delta_p = 4.0$, $p = 3$ as a function of Δ_2, starting from $m(0) = 10^{-10}$. The points are fitted with a power law consistent with a divergence point $1/\Delta_2^{\rm GF} = 1.35$ (vertical dotted line, log-log scale of the fit shown in the inset) while landscape trivialization occurs at $1/\Delta_2^{\rm triv} = 1.57$ (vertical dashed line).

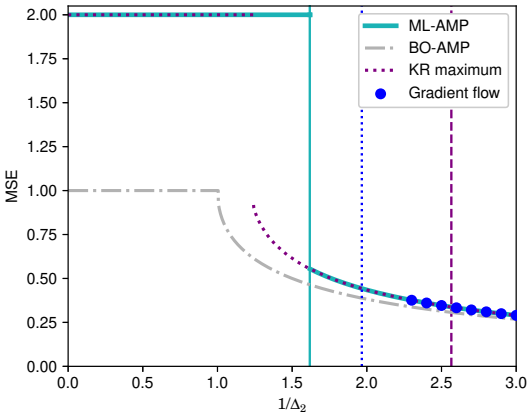


Figure 5. We show the mean-squared error (MSE) as achieved by the analyzed algorithms, for $p = 3$, $\Delta_p = 1.0$ as a function of the signal-to-noise (snr) ratio $1/\Delta_2$. The full cyan line corresponds to the error reached by the ML-AMP algorithm, it jumps discontinuously at $1/\Delta_2^{\rm ML-AMP} = 1.62$. The blue points is the error reached by the gradient flow in time $t < 1000$. The divergence of the convergence time is extrapolated to occur at $1/\Delta_2^{\rm GF} = 1.97$, blue dotted vertical line. The purple dotted line represents the maximum having the largest m of the complexity function $\Sigma(m)$, Eq. (19). The vertical purple dashed line at $1/\Delta_2^{\rm triv} = 2.57$ corresponds to the trivialization of the landscape, beyond which only local minima well correlated with the signal remain. We note that all these approaches agree on the value of the MSE. For the sake of comparison we show (the dashed-dotted grey line) also the minimal-MSE achieved in the Bayes-optimal setting.

time happens only well inside the phase where spurious local minimizers do exist.

5. Maximum-Likelihood Approximate Message Passing

Approximate Message Passing (AMP) is a popular iterative algorithm (Donoho et al., 2009) with a key advantage of being analyzable via a set of equations, called state evolution equations, that have been proved rigorously to follow the average evolution of the algorithm (Javanmard & Montanari, 2013). The maximum-likehood AMP (ML-AMP) algorithm studied in this paper is a generalization of AMP for the pure spiked tensor model from (Richard & Montanari, 2014) to the spiked matrix-tensor model. We will show that its fixed points correspond to stationary points of the loss function (3). This should be contrasted with the Bayes-optimal AMP (BO-AMO) that was studied in (Sarao Mannelli et al., 2018a) and aims to approximate the marginals of the corresponding posterior probability distribution. The ML-AMP instead aims to estimate the maximum-likelihood solution, $\hat{x}$. In information theory the BO-AMP would correspond to the sum-product algorithm, while the present one to the max-sum algorithm. In statistical physics language the BO-AMP corresponds to temperature one, while the present one to zero temperature. In the supporting information we provide a schematic derivation of the ML-AMP as a zero-temperature limit of the BO-AMP, using a scheme similar to (Lesieur et al., 2017).

The ML-AMP algorithm reads

$$B_i^t = \frac{\sqrt{(p-1)!}}{N^{(p-1)/2}} \sum_{k_2<\cdots<k_p} \frac{T_{ik_2\ldots k_p}}{\Delta_p} \hat{x}_{k_2}^t \ldots \hat{x}_{k_p}^t + \frac{1}{\sqrt{N}} \sum_k \frac{Y_{ik}}{\Delta_2} \hat{x}_k^t - \mathrm{r}_t \hat{x}_i^{t-1} , \tag{25}$$

$$\hat{x}_i^{t+1} = \frac{B_i^t}{\frac{1}{\sqrt{N}}||B^t||_2} , \tag{26}$$

$$\hat{\sigma}^{t+1} = \frac{1}{\frac{1}{\sqrt{N}}||B^t||_2} \tag{27}$$

with $||\cdots||_2^2$ the ℓ_2-norm and r_t the Onsager reaction term

$$\mathrm{r}_t = \frac{1}{\Delta_2}\frac{1}{N}\sum_k \hat{\sigma}_k^t + \frac{p-1}{\Delta_p}\frac{1}{N}\sum_k \hat{\sigma}_k^t \left(\frac{1}{N}\sum_k \hat{x}_k^t \hat{x}_k^{t-1}\right)^{p-2} . \tag{28}$$

5.1. ML-AMP & Stationary Points of the Loss

Using an argument similar to Prop. 5.1 in (Montanari, 2012) we can show that a fixed points found by ML-AMP corresponds to finding a stationary point of the loss Eq. (3) with a ridge regularizer.

Property 1. *Given $(\hat{x}^*, \sigma^*)$ a fixed point of ML-AMP, then $\hat{x}^*$ satisfies the stationary condition of the loss.*

Proof sketch. Let us denote B^*, r^* the fixed point of Eqs. (25) and (28). From Eq. (26) and Eq. (25) we have

$$\left(\frac{1}{\sqrt{N}}||B^*||_2 + \mathrm{r}^*\right) x^* = \frac{1}{\sqrt{N}} \sum_k \frac{Y_{ik}}{\Delta_2} \hat{x}_i^* + \frac{\sqrt{(p-1)!}}{N^{(p-1)/2}} \sum_{k_2<\cdots<k_p} \frac{T_{ik_2\ldots k_p}}{\Delta_p} \hat{x}_{k_2}^* \ldots \hat{x}_{k_p}^* \tag{29}$$

which is exactly solution of the derivative of Eq. (3) with respect to x_i when the spherical constraint is enforced by a Lagrange multiplier μ

$$0 = -\mu x_i + \frac{1}{\sqrt{N}} \sum_k \frac{Y_{ik}}{\Delta_2} x_i + \frac{\sqrt{(p-1)!}}{N^{(p-1)/2}} \sum_{k_2<\cdots<k_p} \frac{T_{ik_2\ldots k_p}}{\Delta_p} x_{k_2} \ldots x_{k_p} \,.$$

Moreover ML-AMP by construction preserves the spherical constrain at every time iteration. □

5.2. State Evolution

The evolution of ML-AMP can be tracked through a set of equations called state evolution (SE). The state evolution can be characterized via an order parameter: $m^t = \frac{1}{N}\sum_i \hat{x}_i^t x_i^*$, the correlation of the ML-AMP-estimator with the ground truth signal at time t. According to the SE, as derived in the supporting information, and proven for a general class of models in (Javanmard & Montanari, 2013), this parameter evolves in the large N limit as

$$m^{t+1} = \frac{\frac{m^t}{\Delta_2} + \frac{(m^t)^{p-1}}{\Delta_p}}{\sqrt{\frac{1}{\Delta_2} + \frac{1}{\Delta_p} + \left(\frac{m^t}{\Delta_2} + \frac{(m^t)^{p-1}}{\Delta_p}\right)^2}}\,, \tag{30}$$

and the mean square error correspondingly

$$\mathrm{MSE}^t = 2(1 - m^t). \tag{31}$$

Analysis of the simple scalar SE, Eq. (30), allows to identify the error reached by the ML-AMP algorithm. We first observe that $m = 0$ is always a fixed point. For the performance of ML-AMP is the stability of this fixed point that determines whether the ML-AMP will be able to find a positive correlation with the signal or not. Analyzing Eq. (30) we obtain that the $m = 0$ is a stable fixed point for $\Delta_2 > \Delta_2^{\mathrm{ML-AMP}}$ where

$$\Delta_2^{\mathrm{ML-AMP}}(\Delta_p) = \frac{-\Delta_p + \sqrt{\Delta_p^2 + 4\Delta_p}}{2}\,. \tag{32}$$

Consequently for $\Delta_2 > \Delta_2^{\mathrm{ML-AMP}}$ the ML-AMP algorithm converges to $m = 0$, i.e. zero correlation with the signal. The line $\Delta_2^{\mathrm{ML-AMP}}$ is the line plotted in Fig. 1. For $p = 3$ and $p = 4$, we obtain that for $\Delta_2 < \Delta_2^{\mathrm{ML-AMP}}$ the ML-AMP algorithm converges to a positive $m^* > 0$ correlation with the signal, depicted in Fig. 5. In Fig. 5 we also observe that this correlation agrees with the position of the maximum having largest value of m in the complexity function $\Sigma(m)$. The trivialization of the landscape occurs at $\Delta_2^{\mathrm{triv}} < \Delta_2^{\mathrm{ML-AMP}}$, thus showing that for $\Delta_2^{\mathrm{triv}} < \Delta < \Delta_2^{\mathrm{ML-AMP}}$ the ML-AMP algorithm is able to ignore a good portion of the spurious local minima and to converge to the local minima best correlated with the signal.

In Fig. 5 we also compared to the MSE obtained by the Bayes-optimal AMP that provably minimizes the MSE in the case depicted in the figure (Sarao Mannelli et al., 2018a). We see that the gap between the Bayes-optimal error and the one reached by the loss minimization approaches goes rapidly to zero as Δ_2 decreases.

6. Discussion

We analyzed the behavior of two algorithms for optimizing a rough high-dimensional loss landscape of the spiked matrix-tensor model. We used the Kac-Rice formula to count the average number of minima of the loss function having a given correlation with the signal. Analyzing the resulting formula we defined and located where the energy landscape becomes trivial in the sence that spurious local minima disappear. We analyzed the performance of gradient flow via integro-differential state-evolution-like equations. We delimited a region of parameters for which the gradient flow is able to avoid the spurious minima and obtain a good correlation with the signal in time linear in the input size. We also analyzed the maximum-likelihood AMP algorithm, located the region of parameters in which this algorithm works, which is larger than the region for which the gradient flow works. The relation between existence or absence of spurious local minima in the loss landscapes of a generic optimization problems and the actual performance of optimization algorithm is yet to be understood. Our analysis of the spiked matrix-tensor model brings a case-study where we were able to specify this relation quantitatively.

Acknowledgments

We thank G. Ben Arous, G. Biroli, C. Cammarota, G. Folena, and V. Ros for precious discussions. We acknowledge funding from the ERC under the European Unions Horizon 2020 Research and Innovation Programme Grant Agreement 714608-SMiLe and 307087-SPARCS; from the French National Research Agency (ANR) grant PAIL; and from "Investissements dAvenir" LabEx PALM (ANR-10-LABX-0039-PALM) (SaMURai and StatPhysDisSys). The manuscript was finalized while some of the authors were visiting KITP thus they acknowledge partial support by the National Science Foundation under Grant No. PHY-1748958.

References

Adler, R. J. and Taylor, J. E. *Random fields and geometry*. Springer Science & Business Media, 2009.

Antenucci, F., Franz, S., Urbani, P., and Zdeborová, L. On the glassy nature of the hard phase in inference problems. *arXiv preprint arXiv:1805.05857*, 2018.

Baik, J., Arous Ben, G., Péché, S., et al. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5): 1643–1697, 2005.

Ben Arous, G., Dembo, A., and Guionnet, A. Cugliandolo-Kurchan equations for dynamics of spin-glasses. *Probability theory and related fields*, 136(4):619–660, 2006.

Ben Arous, G., Mei, S., Montanari, A., and Nica, M. The landscape of the spiked tensor model. *arXiv preprint arXiv:1711.05424*, 2017.

Bhojanapalli, S., Neyshabur, B., and Srebro, N. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pp. 3873–3881, 2016.

Crisanti, A., Horner, H., and Sommers, H.-J. The spherical p-spin interaction spin-glass model. *Zeitschrift für Physik B Condensed Matter*, 92(2):257–271, 1993.

Cugliandolo, L. F. and Kurchan, J. Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model. *Physical Review Letters*, 71(1):173, 1993.

Donoho, D. L., Maleki, A., and Montanari, A. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, Nov 2009.

Du, S. S., Lee, J. D., Tian, Y., Poczos, B., and Singh, A. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. *arXiv preprint arXiv:1712.00779*, 2017.

Freeman, C. D. and Bruna, J. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

Fyodorov, Y. V. Complexity of random energy landscapes, glass transition, and absolute value of the spectral determinant of random matrices. *Physical review letters*, 92 (24):240601, 2004.

Fyodorov, Y. V. High-dimensional random fields and random matrix theory. *Markov Processes Relat. Fields*, 21: 483–518, 2015.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pp. 2973–2981, 2016.

Ge, R., Jin, C., and Zheng, Y. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1233–1242, 2017. arXiv preprint arXiv:1704.00708.

Javanmard, A. and Montanari, A. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pp. 295–327, 2001.

Kawaguchi, K. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pp. 586–594, 2016.

Lesieur, T., Krzakala, F., and Zdeborová, L. Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7): 073403, 2017.

Lu, H. and Kawaguchi, K. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.

Montanari, A. Graphical models concepts in compressed sensing. *Compressed Sensing: Theory and Applications*, pp. 394–438, 2012.

Park, D., Kyrillidis, A., Carmanis, C., and Sanghavi, S. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Artificial Intelligence and Statistics*, pp. 65–74, 2017.

Ricci-Tersenghi, F., Semerjian, G., and Zdeborová, L. Typology of phase transitions in bayesian inference problems. preprint:arXiv:1806.11013.

Richard, E. and Montanari, A. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, pp. 2897–2905, 2014.

Ros, V., Ben Arous, G., Biroli, G., and Cammarota, C. Complex energy landscapes in spiked-tensor and simple glassy models: Ruggedness, arrangements of local minima, and phase transitions. *Physical Review X*, 9(1):011003, 2019.

Sarao Mannelli, S., Biroli, G., Cammarota, C., Krzakala, F., Urbani, P., and Zdeborová, L. Marvels and pitfalls of the langevin algorithm in noisy high-dimensional inference. *arXiv preprint arXiv:1812.09066*, 2018a.

Sarao Mannelli, S., Krzakala, F., Urbani, P., and Zdeborová, L. Gradient descent state evolution integrators, 2018b. Available at: https://github.com/sphinxteam/spiked_matrix-tensor_T0.

Soudry, D. and Carmon, Y. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.