



HAL
open science

Passed & Spurious: analysing descent algorithms and local minima in Spiked Matrix-Tensor Models

Stefano Sarao Mannelli, Florent Krzakala, Pierfrancesco Urbani, Lenka Zdeborová

► **To cite this version:**

Stefano Sarao Mannelli, Florent Krzakala, Pierfrancesco Urbani, Lenka Zdeborová. Passed & Spurious: analysing descent algorithms and local minima in Spiked Matrix-Tensor Models. Proceedings of Machine Learning Research, 2019, Proceedings of the 36 th International Conference on Machine Learning, pp.4333-4342. cea-02009727

HAL Id: cea-02009727

<https://cea.hal.science/cea-02009727>

Submitted on 6 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Passed & Spurious: analysing descent algorithms and local minima in spiked matrix-tensor model

Stefano Sarao Mannelli^a, Florent Krzakala^b,
Pierfrancesco Urbani^a, and Lenka Zdeborová^a

^aInstitut de physique théorique, Université Paris Saclay, CNRS, CEA, 91191 Gif-sur-Yvette, France

^bLaboratoire de Physique de l'Ecole normale supérieure, Université PSL, CNRS, Sorbonne Université, Université Paris-Diderot, Sorbonne Paris Cité, Paris, France

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.

Contents

1	Introduction	2
2	Problem definition	2
3	Landscape characterization	4
4	Gradient flow analysis	7
4.1	Performance of the gradient flow	9
5	Maximum-likelihood approximate message passing	9
5.1	ML-AMP & stationary points of the loss	10
5.2	State evolution	11
6	Discussion	12
A	Kac-Rice formula	14
A.1	p -odd cases	14
B	Gradient Flow	15
B.1	Dependence on the initial conditions	15

arXiv:1902.00139v1 [cs.LG] 1 Feb 2019

C AMP	15
C.1 From AMP to ML-AMP	15
C.2 State evolution	16
C.3 Derivation of spinodals	17

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.

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 [2, 3, 4, 5, 6, 7, 8, 9, 8, 10]. The *spuriousity* of a minima is in some works defined by their distance from the global minima, 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: In this work we provide a case-study of a high-dimensional inference problems – the spiked matrix-tensor model – for which we are able to describe and quantitatively compare the following:

- With the use of the Kac-Rice formula [11, 12] we compute the expected number of local minimizers of the associated likelihood at a given correlation with the ground truth signal.
- With the use of the recently introduced *Langevin-state-evolution* [1], that is a generalization of an approach well known in physics for analysis of the Langevin dynamics [13, 14], we give a closed-form description for the behaviour of the gradient flow (GF) algorithm in the limit of large system sizes.
- We analyze the state evolution of the maximum-likelihood version of the approximate message passing algorithm (ML-AMP).

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, up to the degree of accuracy of the extrapolation of the GF performance, the two lines are not the same. We also show that the algorithmic performance is not driven by the absence of spurious local minima. These results are summarized in Fig. 1. These result 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 [1]. 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

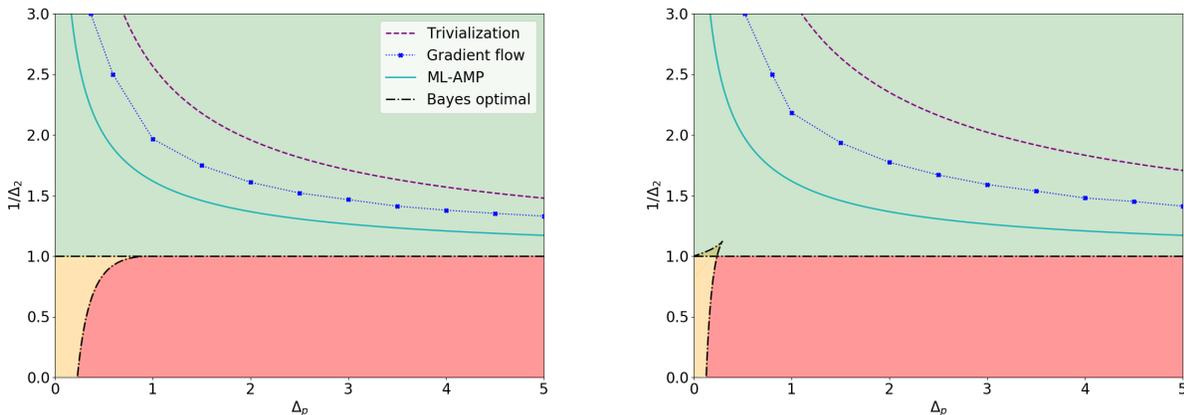


Figure 1: The figure summarizes the main results of this paper for the spiked matrix-tensor model with $p = 3$ (left) and $p = 4$ (right). As a function of the tensor-noise parameter Δ_p on the x-axes, we plot the values of $1/\Delta_2$ above which the following happens (from above): Above Δ_2^{triv} (the dashed purple line) the landscape of the problem becomes trivial in the sense that all spurious local minima disappear. Above Δ_2^{GF} (the dotted blue line) and $\Delta_2^{\text{ML-AMP}}$ (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 Δ_2^{triv} , Δ_2^{GF} , and $\Delta_2^{\text{ML-AMP}}$ converge to 1 as $\Delta_p \rightarrow \infty$, 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 [1]. 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.

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

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

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

for $1 \leq i < j \leq N$ and $1 \leq i_1 < \dots < i_p \leq N$, using symmetries to obtain the other non-diagonal components. Here ξ_{ij} and ξ_{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 [15]. If on the contrary only the tensor is observed then the above model reduces to the spiked tensor model as introduced in [16] and studies 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 [1], 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 $N \rightarrow \infty$ [1].

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

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

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

In a high-dimensional, $N \rightarrow \infty$, 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 [1]. 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

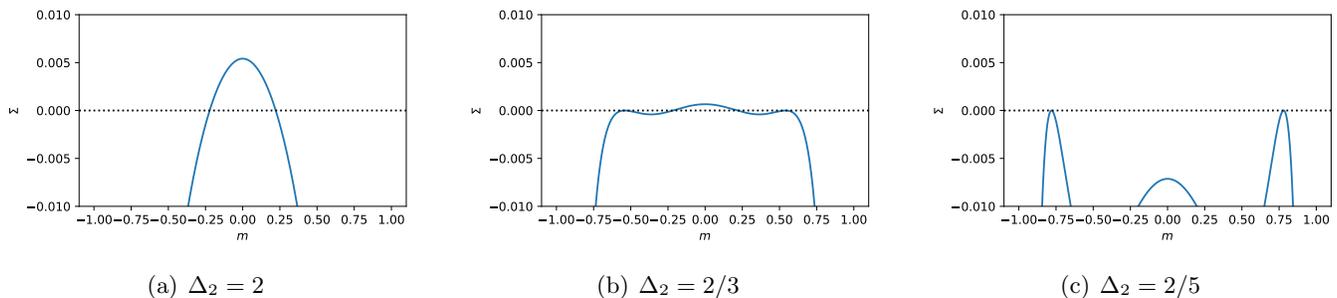


Figure 2: The complexity $\Sigma(m)$, Eq. (19), is shown for different values of parameter Δ_2 at fixed $\Delta_p = 4.0$ in the case $p = 4$. As Δ_2 is decreased (the signal to noise ratio increases) the complexity allows to identify three main scenarios in the topology of the loss landscape. In the 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.

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 $m = \lim_{N \rightarrow \infty} x \cdot x^* / N$. 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 $\Sigma(m)$ [17, 18]. 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 ξ_{i_1, \dots, i_p} , we will consider the expected number of minima obtaining the Kac-Rice formula [17, 18].

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

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

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

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

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 $\mathbb{1}_d$ 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*

$$\frac{f_k}{N} \sim \frac{1}{k\Delta_k} m^k + \frac{1}{\sqrt{k\Delta_k}} \frac{1}{\sqrt{N}} Z_k; \quad (6)$$

$$\frac{g}{N} \sim \left(\frac{1}{\Delta_p} m^{p-1} + \frac{1}{\Delta_2} m \right) \sqrt{1-m^2} \mathbf{e}_1 - \sqrt{\frac{1}{\Delta_p} + \frac{1}{\Delta_2}} \frac{1}{\sqrt{N}} \tilde{\mathbf{Z}}; \quad (7)$$

$$\frac{h}{N} \sim \left(\frac{p-1}{\Delta_p} m^{p-2} + \frac{1}{\Delta_2} \right) (1-m^2) \mathbf{e}_1 \mathbf{e}_1^T + \sqrt{\frac{p-1}{\Delta_p} + \frac{1}{\Delta_2}} \sqrt{\frac{N-1}{N}} \mathbb{W} - (pf_p + 2f_2) \mathbb{1}_{N-1}; \quad (8)$$

with Z_k standard Gaussians and $k \in \{2, p\}$, $\tilde{\mathbf{Z}} \sim \mathcal{N}(0, \mathbb{1}_{N-1})$ a standard multivariate Gaussian and $\mathbb{W} \sim \text{GOE}(N-1)$ 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 .

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

$$g_i(\sigma) = G_i(\sigma) - \mu \sigma_i, \quad (10)$$

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

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 we 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, $\xi_{i_1 \dots i_p}, \xi_{ij}$. For $k \in \{2, p\}$ the first moment leads to

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

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

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

Using standard results for derivatives of Gaussians (see e.g. [17] 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 $\{\mathbf{e}_j\}_{j=1, \dots, N}$. Align the configuration along the last coordinate $\mathbf{e}_N = \sigma$ and the signal with a combination of the first and last coordinates $\sigma^* = \sqrt{1 - m^2} \mathbf{e}_1 + m \mathbf{e}_N$. Finally, project on the sphere by discarding the last coordinate. \square

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

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

with t_N and θ_N given by

$$t_N \rightarrow t = 2 \frac{p\epsilon_p + 2\epsilon_2}{\sqrt{\frac{p-1}{\Delta_p} + \frac{1}{\Delta_2}}}, \quad (15)$$

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

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 [12] and we can thus deduce the expression for the complexity as

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

with

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

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

We note at this point that for the case of the pure spiked tensor model $\Delta_2 \rightarrow \infty$ the above expression reduces exactly to the complexity derived in [12]. 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*

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{E} \mathcal{N}(m, \epsilon_2, \epsilon_p; \Delta_2, \Delta_p) &= \\ &= \tilde{\Sigma}_{\Delta_2, \Delta_p}(m, \epsilon_p, \epsilon_2) \end{aligned} \quad (18)$$

Proof sketch. The proof comes immediately from [12] Thm. 2, see also Sec. 4.1. \square

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

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

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 [19]. It is interesting to notice that in [19] 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)

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

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 $N \rightarrow \infty$.

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

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

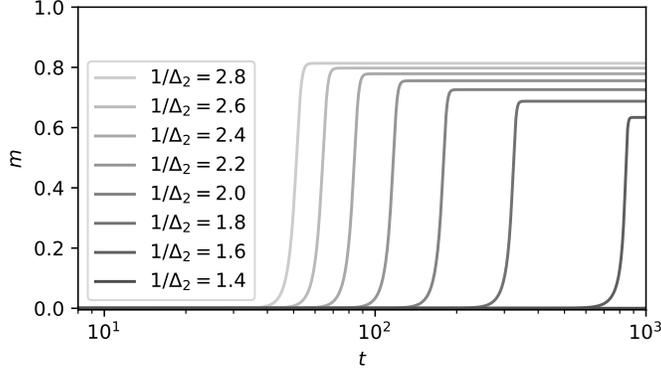


Figure 3: Eq. (24) characterizes the evolution of the correlation of the gradient flow with the ground truth signal, evaluated for several values of Δ_2 , at $\Delta_p = 4.0$ starting from $m(0) = 10^{-10}$. The dynamics displays a fast increase of the convergence time as Δ_2 increases. At large times, 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.

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 ξ_{i_1, \dots, i_p} . As we take the limit $T \rightarrow 0$, the noise becomes peaked around zero, effectively recovering the gradient flow.

The performance of the Langevin algorithm was characterized recently in [1] using equations developed in physics of disordered systems [13, 14]. In [1] this characterization was given for an arbitrary temperature T and compared to the landscape of the Bayes-optimal estimator [20]. Here we hence summarize and use the results of [1] corresponding to the limit $T \rightarrow 0$.

The Langevin dynamics with generic temperature is in the large size limit, $N \rightarrow \infty$, characterized by a set of PDEs for the self-correlation $C(t, t') = \lim_{N \rightarrow \infty} \langle \frac{1}{N} \sum x_i(t)x_i(t') \rangle$, the response function $R(t, t') = \lim_{N \rightarrow \infty} \langle \frac{1}{N} \sum \frac{\delta x_i(t)}{\delta \eta_i(t')} \rangle$, and the correlation with the signal $m(t) = \lim_{N \rightarrow \infty} \langle \frac{1}{N} \sum x_i(t)x_i^* \rangle$. Ref. [1] 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

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

$$\frac{\partial}{\partial t} R(t, t') = -\tilde{\mu}(t)R(t, t') + \int_{t'}^t dt'' R(t, t'')Q''(C(t, t''))R(t'', t'), \quad (23)$$

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

with $Q(f) = f^p/(p\Delta_p) + f^2/(2\Delta_2)$ and $\tilde{\mu}(t) = \lim_{T \rightarrow 0} T\mu(t)$ 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 $\lim_{t' \rightarrow t^-} R(t, t') = 1 \forall t$. 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 [21].

4.1 Performance of the gradient flow

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 will be made available 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 Δ_2^{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 Δ_2^{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 time happens only well inside the phase where spurious local minimizers do exist.

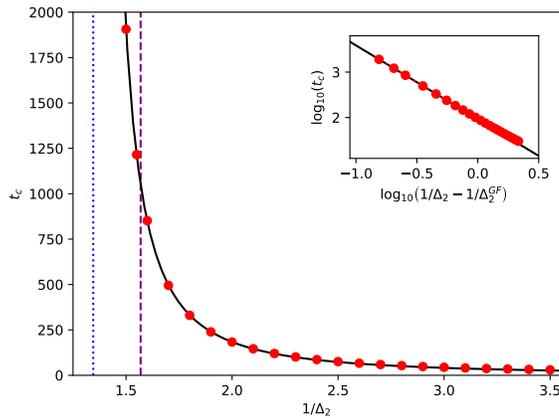


Figure 4: The convergence time the gradient flow takes to find a configuration well correlated with the signal for $\Delta_p = 4.0$, $p = 3$ as a function of Δ_2 , starting from $m(0) = 10^{-10}$. The points are fitted with a power law consistent with a divergence point $1/\Delta_2^{\text{GF}} = 1.35$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

5 Maximum-likelihood approximate message passing

Approximate Message Passing (AMP) is a popular iterative algorithm [22] with a key advantage of being analyzable via its state evolution [23]. The maximum-likelihood AMP (ML-AMP) algorithm studied in this paper is a generalization of AMP for the pure spiked tensor model from [16] 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 [1] 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 appendix, Sec. C, we provide a schematic derivation of the ML-AMP as a zero-temperature limit of the BO-AMP, using a scheme similar to [24].

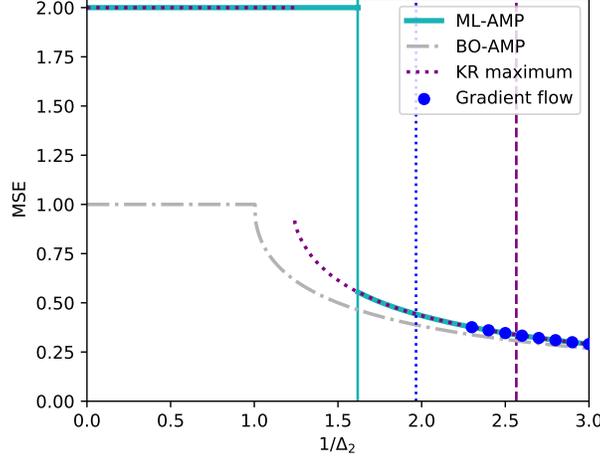


Figure 5: We show the mean-squared error (MSE) as achieved by the analyzed algorithms, for $p = 3$, $\Delta_p = 1.0$ as a function of the signal-to-noise (snr) ratio $1/\Delta_2$. The full cyan line corresponds to the error reached by the ML-AMP algorithm, it jumps discontinuously at $1/\Delta_2^{\text{ML-AMP}} = 1.62$. 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 $1/\Delta_2^{\text{GF}} = 1.97$, 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 $1/\Delta_2^{\text{triv}} = 2.57$ 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 up to numerical precision. For the sake of comparison we show (the dashed-dotted grey line) also the minimal-MSE achieved in the Bayes-optimal setting.

The ML-AMP algorithm reads

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

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

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

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

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

5.1 ML-AMP & stationary points of the loss

Using an argument similar to Prop. 5.1 in [25] 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}^*, σ^*) 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

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

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

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

Moreover ML-AMP by construction preserves the spherical constrain at every time iteration, thus the fixed point value of the Lagrange multiplier in (29) is the one leading to the spherical constraint. \square

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 proven in [23, 16], this parameter evolves in the large N limit as

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

and the mean square error correspondingly

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

A derivation of this state evolution is presented in the appendix, Sec. C.

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 $\Delta_2 > \Delta_2^{\text{ML-AMP}}$ where

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

Consequently for $\Delta_2 > \Delta_2^{\text{ML-AMP}}$ the ML-AMP algorithm converges to $m = 0$, i.e. zero correlation with the signal. The line $\Delta_2^{\text{ML-AMP}}$ is the line plotted in Fig. 1. For $p = 3$ and $p = 4$, we obtain that for $\Delta_2 < \Delta_2^{\text{ML-AMP}}$ 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 (up to the numerical precision) with the position of the maximum having largest value of m in the complexity function $\Sigma(m)$, this is also depicted in the figure. The trivialization of the landscape occurs at $\Delta_2^{\text{triv}} < \Delta_2^{\text{ML-AMP}}$, thus showing that for $\Delta_2^{\text{triv}} < \Delta < \Delta_2^{\text{ML-AMP}}$ 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 [1]. 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 increases.

6 Discussion

We analyzed the behavior of two descent algorithms in 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 sense that spurious local minima disappear. We analyzed the performance of gradient flow via integro-differential state-evolution-like equations. We numerically solved the equations and extrapolated the divergence of their convergence-time. 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 (numerically extrapolated) region for which the gradient flow works. We found that in cases when both the algorithms converge to an informative minima, the corresponding error is the same in both and also corresponds to the position of all the minima well correlated with the signal in the Kac-Rice approach. 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; from the French National Research Agency (ANR) grant PAIL; and from "Investissements d'Avenir" LabEx PALM (ANR-10-LABX-0039-PALM) (SaMURai and StatPhysDisSys). This research was supported in part by the National Science Foundation under Grant No. PHY-1748958.

References

- [1] Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Marvels and pitfalls of the langevin algorithm in noisy high-dimensional inference. *arXiv preprint arXiv:1812.09066*, 2018.
- [2] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.
- [3] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.
- [4] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.
- [5] C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.
- [6] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.
- [7] Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74, 2017.

- [8] Simon S Du, Jason D Lee, Yuandong Tian, Barnabas Poczos, and Aarti Singh. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. *arXiv preprint [arXiv:1712.00779](#)*, 2017.
- [9] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1233–1242, 2017. arXiv preprint [arXiv:1704.00708](#).
- [10] Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima. *arXiv preprint [arXiv:1702.08580](#)*, 2017.
- [11] Yan V Fyodorov. Complexity of random energy landscapes, glass transition, and absolute value of the spectral determinant of random matrices. *Physical review letters*, 92(24):240601, 2004.
- [12] Gerard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica. The landscape of the spiked tensor model. *arXiv preprint [arXiv:1711.05424](#)*, 2017.
- [13] A Crisanti, H Horner, and H-J Sommers. The spherical p -spin interaction spin-glass model. *Zeitschrift für Physik B Condensed Matter*, 92(2):257–271, 1993.
- [14] Leticia F Cugliandolo and Jorge Kurchan. Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model. *Physical Review Letters*, 71(1):173, 1993.
- [15] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.
- [16] Emile Richard and Andrea Montanari. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.
- [17] Robert J Adler and Jonathan E Taylor. *Random fields and geometry*. Springer Science & Business Media, 2009.
- [18] Yan V Fyodorov. High-dimensional random fields and random matrix theory. *Markov Processes Relat. Fields*, 21:483–518, 2015.
- [19] Valentina Ros, Gerard Ben Arous, Giulio Biroli, and Chiara Cammarota. 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.
- [20] Fabrizio Antenucci, Silvio Franz, Pierfrancesco Urbani, and Lenka Zdeborová. On the glassy nature of the hard phase in inference problems. *arXiv preprint [arXiv:1805.05857](#)*, 2018.
- [21] Gerard Ben Arous, Amir Dembo, and Alice Guionnet. Cugliandolo-Kurchan equations for dynamics of spin-glasses. *Probability theory and related fields*, 136(4):619–660, 2006.
- [22] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, Nov 2009.
- [23] Adel Javanmard and Andrea Montanari. 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.
- [24] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403, 2017.
- [25] Andrea Montanari. Graphical models concepts in compressed sensing. *Compressed Sensing: Theory and Applications*, pages 394–438, 2012.
- [26] Federico Ricci-Tersenghi, Guilhem Semerjian, and Lenka Zdeborová. Typology of phase transitions in bayesian inference problems. preprint:[arXiv:1806.11013](#).

A Kac-Rice formula

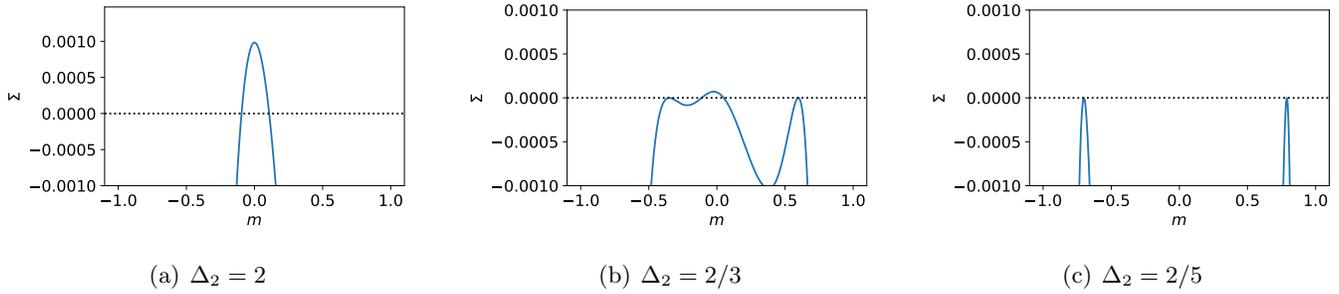


Figure 6: Analogously to Fig. 2, the figures show the complexity, Eq. (19), as a function of the correlation with the signal for different values of parameter Δ_2 at fixed $\Delta_p = 4.0$ in the case $p = 3$.

A.1 p -odd cases

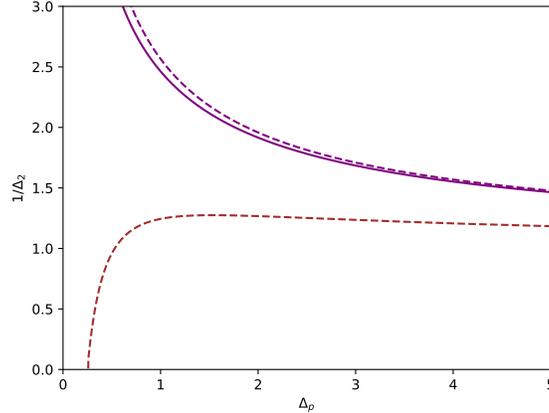


Figure 7: The thresholds representing the trivialization of the landscape (purple) and the point where the support of $\Sigma(m) \geq 0$ become disconnected (brown) for tensors of order $p = 3$. We compare the two definitions of the trivialization threshold described in Sec. A.1: the solid line considers just the positivity of the complexity Eq. (19) at $m = 0$, the dashed line considers the whole non-informative band.

In the cases in which the order of the tensor p is odd we encounter an interesting phenomenon due to the different symmetries of the two types of observation. The matrix is symmetric by inverting the sign of the signal, $\hat{x} \mapsto -\hat{x}$, while the tensor is not symmetric for odd p . This creates an asymmetry in the complexity, Fig. 6 (to be compared with Fig. 2) and causes a shift toward lower correlations of the band characterizing the non-informative minima. Therefore observing when the complexity at $m = 0$ becomes negative does not guarantee that the non-informative minima disappeared. To do so, one must check that the whole non-informative band disappears. This should be contrasted with the case of even p where a maximum of the complexity $\Sigma(m)$ is always at $m = 0$. These two definitions of the threshold have little, but not negligible, difference, see Fig. 7.

Observe that as Δ_p increases the peak of the complexity decreases, since the loss Eq. (3) tends to the simple matrix-factorization problem where the landscape is characterized by two isolated minima. This implies that the two definitions become indistinguishable for large Δ_p . In the main text we use the definition taking into account the maximum (even when it is not strictly at $m = 0$) because gives a more accurate characterization of the trivialization threshold.

B Gradient Flow

B.1 Dependence on the initial conditions

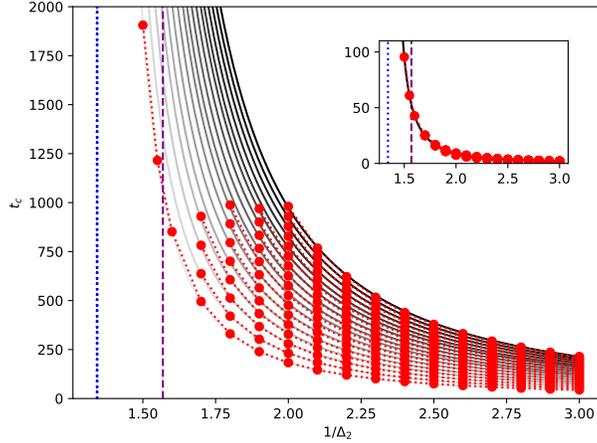


Figure 8: The time corresponding to convergence close to the signal is shown for $\Delta_p = 4.0$ in the case $p = 3$. Different shades of grey correspond to different initial conditions, from $m(0) = 10^{-10}$ (light grey) to $m(0) = 10^{-42}$ (dark grey). The different initializations collapse to a single line when the time is rescaled by $a^{\log m(0)}$ with $a = 1.3$, see inset. In the figure we fit only the case $m(0) = 10^{-10}$ with a power law and use the same parameters for all the other fits with a vertical translation. The divergence point extrapolated is $1/\Delta_2^{\text{GF}} = 1.35$ and is represented by the vertical dotted line, while the dashed line identifies the landscape trivialization predicted with the Kac-Rice formula, $1/\Delta_2^{\text{triv}} = 1.57$.

The dynamics of the gradient flow shows a dependence on the initial conditions, because formally zero correlation is a (unstable) fixed point of the GF state evolution. In practice we observe for both GF and ML-AMP that instability of the fixed point is sufficient for good performance of the algorithm. However, this makes the definition of the convergence time depend of the initial condition.

We observed from our numerical solution of the GF state evolution equations that the initial condition add a factor $a^{\log m(0)}$ to the convergence times. Thus by fitting this term and rescaling the convergence times, the different initializations collapse into a single curve, see inset of Fig. 8. Finally, the collapsed points were used to extrapolate the critical line as shown in the main text, Fig. 4.

C AMP

C.1 From AMP to ML-AMP

In this section we consider the spiked-tensor model in a Bayesian way. We show how the Bayes-optimal AMP leads to the Maximum Likelihood AMP using a temperature-like parameter T . We will introduce the algorithm

AMP for a generic T , and show that as $T \rightarrow 0$ we recover ML-AMP as presented in the main text. The probability distribution we consider is

$$P(X|Y, T) \propto e^{-\mu \|x\|^2} \prod_{i < j} e^{-\frac{1}{2T\Delta_2} \left(Y_{ij} - \frac{x_i x_j}{\sqrt{N}} \right)^2} \prod_{i_1 < \dots < i_p} e^{-\frac{1}{2T\Delta_p} \left(T_{i_1 \dots i_p} - \frac{\sqrt{(p-1)!}}{N^{(p-1)/2}} x_{i_1} \dots x_{i_p} \right)^2}. \quad (33)$$

The scheme for deriving AMP estimating marginals of such a probability distribution can be found in [24, 1] and consist in making a Gaussian assumption on the distribution of the messages in the belief propagation (BP) algorithm and neglecting the node-dependence in the messages. A final consideration to be used in order to derive the algorithm is that the spherical constrain can be imposed by setting $\frac{1}{N} \sum_i (\hat{x}_i^2 + \sigma_i) = 1$ at every iteration. The resulting AMP algorithm will iterate on the following equations:

$$B_i^t = \frac{\sqrt{(p-1)!}}{N^{(p-1)/2}} \sum_{k_2 < \dots < k_p} \frac{T_{ik_2 \dots k_p}}{T\Delta_p} \hat{x}_{k_2}^t \dots \hat{x}_{k_p}^t + \frac{1}{\sqrt{N}} \sum_k \frac{Y_{ik}}{T\Delta_2} \hat{x}_k^t - r_t \hat{x}_i^{t-1} \quad (34)$$

$$\hat{x}_i^{t+1} = 2 \frac{B_i^t}{1 + \sqrt{1 + \frac{4}{N} \|B^t\|_2^2}}, \quad (35)$$

$$\sigma^{t+1} = \frac{2}{1 + \sqrt{1 + \frac{4}{N} \|B^t\|_2^2}}. \quad (36)$$

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

$$r_t = \frac{1}{\Delta_2 T^2} \frac{1}{N} \sum_k \sigma_k^t + \frac{p-1}{\Delta_p T^2} \frac{1}{N} \sum_k \sigma_k^t \left(\frac{1}{N} \sum_k \hat{x}_k^t \hat{x}_k^{t-1} \right)^{p-2}. \quad (37)$$

In the limit $T \rightarrow 0$ AMP defined by Eqs. (34-37) is equivalent to ML-AMP, Eqs. (25-28). To see this we define the rescaled variables $\hat{\sigma}^t \doteq \sigma^t/T$, $\tilde{B}^t \doteq T B^t$ and $\tilde{r}_t \doteq T r_t$. Taking the limit $T \rightarrow 0$ the expression for \hat{x}_i^{t+1} Eq. (35) and the expression for $\hat{\sigma}_i^{t+1}$ Eq. (36) simplify as Eq. (26) and as Eq. (26) respectively. Dropping the tildes we obtain ML-AMP as presented in the main text.

C.2 State evolution

The generic T version of AMP has a slightly more complicated SE that depends of two order parameters: the already introduced $m^t = \frac{1}{N} \sum_i \hat{x}_i^t x_i^*$ and $q^t = \frac{1}{N} \sum_i (\hat{x}_i^t)^2$ the self-overlap of the estimator. The SE equations are:

$$m^{t+1} = 2 \frac{z^t(T)}{1 + \sqrt{1 + 4y^t(T)}}, \quad (38)$$

$$q^{t+1} = 4 \frac{y^t(T)}{\left(1 + \sqrt{1 + 4y^t(T)}\right)^2} \quad (39)$$

and

$$\text{MSE}^t = 1 - 2m^t + q^t, \quad (40)$$

with $y^t(T) = (z^t(T))^2 + \left(\frac{1}{T^2} \frac{q^t}{\Delta_2} + \frac{1}{T^2} \frac{(q^t)^{p-1}}{\Delta_p} \right)$ and $z^t(T) = \frac{1}{T} \frac{m^t}{\Delta_2} + \frac{1}{T} \frac{(m^t)^{p-1}}{\Delta_p}$.

Given $\frac{1}{N} \|\hat{x}^0\|_2^2 \neq 0$, in the limit $T \rightarrow 0$ AMP SE Eqs. (38-39) simplify, to a single equation corresponding to ML-AMP SE Eq. (41). This is seen by taking the limit for Eq. (39) which gives $q^t = 1 \forall t > 0$, implying $\text{MSE}^t = 2(1 - m^t)$. Then, using the result for q^t , we show that Eq. (38) tends to Eq. (30).

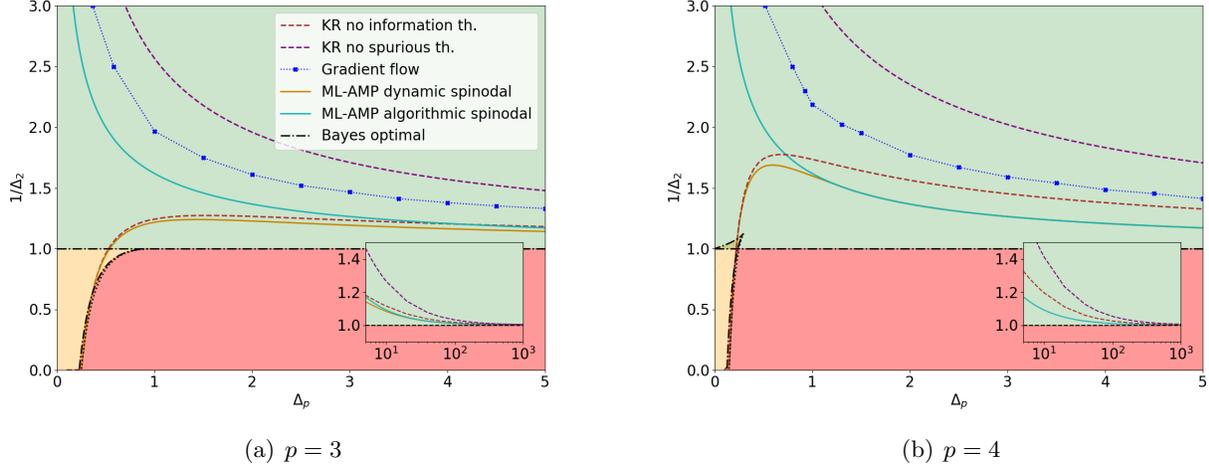


Figure 9: The phase diagram already describe in Fig. 1 with two additional lines. The dashed brown line is the limit predicted by Kac-Rice formula where the support of the $\Sigma(m) \geq 0$ becomes disconnected (above the line). The full orange line is related to the ML-AMP algorithm, is called *dynamical spinodal*, below it the algorithm converges to $m = 0$ even if initialized in the solution. In the insets we show the large Δ_p behaviour of the thresholds, where we can observe that the lines merge at infinity.

C.3 Derivation of spinodals

From SE Eq. (30) we can obtain analytical equations for the *spinodals*, the threshold of stability of the different ML-AMP fixed points. We have $\hat{x}^{t+1} = f_{SE}(z^t)$ with

$$f_{SE}(z) = \frac{z}{\sqrt{z^2 + \gamma}}, \quad (41)$$

with $\gamma = 1/\Delta_2 + 1/\Delta_p$ and $z = m/\Delta_2 + m^{p-1}/\Delta_p$. Observe that: $f'_{SE}(z) = \frac{\gamma}{(z^2 + \gamma)^{\frac{3}{2}}}$. We can now define either $\Delta_p \equiv \Delta_p(z; \Delta_2, \gamma) = \frac{f_{SE}(z)^{p-1}}{z - \frac{f_{SE}(z)}{\Delta_2}}$ or $\Delta_2 \equiv \Delta_2(z; \Delta_p, \gamma) = \frac{f_{SE}(z)}{z - \frac{f_{SE}(z)^{p-1}}{\Delta_p}}$.

As remarked in [1], the spinodals are given by the following condition:

$$0 = \frac{d \log \Delta_2}{dm} \propto \frac{d \log \Delta_2}{dz} = \frac{z \left[(p-2)\gamma \left(\frac{z}{\sqrt{z^2 + \gamma}} \right)^{p-1} - z^3 \Delta_p \right]}{z(z^2 + \gamma) \left[\Delta_p z^2 - z \left(\frac{z}{\sqrt{z^2 + \gamma}} \right)^{p-1} \right]}. \quad (42)$$

A trivial solution is given by $z \rightarrow 0$ corresponding to stability of the non-informative solution $m = 0$, and gives the *algorithmic spinodal* for the cases $p \in \{3, 4\}$. This solution and has a very simple equation for every p : $\Delta_2 = 1/\sqrt{\gamma}$ giving Eq. (32), already presented in the main text. An interesting implication of Eq. (32) is that it is independent from the value of p , it is in some sense *universal* among the $2 + p$ -models.

The expression for the stability of the informative solution, *dynamical spinodal*, is less straightforward, but analytical progresses can be done in the cases $p = 3$ and $p = 6$ (using Cardano formula) and in the case $p = 4$ for which it is equivalent to a second order polynomial

$$z^2 + \gamma = \left(\frac{(p-2)\gamma}{\Delta_p} \right)^{\frac{2}{(p-1)}} = \left(\frac{2\gamma}{\Delta_4} \right)^{\frac{2}{3}}, \quad (43)$$

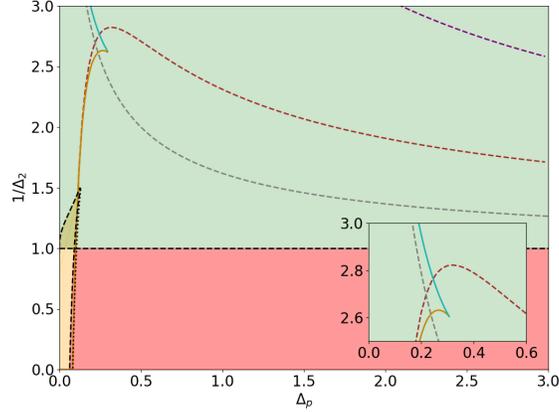


Figure 10: Phase diagram as shown in Fig. 9 for the case $p = 6$. The difference between $p = 3, 4$ and $p > 4$ is that a new phase appears, called hybrid hard phase, where two fixed points of ML-AMP aligned with the signal are present and the convergence to one or the other depends on the initialization. The region is highlighted in the inset. In the phase diagram the grey dashed line represent the threshold above which the non-informative fixed point becomes unstable.

p	Δ_2	Δ_p
4	$\frac{2}{3} \simeq 0.667$	$\frac{4}{3} \simeq 1.333$
5	0.470	0.451
6	0.384	0.305
7	0.322	0.220
8	0.279	0.172
9	0.246	0.147
10	0.220	0.121

Table 1: Table of the values of tricritical points for $p \geq 4$.

that admits a single solution in \mathbb{R}^+ :

$$z = \sqrt{\left(\frac{2\gamma}{\Delta_4}\right)^{\frac{2}{3}} - \gamma}. \quad (44)$$

An important point in the phase diagram is where the algorithmic and dynamical spinodals meet, this is called the *tricritical point*. Its value is obtained for different p , numerically (for $p > 4$) and analytically (for $p = 4$), and is reported in Table 1. The case $p = 3$ does not show any tricritical point for any finite Δ_p , the two lines eventually meet at $\Delta_p = \infty$ when the spiked matrix problem is recovered.

For the cases $p > 4$ we observe additionally the zero temperature analog of what is called *hybrid phase* in AMP in Bayes-optimal regime [26]. The hybrid phase is illustrated in Fig. 10. This phase is defined as a region where the ML-AMP algorithm initialized at random converges to a solution with positive correlation but that is less correlated than the solution achievable starting from the solution. In these cases Eq. (32) does not correspond to the algorithmic spinodal but it is just the stability of the non-informative solution.