Who is afraid of big bad minima? Analysis of gradient-flow in a spiked matrix-tensor model

Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Lenka Zdeborová

▶ **To cite this version:**

Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Lenka Zdeborová. Who is afraid of big bad minima? Analysis of gradient-flow in a spiked matrix-tensor model. Advances in Neural Information Processing Systems, 2019, 32, pp.8676-8686. cea-02529145

HAL Id: cea-02529145
https://cea.hal.science/cea-02529145

Submitted on 2 Apr 2020

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.

arXiv:1907.08226v3 [cs.LG] 20 Jan 2020

Who is Afraid of Big Bad Minima? Analysis of Gradient-Flow in a Spiked Matrix-Tensor Model

Stefano Sarao Mannelli[1], Giulio Biroli[2], Chiara Cammarota[3], Florent Krzakala[2], and Lenka Zdeborová[1]

[1]Institut de physique théorique, Université Paris Saclay, CNRS, CEA, 91191 Gif-sur-Yvette, France
[2]Laboratoire de Physique de l'Ecole normale supérieure ENS, Université PSL, CNRS, Sorbonne Université, Université Paris-Diderot, Sorbonne Paris Cité Paris, France
[3]Department of Mathematics, King's College London, Strand London WC2R 2LS, UK

Abstract

Gradient-based algorithms are effective for many machine learning tasks, but despite ample recent effort and some progress, it often remains unclear why they work in practice in optimising high-dimensional non-convex functions and why they find good minima instead of being trapped in spurious ones. Here we present a quantitative theory explaining this behaviour in a spiked matrix-tensor model. Our framework is based on the Kac-Rice analysis of stationary points and a closed-form analysis of gradient-flow originating from statistical physics. We show that there is a well defined region of parameters where the gradient-flow algorithm finds a good global minimum despite the presence of exponentially many spurious local minima. We show that this is achieved by surfing on saddles that have strong negative direction towards the global minima, a phenomenon that is connected to a BBP-type threshold in the Hessian describing the critical points of the landscapes.

1 Introduction

A common theme in machine learning and optimisation is to understand the behaviour of gradient descent methods for non-convex problems with many minima. Despite the non-convexity, such methods often successfully optimise models such as neural networks, matrix completion and tensor factorisation. This has motivated a recent spur in research attempting to characterise the properties of the loss landscape that may shed some light on the reason of such success. Without the aim of being exhaustive these include [1–11].

Over the last few years, a popular line of research has shown, for a variety of systems, that spurious local minima are not present in certain regimes of parameters. When the signal-to-noise ratio is large enough, the success of gradient descent can thus be understood by a trivialisation transition in the loss landscape: either there is only a single minima, or all minima become "good", and no spurious minima can trap the dynamics. This is what happens, for instance, in the limit of small noise and abundance of data for matrix completion and tensor factorization [3, 8], or for some very large neural networks [1, 2]. However, it is often observed in practice that these guarantees fall short of explaining the success of gradient descent, that is empirically observed to find good minima very far from the regime under mathematical control. In fact, gradient-descent-based algorithms may be able to perform well even when spurious local minima are present because the basins of attraction of the spurious minima may be small and the dynamics might be able to avoid them. Understanding this behaviour requires, however, a very detailed characterisation of the dynamics and of the landscape, a feat which is not yet possible in full generality.

A fruitful direction is the study of Gaussian functions on the N-dimensional sphere, known as p-spin spherical spin glass models in the physics literature, and as isotropic models in the Gaussian process literature [12–16]. In statistics and machine learning, these models have appeared following the studies of spiked matrix and tensor models [17–19]. In particular, a very recent work [20] showed explicitly that for a spiked matrix-tensor model the gradient-flow algorithm indeed reaches global minimum even when spurious local minima are present and the

authors estimated numerically the corresponding regions of parameters. In this work we consider this very same model and explain the mechanism by which the spurious local minima are avoided, and develop a quantitative theoretical framework that we believe has a strong potential to be generic and extendable to a much broader range of models in high-dimensional inference and neural networks.

The Spiked Matrix-Tensor Model. The spiked matrix-tensor model has been recently proposed to be a prototypical model for non-convex high-dimensional optimisation where several non-trivial regimes of cost-landscapes can be displayed quantitatively by tuning the parameters [20, 21]. In this model, one aims at reconstructing a hidden vector (i.e. the spike) $\boldsymbol{\sigma}^*$ from the observation of a noisy version of *both* the rank-one matrix and rank-one tensor created from the spike. Using the following notation: bold lowercase symbols represent vectors, bold uppercase symbols represent matrices or tensors, and $\langle \cdot, \cdot \rangle$ represent the scalar product, the model is defined as follows: given a signal (or spike), $\boldsymbol{\sigma}^*$, uniformly sampled on the N-dimensional hyper-sphere of radius 1, it is given a tensor $\boldsymbol{T}$ and a matrix $\boldsymbol{Y}$ such that

$$T_{i_1 \dots i_p} = \eta_{i_1 \dots i_p} + \sqrt{N(p-1)!}\, \sigma^*_{i_1} \dots \sigma^*_{i_p} \,, \tag{1}$$

$$Y_{ij} = \eta_{ij} + \sqrt{N} \sigma^*_i \sigma^*_j \,, \tag{2}$$

where $\eta_{i_1 \dots i_p}$ and η_{ij} are Gaussian random variables of variance Δ_p and Δ_2 respectively. Neglecting constant terms, the maximum likelihood estimation of the ground truth, $\boldsymbol{\sigma}^*$, corresponds to minimization of the following loss function:

$$\ell(\boldsymbol{\sigma}|\boldsymbol{T},\boldsymbol{Y}) = -\frac{\sqrt{(p-1)!}}{\Delta_p \sqrt{N}} \sum_{i_1 < \dots < i_p} T_{i_1 \dots i_p} \sigma_{i_1} \dots \sigma_{i_p} - \frac{1}{\Delta_2 \sqrt{N}} \sum_{i<j} Y_{ij} \sigma_i \sigma_j = \tag{3}$$

$$-\frac{\sqrt{(p-1)!}}{\Delta_p \sqrt{N}} \sum_{i_1 < \dots < i_p} \eta_{i_1 \dots i_p} \sigma_{i_1} \dots \sigma_{i_p} - \frac{1}{\Delta_2 \sqrt{N}} \sum_{i<j} \eta_{ij} \sigma_i \sigma_j - \frac{\langle \boldsymbol{\sigma}, \boldsymbol{\sigma}^* \rangle^p}{p \Delta_p} - \frac{\langle \boldsymbol{\sigma}, \boldsymbol{\sigma}^* \rangle^2}{2 \Delta_2} \,.$$

The first two contributions of the last equation will be denoted ϵ_p and ϵ_2 in the following. While the matricial observations correspond to a quadratic term and thus to a simple loss-landscape, the additional order-p tensor contributes towards a rough non-convex loss landscape. As $\Delta_p \to \infty$ the information $T_{i_1 \dots i_p}$ becomes irrelevant and the landscape becomes trivial, while in the opposite limit $\Delta_2 \to \infty$, the landscape becomes extremely rough and complex as analyzed recently in [16, 22].

We shall consider the behaviour of the gradient-flow (GF) algorithm aiming to minimise the loss:

$$\dot{\sigma}_i(t) = -\mu(t) \sigma_i(t) - \frac{\partial \ell(\boldsymbol{\sigma}(t)|\boldsymbol{T},\boldsymbol{Y})}{\partial \sigma_i(t)} \,, \tag{4}$$

where $\mu(t)$ enforces that $\boldsymbol{\sigma}(t)$ belongs to the hyper-sphere of radius N and will be referred to as the *spherical constraint*. The algorithm is initialised in a random point drawn uniformly on the hyper-sphere, thus initially having no correlation with the ground-truth signal. We view the gradient-flow as a prototype of gradient-descent-based algorithms that are the work-horse of current machine learning.

Main Contributions. The first main result of this paper is the expression for the threshold below which the gradient-flow algorithm finds a configuration correlated the hidden spike. This threshold is established in the asymptotic limit of large N, fixed p and Δ_p, and reads:

$$\Delta_2^{\rm GF}(\Delta_p, p) \equiv \frac{-\Delta_p + \sqrt{\Delta_p^2 + 4(p-1)\Delta_p}}{2(p-1)} \,. \tag{5}$$

We find that (i) for $\Delta_2 < \Delta_2^{\rm GF}$ the gradient flow reaches in finite time the global minimum, well correlated with the signal, while (ii) for $\Delta_2 > \Delta_2^{\rm GF}$ the algorithm remains uncorrelated with the signal for all times that do not grow as N grows. We contrast it with the threshold $\Delta_2^{\rm triv} < \Delta_2^{\rm GF}$, established in [20], below which the energy landscape does not present any spurious local minima. Note that $\Delta_2^{\rm GF}$ is less than $\Delta_2^{\rm AMP} = 1$ [21], below which the best known algorithm, specifically the approximate message passing, works.

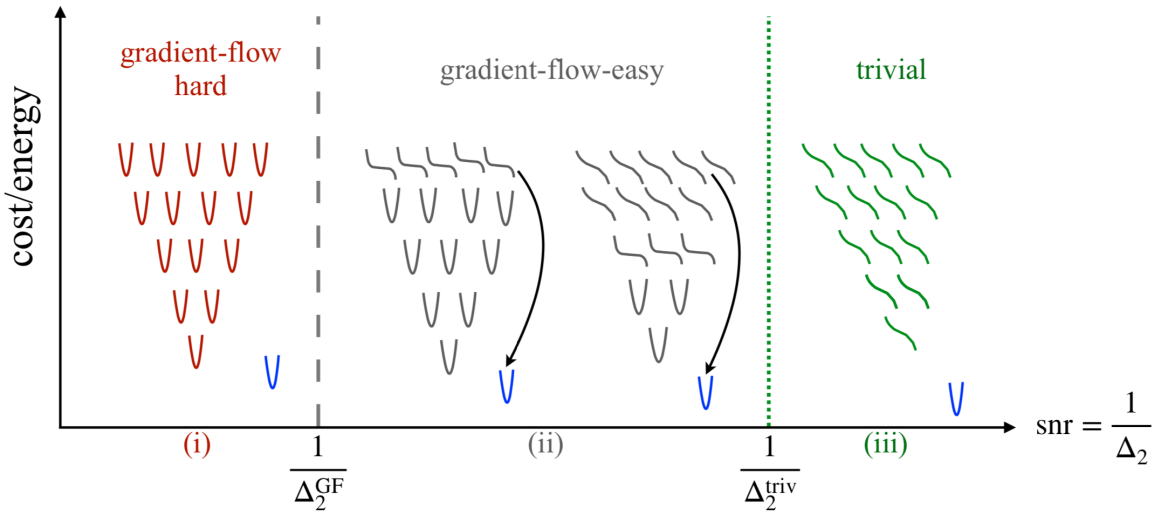


Figure 1: Cartoon illustrating the mechanism by which gradient-flow avoids spurious local minima in the spiked matrix-tensor model. As the signal-to-noise (snr) ratio $1/\Delta_2$ is increased, the spurious local minima that attract the randomly initialised GF algorithm develop a single negative direction towards the global minimum before the others, in particular the lower-cost spurious local minima, do. This has drastic consequences on the GF algorithm. In region (i), for snr $< 1/\Delta_2^{\rm GF}$, the algorithm goes down the landscape, eventually reaches the high-energy *threshold* minima and remains stuck. In region (ii), however, these threshold minima are turned into saddles with a strong negative direction towards the signal. The algorithm is initially reaching these minima-turned-saddles, surfing on the negative slope, it then turns towards the "good" minima correlated with the signal, avoiding the exponentially many spurious minima at lower energies. The main technical contribution of this paper is a quantitative description of this scenario, including a simple formula for the corresponding threshold $\Delta_2^{\rm GF}$, eq. (5). As the snr is further increased, the negative direction appears in lower and lower minima until the trivialization transition in region (iii): for snr $> 1/\Delta_2^{\rm triv}$, all the spurious minima have been turned into saddles.

The second main result of this paper, is the insight we obtain on the behaviour of the gradient-flow in the loss landscape, that is summarised in Fig. 1. The key point is to consider the fate of the spurious local minima that attract the GF algorithm when the signal to noise ratio snr $= 1/\Delta_2$ is increased. As the snr increases, these minima turn into saddles with a single negative direction towards the signal (a phenomenon that we analyze in the next section, and that turns out to be linked to the BBP transition [23] in random matrix theory), all that well before all the other spurious local minima disappear. We present two ways to quantify this insight:

(a) We use the Kac-Rice formula for the number of stationary points, as derived for the present model in [20]. In [20] this formula is used to quantify the region with no spurious local minima. Here we focus on a BBP-type of phase transition that is crucial in the derivation of this formula and deduce the GF threshold (5) from it.

(b) We use the CHSCK equations [24, 25] for closed-form description of the behaviour of the gradient-flow, as derived and numerically solved in [20, 21]. Building on dynamical theory of mean-field spin glasses we determine precisely when and how the algorithm escapes the manifold of zero overlap with the signal, leading again to the threshold (5).

Both these arguments are derived using reasoning common in theoretical physics. From a mathematically rigorous point of view the threshold (5) remains a conjecture and its rigorous proof is an interesting challenge for future work. We note that both the Kac-Rice approach [16] and the CHSCK equations [26] have been made rigorous in closely related problems. We believe that the reported results are not limited to the present model and we will investigate analytically and numerically other models and real-data-based learning in order to validate this theory and to understand its limitations.

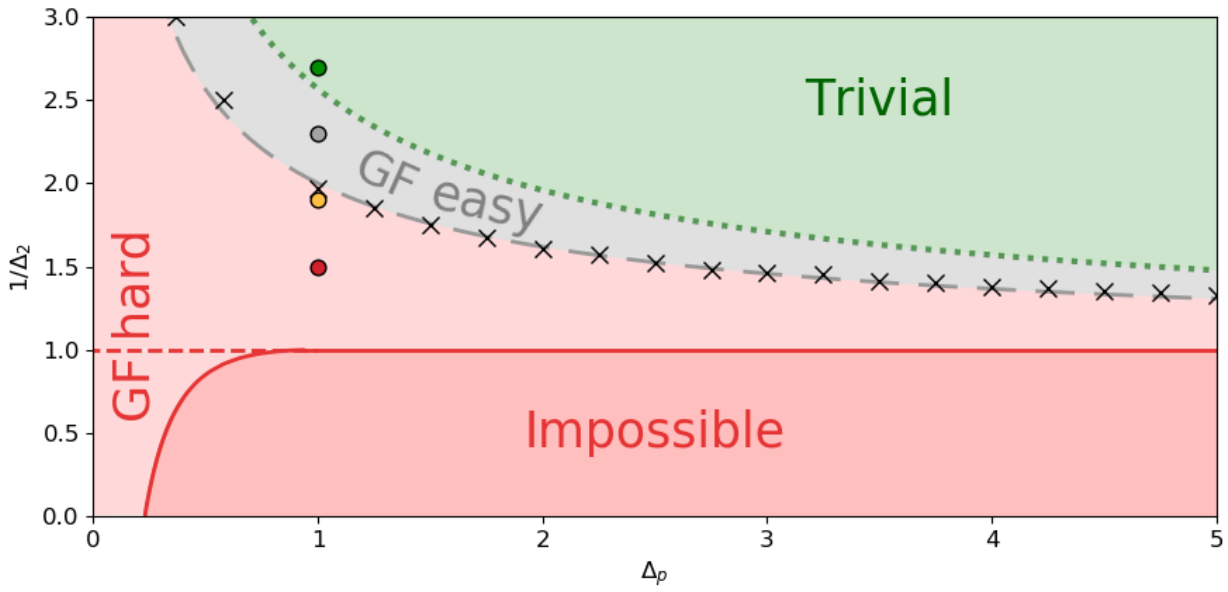


Figure 2: The phase diagram shows the different regions for gradient-flow behaviour, for the spiked matrix-tensor model with $p = 3$. In the region shaded in red (light and dark), GF does not correlate with the signal, while it does in the grey and green regions. In the dark-red region obtaining correlation with the signal is *impossible* information-theoretically [21]. The possible region is divided by a red dashed line, below that line even best known algorithms are unable to obtain correlation with the signal [21]. The green region is characterised by a *trivial* landscape, i.e. all the spurious minima disappear [20]. The grey region is where gradient-flow succeeds to converge despite the presence of spurious minima. We marked with black crosses points predicting the gradient-flow threshold obtained numerically in [20], they perfectly agree with our theoretical prediction of the threshold (5), marked by the grey dashed line. The circles in colours are points that we will use to illustrate the different features of these regions.

2 Probing the Landscape by the Kac-Rice Method

The statistical properties of the landscape associated to the loss function (3) can be studied by the Kac-Rice method, which traces back to the statistical physics literature, see [27] for an overview, and was developed mathematically in [13, 14] and recently extended in [22].

The quantities of interest are the number of critical points at a given energy, $\mathcal{N}(\epsilon_p, \epsilon_2)$, and the Hessian matrix evaluated at those critical points. We analyse the logarithm of $\mathcal{N}(\epsilon_p, \epsilon_2)$, called the *complexity*. Since the complexity is a random quantity we compute its upper bound $\Sigma_a(\epsilon_p, \epsilon_2) = \ln \mathbb{E}[\mathcal{N}(\epsilon_p, \epsilon_2)]$, along the lines of [16, 20]. We have also computed its typical value $\Sigma_q(\epsilon_p, \epsilon_2) = \mathbb{E}[\ln \mathcal{N}(\epsilon_p, \epsilon_2)]$ along the lines of [22], i.e. non-rigorously using the replica symmetry assumption (see SM Sec. A). In what follows we focus on complexity of stationary points with no correlation with the signal, in which case analytical and numerical arguments (see SM Sec. A.2.1) indicate that $\Sigma_a(e_p, e_2)$ and $\Sigma_q(e_p, e_2)$ are either very close numerically or possibly equal. Thus, in the following, we will simply refer to the complexity $\Sigma(\epsilon_p, \epsilon_2)$ without further specification.

In the Kac-Rice analysis the statistics of the Hessian, $\mathcal{H}$, of critical points plays a key role. It was shown in [20], and the argumentation is reproduced in the SM Sec. A.2, that $\mathcal{H}$ has a simple form for the loss (3). It is a $(N-1) \times (N-1)$ matrix formed by the sum of three contributions: a random matrix $\mathbb{W}_{N-1}$ belonging to the Gaussian orthogonal ensemble (GOE), a matrix proportional to the identity, and a rank one projector in the direction of the signal. The expression of $\mathcal{H}$ for critical points with null overlap m with the signal and with energies ϵ_p and ϵ_2 reads:

$$\mathcal{H} = \sqrt{Q''(1)} \left[\mathbb{W}_{N-1} + t\, \mathbb{I}_{N-1} - \theta\, \boldsymbol{e}_1 \boldsymbol{e}_1^T \right] \tag{6}$$

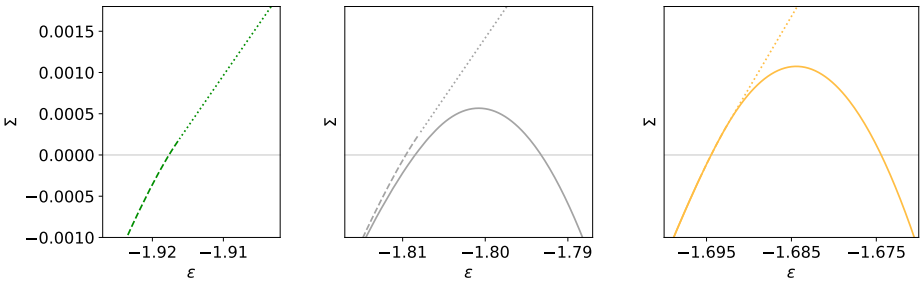


Figure 3: Complexity curves for the number of critical points for an overlap value $m = 0$ at fixed $\Delta_p = 1.0$ for (from left to right) $1/\Delta_2 = 2.7, 2.3, 1.9$. The lines are dotted when the complexity is dominate by critical points having an extensive number of eigenvalues are negative, dashed when only one eigenvalue is negative, full when the points have only positive (or null) eigenvalues i.e. they are minima. The complexity of the minima is drawn in full lines with the same colours, and it merges with the complexity of stationary points when it becomes dominant.

with $Q(x) = \frac{x^p}{p\Delta_p} + \frac{x^2}{2\Delta_2}$, $t = -\left(p\epsilon_p + 2\epsilon_2\right)/\sqrt{Q''(1)}$, and $\theta = Q''(0)/\sqrt{Q''(1)}$. The normalisation of $\mathbb{W}_{N-1}$ is chosen such that $\mathrm{Tr}\mathbb{E}[\mathbb{W}_{N-1}^2] = 1$.

The Fate of the Spurious: The initial condition for the gradient-flow algorithm is a random configuration $\boldsymbol{\sigma}_0$ uniformly drawn on the hyper-sphere. Such an initial condition clearly belongs to the large manifold of configurations uncorrelated with the ground-truth signal. We aim to investigate how does the gradient flow manage to escape from this initial manifold. For this purpose we focus on the properties of the landscape in the subspace where the overlap with the signal is zero, $m = 0$.

In Fig. 3, we plot the complexity at $m = 0$ as a function of the energy ϵ

$$\Sigma(\epsilon) = \sup_{\substack{\epsilon_p,\epsilon_2 \\ \text{s.t. } \epsilon_p+\epsilon_2=\epsilon}} \Sigma(\epsilon_p, \epsilon_2)\Big|_{m=0}$$

for the points $1/\Delta_2 = 1.9, 2.3, 2.7$ and $\Delta_p = 1.0$ ($p = 3$), which are marked with circles of the corresponding colour in Fig. 2. We use discontinuous lines for the complexity of critical points that have at least one negative direction, and full lines for the complexity of local minima. A finding of [20], that holds for any value of Δ_p, is that for small $1/\Delta_2$ the majority of critical points with zero overlap with the signal at low enough energies are spurious minima; they disappear increasing $1/\Delta_2$ above a Δ_p-dependent value $1/\Delta_2^{\text{triv}}$ corresponding to the green region of Fig. 2. In this part of the phase diagram, there are no spurious minima and the global minimum is correlated with the signal; this is an "easy" landscape for gradient flow which is therefore expected to succeed there. The main open question concerns the behavior for smaller values of $1/\Delta_2$: When does the existence of spurious minima, appearing in panel (b) and (c) of Fig. 3, start to be harmful to gradient flow?

In order to answer this question, we investigate more closely the nature of the spurious minima at different energies. We focus in particular on their Hessian, which plays a crucial role in order to understand which spurious minima have the largest basin of attraction and, hence, can trap the algorithm. For low signal-to-noise ratio, large Δ_p and large Δ_2, the spectrum of (6) is a shifted Wigner semicircle with support $[\sqrt{Q''(1)}(-2+t), \sqrt{Q''(1)}(2+t)]$. The effect of the projector (third contribution to the RHS of (6)) on the support of the spectrum is negligible as long as $\theta \leq 1$, as follows from the work on low-rank perturbations of random GOE matrices [23]. Moreover, the most numerous critical points at fixed energy ϵ are characterized by a $t(\epsilon)$ that is a monotonously decreasing function of ϵ, see Fig. 6 in the SM. Thus, moving towards higher energies, the spectrum of the Hessian shifts to the left, which indicates smaller curvature and wider minima. The transition between minima and saddles takes place at the *threshold energy* at which $t(\epsilon_{\text{th}}) = 2$, i.e. where the left edge of the Wigner semi-circle law touches zero, the numerical value is obtained in the Appendix Sec. A.2.3. Above this energy, critical points have an extensive number of downward directions, as found also in spin-glass models [14, 27]. Putting the above findings together, minima at $\epsilon = \epsilon_{\text{th}}$ are the most numerous and *the marginally stable ones*. Therefore, they

are the natural candidates for having the largest basin of attraction and the highest influence on the randomly initialised algorithm. This reasonable guess is at the basis of the theory of glassy dynamics in physics [25]. We take it as a working hypothesis for now, and we confirm it analytically and numerically in what follows.

Going back to the algorithm, when the signal-to-noise ratio is small we therefore expect that the configuration $\boldsymbol{\sigma}(t)$ slowly approaches at long times the ubiquitous "threshold minima" characterised by energy $\epsilon_{\rm th}$ and zero overlap with the signal. The last missing piece is unveiling what makes those minima unstable for large snr. We show below that it is a transition, called BBP (Baik-Ben Arous-Péché) [23], which takes place in the spectrum of the Hessian. Increasing $1/\Delta_2$ at fixed Δ_p, as in Fig. 3, leads to a larger θ. When θ becomes larger than one, an eigenvalue, equal to $\sqrt{Q''(1)}\,(-\theta-\theta^{-1}+t)$ pops out on the left of the Wigner semi-circle, and its corresponding eigenvector develops a finite overlap with the signal [23]. This implies the development of an unstable direction for the threshold minima trapping the dynamics, as they were already at the verge of instability. Indeed, as soon as the isolated eigenvalue pops out, an unstable downward direction towards the signal emerges in the landscape around them, at which point the algorithmic threshold for gradient flow takes place. Interestingly, many other spurious minima at lower energy also undergo the BBP transition, but they remain stable for longer as the isolated eigenvalue is positive when it pops out from the semi-circle. In conclusion, our analysis of the landscape suggests a dynamical transition for signal estimation by gradient flow given by

$$\theta = Q''(0)/\sqrt{Q''(1)} = 1 \tag{7}$$

which leads to a very simple expression for the transition line $\Delta_2^{\rm GF}$, Eq. (5). This theoretical prediction is shown in Fig. 2 as a dashed grey line: The agreement with the numerical estimation from [20] (black crosses) is perfect.

Our analysis unveils that the key property of the loss-landscape determining the performance of the gradient-flow algorithm, is the (in)stability in the direction of the signal of the minima with largest basin of attraction. These are the most numerous and the highest in energy, a condition that likely holds for many high-dimensional estimation problems.

The other spurious minima, which are potentially more trapping than the threshold ones and still stable at the algorithmic transition just derived, are actually completely innocuous since a random initial condition does not lie in their basin of attraction with probability one in the large N limit. This benign role of very bad spurious minima might appear surprising; it is due to the high-dimensionality of the non-convex loss function. Indeed it does not happen in finite dimensional cases, in which a random initial condition has instead a finite probability to fall into bad minima if those are present.

3 Probing the Gradient-Flow Dynamics

3.1 Closed-Form Dynamical Equations

In the large N limit gradient-flow dynamics for the spiked matrix-tensor model can be analysed using techniques originally developed in statistical physics studies of spin-glasses [28–30] and later put on a rigorous basis in [26]. Three observables play a key role in this theory:

(i) The overlap (or correlation) of the estimator at two different times: $C(t,t') = \langle \boldsymbol{\sigma}(t), \boldsymbol{\sigma}(t') \rangle$.

(ii) The change (or response) of the estimator at time t due to an infinitesimal perturbation in the loss at time t', i.e. $\ell \to \ell + \langle \boldsymbol{\sigma}(t'), \boldsymbol{h}(t') \rangle$ in Eq. (4): $R(t,t') = \sum_{i=1}^{N} \frac{\delta \sigma_i(t)}{\delta h_i(t')}\Big|_{h_i=0}$.

(iii) The average overlap of the estimator with the ground truth $m(t) = \langle \boldsymbol{\sigma}^*, \boldsymbol{\sigma}(t) \rangle$.

For $N \to \infty$ the above quantities converge to a non-fluctuating limit, i.e. they concentrate with respect to the randomness in the initial condition and in the generative process, and satisfy closed equations. Following works of Crisanti-Horner-Sommers-Cugliandolo-Kurchan (CHSCK) [29, 30] and their recent extension to the

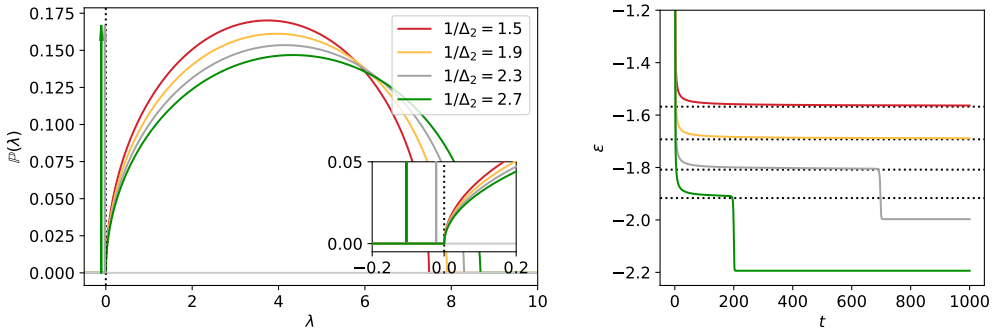


Figure 4: Right panel: energy as a function of time for the set of parameters indicated by small circles in Fig. 2. The horizontal dotted lines correspond to value of the threshold energy $\epsilon_{\rm th}$, as derived both from the Kac-Rice approach in Appendix Sec. A.2.3 and from the large time behaviour of the dynamics in Appendix Sec. B.2.6. Left panel: Eigenvalue distribution of the Hessian of the threshold states for the same set of parameters. When $1/\Delta_2$ becomes smaller than 2 an isolated eigenvalue appears; it has been highlighted using vertical arrows. Concomitantly, the energy as a function of time first approaches the plateau and eventually departs from it and reaches the energy of the global minimum.

spiked matrix-tensor model [20, 21] the above quantities satisfy:

$$\frac{\partial}{\partial t}C(t,t') = -\mu(t)\,C(t,t') + Q'(m(t))m(t') + \int_0^t R(t,t'')Q''(C(t,t''))C(t',t'')dt'' + \int_0^{t'} R(t',t'')Q'(C(t,t''))dt'', \tag{8}$$

$$\frac{\partial}{\partial t}R(t,t') = -\mu(t)\,R(t,t') + \int_{t'}^t R(t,t'')Q''(C(t,t''))R(t'',t')dt'', \tag{9}$$

$$\frac{d}{dt}m(t) = -\mu(t)\,m(t) + Q'(m(t)) + \int_0^t R(t,t'')m(t'')Q''(C(t,t''))dt'', \tag{10}$$

$$\mu(t) = Q'(m(t))m(t) + \int_0^t R(t,t'')\left[Q'(C(t,t'')) + Q''(C(t,t''))\,C(t,t'')\right]dt'', \tag{11}$$

with initial conditions $C(t,t) = 1\ \forall t$ and $R(t,t') = 0$ for all $t < t'$ and $\lim_{t'\to t^-} R(t,t') = 1\ \forall t$. The additional function $\mu(t)$, and its associated equation, are due to the spherical constraint; $\mu(t)$ plays the role of a Lagrange multiplier and guarantees that the solution of the previous equations is such that $C(t,t) = 1$. The derivation of these equations can be found in [21] and in the SM Sec. B. It is obtained using heuristic theoretical physics approach and can be very plausibly made fully rigorous generalising the work of [26, 31].

This set of equations can be solved numerically as described in [21]. The numerical estimation of the algorithmic threshold of gradient-flow, reproduced in Fig. 2, was obtained in [20]. We have also directly simulated the gradient flow Eq. (4) and compare the result to the one obtained from solving Eqs. (8-11). As shown in the SM Sec. C, for $N = 65535$, we find a very good agreement even for this large yet finite size.

Surfing on saddles: Armed with the dynamical equations, we now confirm the prediction of the threshold (5) based on the Kac-Rice-type of landscape analysis. In the SM we check that the minima trapping the dynamics are indeed the marginally stable ones ($t = 2$), see Figs. 7 and 8 in the SM, and we show the energy can be expressed in terms of C, R and m. In the right panel of Fig. 4 we then plot the energy as a function of time obtained from the numerical solution of Eqs. (8-11) for $1/\Delta_2 = 1.5, 1.9, 2.3, 2.7$ and $\Delta_p = 1$ (same points and colour code of Figs. 2 and 3). For the two smaller values of $1/\Delta_2$ the energy converges to a plateau value at $\epsilon_{\rm th}$ (dotted line), whereas for $1/\Delta_2 = 2.3, 2.7$ the energy plateaus close to $\epsilon_{\rm th}$ but then eventually drifts away and reaches a lower value, corresponding to the global minimum correlated with the signal. This behaviour can be understood in terms of the spectral properties of the Hessian (6) of the minima trapping the dynamics. In the left

panel of Fig. 4 we plot the corresponding density of eigenvalues of $\mathcal{H}$ for the same values of $1/\Delta_2$ and Δ_p used in the right panel. This is an illustration of the dynamical phenomenon explained in the previous section: when the signal-to-noise ratio is large enough threshold minima become unstable because a negative eigenvalue, associated to a downward direction toward the signal, emerges. In this case $\boldsymbol{\sigma}(t)$ first seems to converge to the threshold minima and then, at long times, drifts away along the unstable direction. The larger is the signal-to-noise ratio the more unstable is the downward direction and, hence, the shortest is the intermediate trapping time.

3.2 Gradient-flow Threshold from Dynamical Theory

We now show that the very same prediction (5) for the algorithmic threshold of gradient-flow can be directly obtained analysing the dynamical equations (8-11), without directly using results from the Kac-Rice analysis, thus establishing a firm and novel connection between the behaviour of the gradient-flow algorithm and Kac-Rice landscape approaches.

For small signal-to-noise ratios, when m remains zero at all times, the dynamical equations (8-11) are identicall to the well-known one in spin glasses theory, for reviews see [32,33]. These equations have been studied extensively for decades in statistical physics and a range of results about their behaviour has been established. Here we describe the results which are important for our analysis and devote the SM Sec. B.2 to a more extended presentation. It was shown analytically in [29] that the behaviour of the dynamics at large times is captured by an asymptotic solution of Eqs. (8-11) that verifies several remarkable properties. The ones of interest to us are that for t and t' large:

(i) $C(t,t') = 1$ when $t - t'$ finite; $C(t,t')$ becomes less than one when $t - t'$ diverges with t and t'.

(ii) $R(t,t') = R_{\rm TTI}(t-t') + R_{\rm ag}(t,t')$, where TTI stands for time-translational-invariance, ag stand for aging. Here $R_{\rm TTI}(t-t')$ goes to zero on a finite time-scale, whereas $R_{\rm ag}(t,t')$ varies on timescales diverging with t and t'. Moreover, $R_{\rm ag}(t,t')$ verifies the so called "weak-long term memory" property: for any finite t_0, $\int_{t-t_0}^{t} R_{\rm ag}(t,t'')dt''$ is arbitrarily small. We refer to this function form for $R(t,t')$ as the *aging ansatz*, adopting the physics terminology.

These properties are confirmed to hold by our numerical solution, see for instance Fig. 9 in the SM. The interpretation of these dynamical properties is that at long times $\boldsymbol{\sigma}(t)$ decreases in the energy landscape and approaches the marginally stable minima. Concomitantly, dynamics slows down and takes place along the almost flat directions associated to the vanishing eigenvalues of the Hessian.

We remind that in the previous paragraphs we assume null correlation with the signal, $m = 0$. In order to find the algorithmic threshold beyond which the gradient-flow develops a positive correlation, we study the instability of the aging solution as a function of the signal-to-noise ratio. Our strategy is to start with an arbitrarily small overlap, $m(0) = \delta$, and determine whether it grows at long times thus indicating an instability towards the signal. Since the initial condition for the overlap is uncorrelated with the signal, then, for sufficiently small δ, C and R reach their asymptotic form before m becomes of order one. We can thus plug the asymptotic aging ansatz for R in the dynamical equation for m:

$$\begin{aligned}\frac{d}{dt}m(t) = -\mu(t)m(t) + Q'(m(t)) + \int_0^t R_{\rm TTI}(t-t'')Q''(1)m(t'')dt'' + \\ + \int_0^t R_{\rm ag}(t,t'')Q''(C_{\rm ag}(t,t''))m(t'')dt''\end{aligned} \tag{12}$$

In the linear approximation the solution has the form $m(t) = \delta\exp(\Lambda t)$ and we assume Λ arbitrarily small since we want to find the algorithmic threshold where $\Lambda = 0$. The term $Q'(m(t))$ becomes $Q''(0)m(t)$. Since $m(t)$ has an arbitrarily slow evolution, whereas $R_{\rm TTI}(t-t'')$ relaxes to zero on a finite timescale, the second term of the RHS of eq. (12) simplifies to:

$$\delta\exp(\Lambda t)Q''(1)\int_0^t R_{\rm TTI}(t-t'')\exp(-\Lambda(t-t''))dt'' \simeq m(t)Q''(1)\overline{R}$$

where $\overline{R} = \int_0^t R_{\rm TTI}(t-t'')dt''$ does not depend on t (since t can be taken arbitrarily large and $R_{\rm TTI}(t-t'')$ relaxes to zero on finite time-scales). The contribution of to the last term on (12) reads:

$$\delta \exp(\Lambda t)\int_0^t R_{\rm ag}(t,t'')Q''(C_{\rm ag}(t,t''))e^{-\Lambda(t-t'')}dt'' = m(t)\int_0^t R_{\rm ag}(t,t'')Q''(C_{\rm ag}(t,t''))e^{-\Lambda(t-t'')})dt''.$$

Using that $Q''(C_{\rm ag}(t,t''))$ is bounded by $Q''(1)$ and that Λ cuts-off the integral on a time $t_0 \sim 1/\Lambda$ that does not diverge with t, we can use the "weak-long term memory" property to conclude that the last term is arbitrarily small compared to $m(t)$ and hence can be neglected with respect to the previous ones. Collecting all the pieces together we find:

$$\frac{d}{dt}m(t) = \left[-\mu_\infty + Q''(0) + Q''(1)\overline{R}\right] m(t) + O(\delta^2)\,. \tag{13}$$

This is solved by $m(t) = \delta \exp(\Lambda t)$ with $\Lambda = -\mu_\infty + Q''(0) + Q''(1)\overline{R}$, which therefore justifies a posteriori our assumption of exponential growth. The condition for the instability of the aging solution towards the signal solution is therefore given by

$$0 = -\mu_\infty + Q''(0) + Q''(1)\overline{R}\,. \tag{14}$$

From the analysis of the asymptotic aging solution presented in SM Sec. B.2 one finds that $\mu_\infty = 2\sqrt{Q''(1)}$ and $\overline{R} = 1/\sqrt{Q''(1)}$, therefore obtaining $Q''(0) = \sqrt{Q''(1)}$. This condition is the same one found from the study of the landscape, and thus leads to the transition line eq. (5).

Acknowledgments

We thank Pierfrancesco Urbani for many related discussions. We acknowledge funding from the ERC under the European Union's Horizon 2020 Research and Innovation Programme Grant Agreement 714608-SMiLe; from the French National Research Agency (ANR) grant PAIL; and from the Simons Foundation (#454935, Giulio Biroli).

References

[1] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[2] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[3] 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.

[4] C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *ICLR 2017*, 2017. preprint arXiv:1611.01540.

[5] 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.

[6] 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.

[7] Simon S Du, Jason D Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima. In *International Conference on Machine Learning*, pages 1338–1347, 2018.

[8] Rong Ge and Tengyu Ma. On the optimization landscape of tensor decompositions. In *Advances in Neural Information Processing Systems*, pages 3653–3663, 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.

[10] Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.

[11] Shuyang Ling, Ruitu Xu, and Afonso S Bandeira. On the landscape of synchronization networks: A perspective from nonconvex optimization. *arXiv preprint arXiv:1809.11083*, 2018.

[12] David J Gross and Marc Mézard. The simplest spin glass. *Nuclear Physics B*, 240(4):431–452, 1984.

[13] 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.

[14] Antonio Auffinger, Gérard Ben Arous, and Jiří Černý. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201, 2013.

[15] Levent Sagun, V Ugur Guney, Gerard Ben Arous, and Yann LeCun. Explorations on high dimensional landscapes. *arXiv preprint arXiv:1412.6615*, 2014.

[16] Gerard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica. The landscape of the spiked tensor model. *arXiv preprint arXiv:1711.05424*, 2017.

[17] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[18] Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse PCA. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 2197–2201. IEEE, 2014.

[19] Emile Richard and Andrea Montanari. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

[20] Stefano Sarao Mannelli, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborova. Passed & spurious: Descent algorithms and local minima in spiked matrix-tensor models. In *International Conference on Machine Learning*, pages 4333–4342, 2019.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] Tommaso Castellani and Andrea Cavagna. Spin glass theory for pedestrians. *Journal of Statistical Mechanics: Theory and Experiment*, 2005:P05012, 2005.

[28] Andrea Crisanti and H-J Sommers. The spherical p-spin interaction spin glass model: the statics. *Zeitschrift für Physik B Condensed Matter*, 87(3):341–354, 1992.

[29] Leticia F Cugliandolo and Jorge Kurchan. On the out-of-equilibrium relaxation of the Sherrington-Kirkpatrick model. *Journal of Physics A: Mathematical and General*, 27(17):5749, 1994.

[30] Leticia F Cugliandolo and Jorge Kurchan. Weak ergodicity breaking in mean-field spin-glass models. *Philosophical Magazine B*, 71(4):501–514, 1995.

[31] Pompiliu Manuel Zamfir. Limiting dynamics for spherical models of spin glasses with magnetic field. *arXiv preprint arXiv:0806.3519*, 2008.

[32] Jean-Philippe Bouchaud, Leticia F Cugliandolo, Jorge Kurchan, and Marc Mézard. Out of equilibrium dynamics in spin-glasses and other glassy systems. *Spin glasses and random fields*, pages 161–223, 1998.

[33] Leticia F Cugliandolo. Course 7: Dynamics of glassy systems. In *Slow Relaxations and nonequilibrium dynamics in condensed matter*, pages 367–521. Springer, 2003.

[34] Eliran Subag. The complexity of spherical p-spin models—a second moment approach. *The Annals of Probability*, 45(5):3385–3450, 2017.

[35] Robert J Adler and Jonathan E Taylor. *Random fields and geometry*. Springer Science & Business Media, 2009.

[36] Andrea Crisanti and Luca Leuzzi. Spherical 2+ p spin-glass model: An exactly solvable model for glass to spin-glass transition. *Physical review letters*, 93(21):217203, 2004.

[37] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor PCA. *arXiv preprint arXiv:1808.00921*, 2018.

[38] A Barrat. The p-spin spherical spin glass model. *arXiv preprint cond-mat/9701031*, 1997.

[39] Jorge Kurchan and Laurent Laloux. Phase space geometry and slow dynamics. *Journal of Physics A: Mathematical and General*, 29(9):1929, 1996.

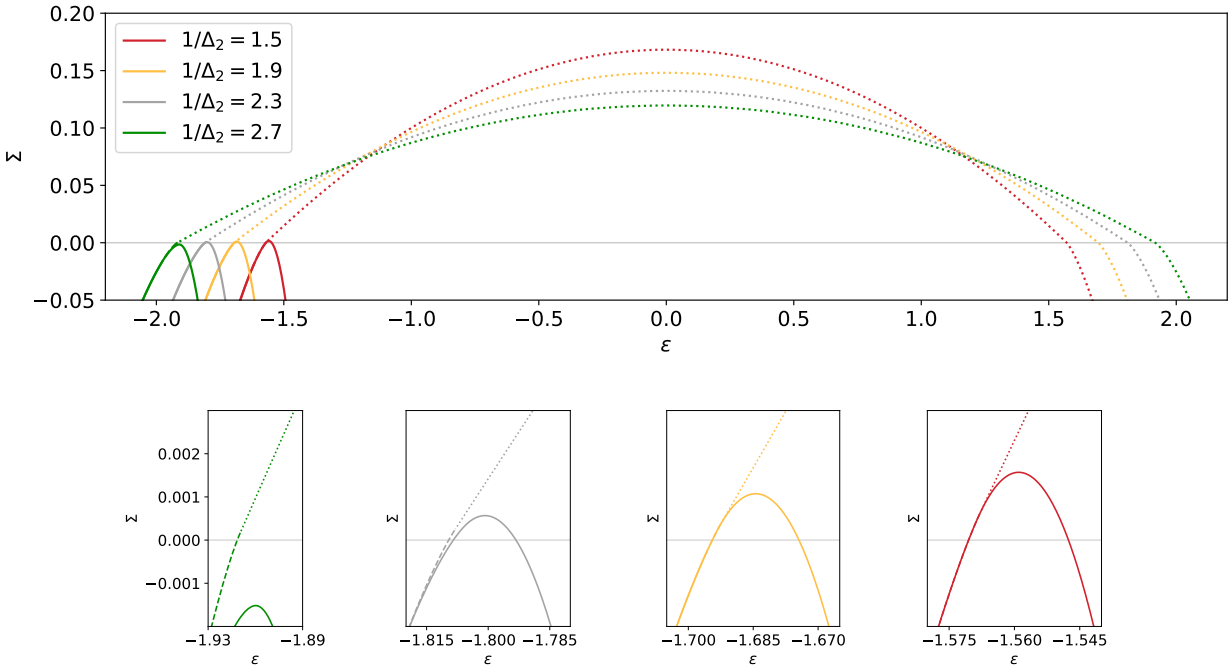
[40] Florent Krzakala and Lenka Zdeborová. Performance of simulated annealing in p-spin glasses. In *Journal of Physics: Conference Series*, volume 473, page 012022. IOP Publishing, 2013.

[41] Guilhem Semerjian, Leticia F Cugliandolo, and Andrea Montanari. On the stochastic dynamics of disordered spin models. *Journal of statistical physics*, 115(1-2):493–530, 2004.

A Kac-Rice method

A.1 Summary of the Kac-Rice complexity

In this section we introduce the Kac-Rice formula and we show how to reduce it to an explicit expression for the spiked matrix-tensor model. The Kac-Rice formula evaluates the expected number of critical points of a rough function subject to a number of conditions. For an inference problem it is interesting to focus on the



(a) (b) (c) (d)

Figure 5: Curves of the complexity of critical points, dotted and dashed curves from Eq. (17), and of minima, full curve from Eq. (19), at overlap value $m = 0$ at fixed $\Delta_p = 1.0$ for different Δ_2. The figure shows qualitatively the same features as Fig. 3, but displays the full positive part of the complexity for the four cases discussed in the main text, $1/\Delta_2 \in \{1.5, 1.9, 2.3, 2.7\}$. Zooms of the curves of the annealed complexity of critical points and minima when they cross zero at negative loss are in the panels labelled from (a) to (d) for increasing $1/\Delta_2$.

expected number of critical points constrained to have of given loss and a given overlap with the ground truth. For convenience reasons we consider the rescaled loss $\mathcal{L}(\boldsymbol{\sigma}) = N\ell(\boldsymbol{\sigma})$. The Kac-Rice formula then reads

$$\mathbb{E}_\eta[\mathcal{N}(\epsilon, m|\Theta)] = \int_{\mathbb{S}^{N-1}} \delta\left(\langle \boldsymbol{\sigma}, \boldsymbol{\sigma}^* \rangle - m\right) \mathbb{E}_\eta\left[|\det \mathcal{H}| \Big| \mathcal{L} = N\,\epsilon, \partial_i \mathcal{L} = 0\ \forall i, \lambda_{\min} > 0 \right] \times \\ \times\, \phi_{\mathcal{L}, \partial_i \mathcal{L}}(\boldsymbol{\sigma}, \mathbf{0}, \epsilon) d\boldsymbol{\sigma}\,, \tag{15}$$

where η represents the noise in the problem, Θ the parameters and ϕ the joint probability density of the loss and its gradient.

The quantity of interest is the density of the logarithm of the number of critical points $\log \mathcal{N}(\epsilon, m|\Theta)/N$. It should be noted that, since the random variable representing the number of critical point fluctuates at the exponential scale, a correct estimation of the expected value of this quantity is not $\log \mathbb{E}_\eta[\mathcal{N}(\epsilon, m|\Theta)]$, as it would be immediately obtained by using the result of the Kac-Rice formula [20], but $\mathbb{E}_\eta[\log \mathcal{N}(\epsilon, m|\Theta)]$. These two quantities are called respectively *annealed* and *quenched complexities*. Using Jensen inequality one observes that the annealed complexity is just an upper bound of the quenched one. However, for mathematical convenience most of the studies have been focused on the former. Eventually the second moment of the number of critical points has been evaluated [34], by an extension of the Kac-Rice formula to higher moments [35], just to prove that the two are equivalent in some models [34]. The quenched complexity has been evaluated in a related model in a non rigorous way by studying the n-th moment and applying replica trick, the so-called replicated Kac-Rice [22]. Given a random variable Y replica trick says

$$\mathbb{E}_\eta[\log Y] = \lim_{n \to 0^+} \frac{\mathbb{E}_\eta[Y^n] - 1}{n} \tag{16}$$

but instead of considering an arbitrary $n \in \mathbb{R}^+$, the study is done using $n \in \mathbb{N}$ and performing an analytic continuation of the result to 0^+. The replica trick has already been used in a plethora of applications and, although not rigorous, it was found correct in all naturally motivated cases that have been later approached by other techniques. An important mathematical literature has developed in order to understand the method.

In the next section we sketch the derivation of the quenched Kac-Rice and we provide all the information to determine the annealed one. Since the threshold is determined considering the configuration with arbitrarily small overlap $m \ll 1$, we focus on that case. Remarkably we found that as $m \to 0$ the quenched complexity computed within the replica symmetric (RS) approximation is equal to the annealed one[1]. We show that the corresponding Hessian is Eq. (6) in the main text, *i.e.* it is proportional to a GOE translated by t and perturbed by a rank n perturbation of strength θ that in the annealed case is of rank 1. Thus we find that the complexity for the stationary points is [20]

$$\begin{aligned}\Sigma_a^{\rm sta}(m,\epsilon|\Delta_p,\Delta_2) = \max_{\substack{\epsilon_p,\epsilon_2 \\ \text{s.t. } \epsilon_p+\epsilon_2=\epsilon}} & \frac{1}{2}\log\frac{Q''(1)}{Q'(1)} + \frac{1}{2}\log(1-m^2) - \frac{1}{2}\frac{(Q''(m))^2}{Q'(1)}(1-m^2) + \\ & - \frac{p\Delta_p}{2}\left(\epsilon_p + \frac{m^p}{p\Delta_p}\right)^2 - \Delta_2\left(\epsilon_2 + \frac{m^2}{2\Delta_2}\right)^2 + \Phi(t),\end{aligned} \tag{17}$$

with

$$\Phi(t) = \begin{cases} \frac{t^2}{4} & \text{if } |t| \le 2 \\ \frac{t^2}{4} + \log\left(\sqrt{\frac{t^2}{4}-1} + \frac{|t|}{2}\right) - \frac{|t|}{4}\sqrt{t^2-4} & \text{otherwise} \end{cases} \tag{18}$$

and $t = -\left(p\epsilon_p + 2\epsilon_2\right)/\sqrt{Q''(1)}$ as already introduced in the main text. Finally studying the eigenvalue of the Hessian to constrain them in the positive semi-axis, we find the complexity of minima [20]

$$\Sigma_a(m,\epsilon|\Delta_p,\Delta_2) = \Sigma_a^{\rm sta}(m,\epsilon_p,\epsilon_2|\Delta_p,\Delta_2) - L(\theta,t)\,. \tag{19}$$

with

$$L(\theta,t) = \begin{cases} \frac{1}{4}\int_{\theta+\frac{1}{\theta}}^{t}\sqrt{y^2-4}dy - \frac{\theta}{2}\left(t - \left(\theta+\frac{1}{\theta}\right)\right) + \frac{t^2-\left(\theta+\frac{1}{\theta}\right)^2}{8} & \theta > 1,\ 2 \le t < \frac{\theta^2+1}{\theta} \\ \infty & t < 2 \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

and $\theta = Q''(m)\,(1-m^2)/\sqrt{Q''(1)}$. In Fig. 5, we show the two complexities of the stationary points and of the minima in the parameter space discussed in the main text with discontinuous lines Eq. (17) and full lines Eq. (19), respectively. A positive complexity means an exponential number of critical points (minima). The region where exponentially many minima appear is highlighted in the small figures, showing the coexistence of exponentially many minima and saddles.

A.2 Derivation of the quenched complexity

We proceed with the computation of the quenched Kac-Rice complexity for the spiked matrix-tensor model, using replicated Kac-Rice prescription for the spiked pure-tensor model [22]. This implies, following replica trick Eq. (16), the evaluation of the n-th moment of number of minima using Kac-Rice formula which is given by [35]

$$\begin{aligned}\mathbb{E}_\eta[\mathcal{N}(\epsilon,m|\Theta)^n] = & \int_{\mathbb{S}^{N-1}}\cdots\int_{\mathbb{S}^{N-1}}\mathbb{E}_\eta\left[\left(\prod_{a=1}^{n}|\det\mathcal{H}[\boldsymbol{\sigma}^a]|\right)\middle|\forall b,c\ \mathcal{L}[\boldsymbol{\sigma}^b] = N\epsilon, \partial_i\mathcal{L}[\boldsymbol{\sigma}^c] = 0\ \forall i, \lambda_{\min} > 0\right] \\ & \times \phi_{\mathcal{L},\partial_i\mathcal{L}}\left(\{\boldsymbol{\sigma}^a\},\mathbf{0},\epsilon\right)\prod_{a=1}^{n}\delta\left(\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^*\rangle - m\right)d\boldsymbol{\sigma}^a\,.\end{aligned} \tag{21}$$

[1] We expect the RS approximation to be correct for $p = 3$ in the whole $(\Delta_p, 1/\Delta_2)$ phase diagram, and hence that quenched and annealed complexities coincide. For $p > 3$, results from replica theory [36] suggest that one needs to go beyond the RS approximation at least in some parts of the phase diagram.

Where ϕ is the joint probability density of the loss and its gradients evaluated on the n replicated configurations. We hereby sketch the main computation steps and present the results that are the most relevant for the theory presented in this paper, i.e. for $m = 0$. The details of the computation and further results for $m > 0$ will be presented in a dedicated work elsewhere.

It is convenient to consider free variables on $\mathbb{R}^N$ and constrain them using a Lagrange multiplier γ. Thus $\mathcal{L}(\boldsymbol{\sigma}^a) \mapsto \mathcal{L}(\boldsymbol{\sigma}^a) - \frac{\gamma}{2}\left(\sum_i (\sigma_i^a)^2 - 1\right)$. Using the fact that the gradient must be zero on the sphere, *i.e.* that $\nabla_i \mathcal{L}(\boldsymbol{\sigma}^a) - \gamma\, \sigma_i^a = 0 \quad \forall i$, we obtain a simple expression for the multiplier: $\gamma = \langle \boldsymbol{\nabla}\mathcal{L}(\boldsymbol{\sigma}^a), \boldsymbol{\sigma}^a \rangle$. Moreover by separating the two terms in the loss that represent the contribution of the two channels, $\mathcal{L} = \mathcal{L}_p + \mathcal{L}_2$, we define $\mathcal{L}_p = N\epsilon_p$ and $\mathcal{L}_2 = N\epsilon_2$ so to obtain for the multiplier the even simpler equation $\gamma = p\epsilon_p + 2\epsilon_2$. To take advantage of this simple formula, in the following we work with the contributions of the two channels to the loss function separately and we impose the constraint on their sum, $N\epsilon = N(\epsilon_p + \epsilon_2)$, only at the end.

The use of Cartesian coordinates allows us to evaluate easily the moments and covariances by means of standard derivatives.

$$\mathbb{E}_\eta \left[\mathcal{L}[\boldsymbol{\sigma}^a]\right] = -NQ(\langle \boldsymbol{\sigma}^a, \boldsymbol{\sigma}^* \rangle)\,, \tag{22}$$

$$\mathrm{Cov}\left[\mathcal{L}[\boldsymbol{\sigma}^a], \mathcal{L}[\boldsymbol{\sigma}^b]\right] = N\, Q\left(\langle \boldsymbol{\sigma}^a, \boldsymbol{\sigma}^b \rangle\right)\,. \tag{23}$$

Taking derivatives of these equation gives all the covariances of loss, gradient and Hessian. For instance, we can easily see that the covariance of the Hessian is given by:

$$\begin{aligned}
&\frac{\partial^4}{\partial\sigma_i^a \partial\sigma_j^a \partial\sigma_k^b \partial\sigma_l^b} \frac{\mathrm{Cov}\left[\mathcal{L}[\boldsymbol{\sigma}^a], \mathcal{L}[\boldsymbol{\sigma}^b]\right]}{N} = Q''''\left(\langle \boldsymbol{\sigma}^a, \boldsymbol{\sigma}^b \rangle\right) \langle \boldsymbol{\sigma}^b, \boldsymbol{e}_i^a \rangle \langle \boldsymbol{\sigma}^b, \boldsymbol{e}_j^a \rangle \langle \boldsymbol{\sigma}^a, \boldsymbol{e}_k^b \rangle \langle \boldsymbol{\sigma}^a, \boldsymbol{e}_l^b \rangle + \\
&+ Q'''\left(\langle \boldsymbol{\sigma}^a, \boldsymbol{\sigma}^b \rangle\right)\Big(\langle \boldsymbol{e}_i^a, \boldsymbol{e}_k^b \rangle \langle \boldsymbol{\sigma}^b, \boldsymbol{e}_j^a \rangle \langle \boldsymbol{\sigma}^a, \boldsymbol{e}_l^b \rangle + \langle \boldsymbol{e}_j^a, \boldsymbol{e}_k^b \rangle \langle \boldsymbol{\sigma}^b, \boldsymbol{e}_i^a \rangle \langle \boldsymbol{\sigma}^a, \boldsymbol{e}_l^b \rangle + \langle \boldsymbol{e}_i^a, \boldsymbol{e}_l^b \rangle \langle \boldsymbol{\sigma}^b, \boldsymbol{e}_j^a \rangle \langle \boldsymbol{\sigma}^a, \boldsymbol{e}_k^b \rangle + \\
&+ \langle \boldsymbol{e}_j^a, \boldsymbol{e}_l^b \rangle \langle \boldsymbol{\sigma}^b, \boldsymbol{e}_i^a \rangle \langle \boldsymbol{\sigma}^a, \boldsymbol{e}_k^b \rangle\Big) + Q''\left(\langle \boldsymbol{\sigma}^a, \boldsymbol{\sigma}^b \rangle\right)\Big(\langle \boldsymbol{e}_i^a, \boldsymbol{e}_k^b \rangle \langle \boldsymbol{e}_j^a, \boldsymbol{e}_l^b \rangle + \langle \boldsymbol{e}_i^a, \boldsymbol{e}_l^b \rangle \langle \boldsymbol{e}_j^a, \boldsymbol{e}_k^b \rangle \Big)\,,
\end{aligned} \tag{24}$$

where $\{\boldsymbol{e}_i^a\}_i$ and $\{\boldsymbol{e}_k^b\}_k$ are the reference frames associated to replica a and b respectively.

Remark 1 (annealed Hessian) *In particular notice that if $n = 1$ there is only one replica and using an orthogonal basis where the N-th direction is aligned with the replica and projecting on the sphere by discarding the last coordinates we obtain a simple expression:*

$$\frac{1}{N} Cov\left[\mathcal{H}_{ij}[\boldsymbol{\sigma}], \mathcal{H}_{kl}[\boldsymbol{\sigma}]\right] = Q''(1)\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) \tag{25}$$

with the delta representing Kronecker's deltas. This is the expression of a GOE. We can as well compute the mean deriving twice in the i-th and j-th coordinate. Following [16] we make another convenient choice for the basis imposing that the signal lies in the space spanned by the $\boldsymbol{e}_1$ and $\boldsymbol{e}_N = \boldsymbol{\sigma}$. This gives,

$$\frac{1}{N}\mathbb{E}\left[\mathcal{H}_{ij}[\boldsymbol{\sigma}]\right] = Q''(\langle \boldsymbol{\sigma}^a, \boldsymbol{\sigma}^* \rangle) \langle \boldsymbol{\sigma}^*, \boldsymbol{e}_i \rangle \langle \boldsymbol{\sigma}^*, \boldsymbol{e}_j \rangle = Q''(\langle \boldsymbol{\sigma}^a, \boldsymbol{\sigma}^* \rangle) \langle \boldsymbol{\sigma}^*, \boldsymbol{e}_i \rangle \langle \boldsymbol{\sigma}^*, \boldsymbol{e}_j \rangle \delta_{i1}\delta_{j1} \tag{26}$$

that, when $m = 0$, equals

$$\frac{1}{N}\mathbb{E}\left[\mathcal{H}[\boldsymbol{\sigma}]\right] = Q''(0)\, \boldsymbol{e}_1 \boldsymbol{e}_1^T\,. \tag{27}$$

Wrapping together Eq. (25), Eq. (27) and the expression for the Langrange multiplier that acts as a translation, we obtain the Hessian presented in the main text Eq. (6). Observe, however, that the Hessian in which we are interested in is not the simple Hessian of the loss but rather the Hessian of the loss conditioned to a given loss and a given gradient. Using Eq. (23) to compute the covariance of Hessian and loss, and of Hessian and gradient under this basis, we can observe that these random variables are unconditioned. Thus the conditioning does not affect the distribution of the Hessian of the loss and therefore Eq. (6) is recovered.

Eq. (22) and Eq. (23) are basic ingredients required to continue with the analysis. In the next two sections we first compute the joint density of the loss and its gradient, and second compute the expected value of the determinant of the Hessians. In the final section we put together the results obtaining the complexities already presented in the summary A.1.

A.2.1 Joint probability density.

In order to evaluate the joint probability density ϕ we focus on the covariance matrix of the loss and its gradient, that using Eq. (23) is given by:

$$\frac{1}{N}\left[\boldsymbol{C}_{\mathcal{L},\nabla\mathcal{L}}\right]^{a,b} = \begin{bmatrix} Q''\left(\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^b\rangle\right)\boldsymbol{\sigma}^a\otimes\boldsymbol{\sigma}^b + Q'\left(\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^b\rangle\right)\mathbb{I}_N & Q'\left(\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^b\rangle\right)\boldsymbol{\sigma}^{b,T} \\ Q'\left(\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^b\rangle\right)\boldsymbol{\sigma}^a & Q\left(\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^b\rangle\right) \end{bmatrix} . \tag{28}$$

The joint density corresponds to the probability of observing a zero gradient on the sphere and a given loss, $(\gamma\,\boldsymbol{\sigma}^T,\epsilon)^T$, in the multivariate Gaussian variable $(\boldsymbol{\nabla}\mathcal{L}^T,\mathcal{L})^T$. Thus taking into account the first moments of loss and gradient, obtained from Eq. (22), we define the auxiliary vector

$$[\boldsymbol{\mu}(\epsilon_p,\epsilon_2)]^a = \left((p\epsilon_p+2\epsilon_2)\boldsymbol{\sigma}^{a,T} + Q'(\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^*\rangle)\boldsymbol{\sigma}^{*,T},\ \epsilon + Q(\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^*\rangle)\right)^T . \tag{29}$$

The probability density is given by

$$\phi_{\mathcal{L},\,\partial_i\mathcal{L}}\left(\{\boldsymbol{\sigma}^a\},\mathbf{0},\epsilon\right) \propto \iint \delta(\epsilon-\epsilon_p-\epsilon_2)\exp\left[-\frac{1}{2}\sum_{a,b}\left[\boldsymbol{\mu}(\epsilon_p,\epsilon_2)\right]^{a,T}\left[\boldsymbol{C}^{-1}_{\mathcal{L},\nabla\mathcal{L}}\right]^{a,b}\left[\boldsymbol{\mu}(\epsilon_p,\epsilon_2)\right]^b\right] d\epsilon_p d\epsilon_2 . \tag{30}$$

This expression can be evaluated by observing that there is a set of $(N+1)n$-dimensional vectors that forms a closed group under the action of the covariance matrix Eq. (28). This set is composed by the following four vectors

$$\boldsymbol{\xi}_1^T = \left(\boldsymbol{\sigma}^{1,T},0,\boldsymbol{\sigma}^{2,T},0,\ldots,\boldsymbol{\sigma}^{n,T},0\right) , \tag{31}$$

$$\boldsymbol{\xi}_2^T = \left(\sum_{e\neq 1}\boldsymbol{\sigma}^{e,T},0,\sum_{e\neq 2}\boldsymbol{\sigma}^{e,T},0,\ldots,\sum_{e\neq n}\boldsymbol{\sigma}^{e,T},0\right) , \tag{32}$$

$$\boldsymbol{\xi}_3^T = \left(\mathbf{0}^T,1,\mathbf{0}^T,1,\ldots,\mathbf{0}^T,1\right) , \tag{33}$$

$$\boldsymbol{\xi}_4^T = \left(\boldsymbol{\sigma}^{*,T},0,\boldsymbol{\sigma}^{*,T},0,\ldots,\boldsymbol{\sigma}^{*,T},0\right) , \tag{34}$$

where $\mathbf{0}$ is an N dimensional null vector. Indeed the auxiliary vector can be rewritten in terms of the elements of this set of newly defined vectors as follows

$$[\boldsymbol{\mu}(\epsilon_p,\epsilon_2)]^a = (p\epsilon_p+2\epsilon_2)[\boldsymbol{\xi}_1]^a + (\epsilon+Q(\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^*\rangle)\,[\boldsymbol{\xi}_3]^a + Q'(\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^*\rangle)\,[\xi_4]^a . \tag{35}$$

At this point we exploit the fact that the set of these vectors forms a closed group under the action of the covariance matrix. In fact we can invert its action on the set $\{\boldsymbol{\xi}_k\}_{k=1}^4$ only, without the need to evaluate the inverse of the full covariance matrix. Using this trick, the integrand in Eq. (30) can be evaluated. The result for the integrand in Eq. (30) contains the dependence on the configurations of replicas only in terms of the overlaps $q_{a,b}=\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^b\rangle$ with each other, and of the overlap of each of them with the ground truth, *i.e.* the magnetisation $m_a=\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^*\rangle$. In this formulation, hence, the problem of evaluating a free integral over n vectors on the sphere has been translated into the task of evaluating an integral over the possible choices of the $n\times n$ matrix of the overlaps provided that we consider the multiplying factor that comes from the volume $V(\{q_{a,b}\},\{m_a\})$ of configurations that are compatible with that choice and the condition on the magnetisations.

The next step is to make an ansatz on the form of the matrix of these overlaps which must be consistent with the condition on the vector of magnetisations required in the Kac-Rice formula. The simplest ansatz is called *replica symmetric* ansatz and assumes that the overlaps of different replicas are independent of the indices a and b, i.e.

$$\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^b\rangle = \delta_{ab} + (1-\delta_{ab})\,q\,. \tag{36}$$

Note that the replica symmetric ansatz is compatible with the condition $\langle\boldsymbol{\sigma}^a,\boldsymbol{\sigma}^*\rangle = m\ \forall a$ imposed in the Kac-Rice formula. Within this ansatz the probability density can be evaluated as a function of q and m for arbitrary n and

the analytic continuation at $n \to 0^+$ can be finally taken to evaluate the quenched complexity. The expression for generic n is too long and convoluted to be reported here. However in the limit $n \to 1$ it corresponds to the expression of the probability density of losses and gradients evaluated in the annealed computation which has the following nice expression

$$\begin{aligned}
\phi_{\mathcal{L},\partial_i\mathcal{L}}\left(\{\boldsymbol{\sigma}^a\},\mathbf{0},\epsilon\right) &\propto \int\int \delta(\epsilon-\epsilon_p-\epsilon_2)\exp\left[-\frac{N}{2}\frac{\left(Q''(m)\right)^2}{Q'(1)}(1-m^2)\right]\times \\
&\times\exp\left[-\frac{Np}{2}\Delta_p\left(\epsilon_p+\frac{m^p}{p\Delta_p}\right)^2 - N\Delta_2\left(\epsilon_2+\frac{m^2}{2\Delta_2}\right)^2\right]d\epsilon_p d\epsilon_2 \\
&\simeq \max_{\substack{\epsilon_p,\epsilon_2 \\ \text{s.t. } \epsilon_p+\epsilon_2=\epsilon}} \exp\left[-\frac{N}{2}\frac{\left(Q''(m)\right)^2}{Q'(1)}(1-m^2) - \frac{Np}{2}\Delta_p\left(\epsilon_p+\frac{m^p}{p\Delta_p}\right)^2 - N\Delta_2\left(\epsilon_2+\frac{m^2}{2\Delta_2}\right)^2\right].
\end{aligned} \tag{37}$$

We must also consider the normalisation of the density that is given by

$$\exp\left[-\frac{Nn}{2}\log(2\pi Q'(1))\right]. \tag{38}$$

Finally we come back to the volume term $V(\{q_{a,b}\},\{m_a\})$. Constraining the configurations to a given overlap q with each other and m with the ground truth produces a volume term that can be easily evaluated as

$$V(q,m) \simeq \exp\left[\frac{Nn}{2}\left(\log\frac{2\pi e(1-q)}{N} - \frac{m^2-q}{1-q}\right)\right], \tag{39}$$

and for one single replica (which is useful in the computation of the annealed complexity) is simply

$$V(m) \simeq \exp\left[\frac{N}{2}\left(\log\frac{2\pi e}{N} + \log(1-m^2)\right)\right]. \tag{40}$$

Under the replica symmetric assumption we make a Laplace approximation that allows to evaluate the quenched complexity as an extremisation of the entire expression that depends on the overlap variable q through the volume term $V(q,m)$ and the probability density ϕ. An interesting remark concerns the limit $q \to 0$ in quenched joint probability density. Indeed in that case the two joint probability coincide module a factor n. We checked numerically that as $m \to 0$ the optimal q goes to 0, which implies that the equations of the annealed and quenched complexities do correspond on the equator $m = 0$.

A.2.2 Expected value of the Hessian.

As discussed introducing Eq. (23), the Hessian is a matrix-valued random variable with multivariate Gaussian distribution. In evaluating the Kac-Rice formula we must consider the distribution of the Hessian conditioned to the loss and its gradient, this can be imposed using the formula for conditioning of Gaussian random variables. Given $\boldsymbol{X},\boldsymbol{Y}$ Gaussian random variables with covariance $\boldsymbol{C}$ and mean $\boldsymbol{\mu}$ the distribution of $\boldsymbol{X}$ conditioned to $\boldsymbol{Y}=\boldsymbol{y}^*$ is still Gaussian with covariance and mean

$$\begin{aligned}
\boldsymbol{C}_{\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}^*} &= \boldsymbol{C}_{\boldsymbol{X}} - \boldsymbol{C}_{\boldsymbol{XY}}\boldsymbol{C}_{\boldsymbol{Y}}^{-1}\boldsymbol{C}_{\boldsymbol{XY}}\,; \\
\boldsymbol{\mu}_{\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}^*} &= \boldsymbol{\mu}_{\boldsymbol{X}} + \boldsymbol{C}_{\boldsymbol{XY}}\boldsymbol{C}_{\boldsymbol{Y}}^{-1}(\boldsymbol{y}^*-\boldsymbol{\mu}_{\boldsymbol{Y}})\,.
\end{aligned}$$

In the annealed case, by using Eq. (23) and the expression for the Langrange multiplier γ we get that the Hessian corresponds to a shifted GOE subject to a rank one perturbation, as already discussed in the main text (see Eq. (6)). In the replicated Kac-Rice formula a more complicated expression appears that depends on the product of the determinants of Hessians associated to different replicas. However, using a proper reference frame, it was already noticed [22, 34] that each Hessian corresponds also to a GOE since it is dominated by a

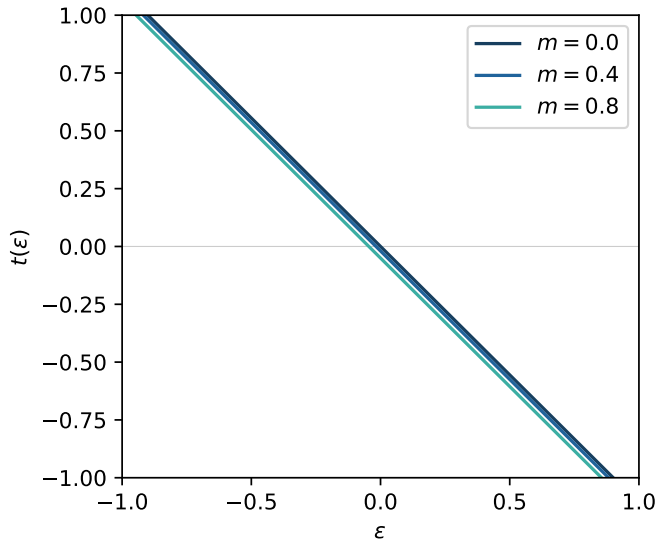


Figure 6: Shift of the Hessian, from Eq. (6), as function the loss density for different values of m. The qualitative behaviour shown in the figure does not change varying the parameter of the systems, i.e. it is always a decreasing function. The figure shows results obtained using $p = 3$, $\Delta_p = 1.0$ and $1/\Delta_2 = 2.3$.

$(N-n) \times (N-n)$ GOE block as $n \ll N$. Moreover it has also been shown [22,34] that the expectation value of the product the Hessian determinants is equivalent to the product of the expectation values of each determinant. Thus we can still use standard results on the spectrum of GOE random matrices to evaluate the term in the Kac-Rice that depends on the Hessian. The distribution of the spectrum of the eigenvalue is given by

$$\rho(\lambda)d\lambda = \frac{\sqrt{4Q''(1) - (\lambda+\gamma)^2}}{2\pi Q''(1)} d\lambda \,, \tag{41}$$

thus the determinant is given by

$$\int \rho(\lambda) \log|\lambda| \, d\lambda = \frac{1}{2}\log[2Q''(1)] + \frac{1}{\pi}\int \sqrt{2-\lambda^2} \log\left|\lambda - \gamma/\sqrt{Q''(1)}\right| d\lambda \,. \tag{42}$$

where we recognise $t = \gamma/\sqrt{Q''(1)}$. After some algebra we find:

$$\frac{1}{2}\log[Q''(1)] - \frac{1}{2} + \Phi(t) \tag{43}$$

with $\Phi(t)$ defined in Eq. (18).

A.2.3 Complexities: Putting pieces together

By putting the above pieces together we obtain the annealed complexity of stationary points Eq. (17) where we can finally distinguish the origin of the various terms: the first term comes from the normalisation of the density and the determinant of the Hessian, the second comes from the volume prefactor, the third, fourth and fifth terms are originated by the probability density of loss and gradient and the last term comes from the product of Hessians.

In order to select only the minima in the study of the complexity we must impose that the smallest eigenvalue is positive. There are two possible scenarios: either the smallest eigenvalue is determined by the left edge of the bulk of the spectrum (the perturbation does not induces any BBP transition), or it is outside the bulk of the spectrum. In the first case the probability that the smallest eigenvalue is positive is suppressed by a factor

e^{-N^2} and the corresponding large deviation function is infinite. In the second case the large deviation function associated to the shift in the position of the smallest eigenvalue, that would allow to keep it positive, is finite and must be evaluated. The problem can be addressed with a replica computation [22] and focuses only on the typical value of the eigenvalue, missing the large deviation function. As we already discussed as $m \to 0$ we found numerically that the overlap q that extremises the complexity is $q = 0$, which leads back to the annealed complexity as we have shown computing the density. Since the main focus in the paper is on the critical points at the equator we do not compute the isolated eigenvalue in a quenched approach but we rather use the large deviation function as it can be obtained in the annealed approximation [37] of which we report here the result. The condition on having a positive minimum eigenvalue suppresses the number of critical points by a factor $e^{-NL(\theta,t)}$, with $L(\theta, t)$ given in Eq. (20), that enters in Eq. (17) leading to Eq. (19).

Remark 2 (threshold loss) *Considering Eq.* (17) *and changing variables from* (ϵ_p, ϵ_2) *to* (ϵ, t) *in the maximisation. We can now specialise on the region with no overlap with the signal,* $m = 0$*, and consider that the Hessian of the loss touches the zero,* $t = 2$ *which simplifies the expression of the piecewise term* $\Phi(t)$*. The result is a quadratic equation in* t *which is equal to* 2 *because of the marginality condition and can be solved in the energy* ϵ*. The resulting energy is the Kac-Rice threshold energy,*

$$\epsilon_{\text{th}}^{\text{KR}} = \frac{1}{\sqrt{Q''(1)}Q'(1)} \left[\frac{(p-2)^2}{2p\Delta_2\Delta_p} - 2Q''(1)Q(1) \right] . \tag{44}$$

B CHSCK Equations

In this section we present a derivation of CHSCK equations for the spiked matrix-tensor model using the generating functional formalism and later the asymptotic analysis under the hypothesis presented in the main text. The starting point is the loss, Eq. (3), expliciting the observations, Eqs. (1-2),

$$\begin{aligned}\ell(\boldsymbol{\sigma}|\boldsymbol{T},\boldsymbol{Y}) = &-\frac{\sqrt{(p-1)!}}{\Delta_p\sqrt{N}}\sum_{i_1<\cdots<i_p}\eta_{i_1\dots i_p}\sigma_{i_1}\dots\sigma_{i_p} - \frac{1}{p\Delta_p}\langle\boldsymbol{\sigma},\boldsymbol{\sigma}^*\rangle^p + \\ &- \frac{1}{\Delta_2\sqrt{N}}\sum_{i<j}\eta_{ij}\sigma_i\sigma_j - \frac{1}{2\Delta_2}\langle\boldsymbol{\sigma},\boldsymbol{\sigma}^*\rangle^2\end{aligned} \tag{45}$$

and the gradient flow Eq. (4) that for mathematical convenience we associate to an auxiliary function $\boldsymbol{f}(\boldsymbol{\sigma})$

$$\dot{\sigma}_i(t) = -\mu(t)\sigma(t) - \frac{\partial\ell(\boldsymbol{\sigma}(t)|\boldsymbol{T},\boldsymbol{Y})}{\partial\sigma_i(t)} \doteq f_i\left(\boldsymbol{\sigma}(t)\right) . \tag{46}$$

The next section shows in detail the derivation that proceeds by introducing a probability distribution for the different evolutions, or trajectories, of the dynamics at a fixed realization of the noise. Then the distribution is averaged over the noise and this implies some technical steps before obtaining the final form. The resulting distribution is used to average correlation, response function and magnetisation over all the trajectories giving the CHSCK Eqs. (8-11). In the analysis an important role is played by the normalisation constant of the distribution of the trajectories, that is used in the final steps to derive with simplicity the equations.

After deriving the equations we show how to apply the hypothesis on the large time behaviour of t and t' to the CHSCK Eqs. In the last part this analysis provides the constants $\overline{R}$ and μ_∞ used to the derive the threshold in the main text, and some interesting additional information, such as the value of the loss at the threshold shown in the right panel of Fig. 4.

B.1 Derivation of CHSCK Equations

The first step is to discretise the time in M time steps of length h. We want the trajectories to be a solution at every time step of Eq. (4), which discretized looks as $\sigma_i^{a+1} - \sigma_i^a = f_i\left(\boldsymbol{\sigma}^a\right)h$ with a the time index. Let's call y^{a+1} a solution to this equation. We can define the probability density of observing a trajectory satisfying Eq. (4) at a fixed noise:

$$p(\boldsymbol{\sigma}^1,\dots,\boldsymbol{\sigma}^M) = \int\prod_{ai}\delta\left(\sigma_i^{a+1} - y_i^a(\boldsymbol{\sigma}^a)\right)\prod_{a=0}^{M-1}d\mu_{\mathbb{S}}^a\,. \tag{47}$$

where $\mu_{\mathbb{S}}$ is the measure over $\mathbb{S}^{N-1}$.

The normalisation constant is the integral of this probability and is called *generating functional* $\mathcal{Z}$ and since the previous object is already properly normalised it is equal to 1. Rewriting the δs as Fourier transforms and therefore including the auxiliary variables $\tilde{\boldsymbol{\sigma}}^a$,

$$1 = \int\prod_{ai}\exp\left[N\tilde{\sigma}_i^a\left(\sigma_i^{a+1} - \sigma_i^a - f_i(\boldsymbol{\sigma}^a)h\right)\right]\prod_{a=0}^{M-1}d\mu_{\mathbb{S}}^a\frac{d\tilde{\boldsymbol{\sigma}}^a}{2\pi i} \tag{48}$$

where in order to have mathematically well-defined quantities in the large N limit we have a factor in the exponential. Moving to the continuum, the generating functional appears as a path integral

$$1 = \mathcal{Z} = \int\mathcal{D}\left[\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}\right]\prod_i\exp\left[N\int\tilde{\sigma}_i(t)\left(\partial_t\sigma_i(t) - f_i(\boldsymbol{\sigma}(t))\right)dt\right] . \tag{49}$$

So far the object we derived is a distribution that tells whether a trajectory from arbitrary initial condition respects or not gradient-flow dynamics, however, our interest is in average trajectories with respect to the

realization of the disorder. Therefore the distribution has to be averaged and after some algebraic manipulation gives the average generating functional in Eq. (51),

$$
\begin{aligned}
1 = \mathbb{E}_\eta\left[\mathcal{Z}\right] = \int \mathcal{D}\left[\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}\right] \prod_i \exp\left[N\int \tilde{\sigma}_i(t)\left(\partial_t\sigma_i(t) + \mu(t)\sigma_i(t) - Q'(\langle\boldsymbol{\sigma}(t),\boldsymbol{\sigma}^*\rangle^{p-1})\sigma_i^*\right)dt\right]\times \\
\times\,\mathbb{E}_\eta\left\{\prod_i \exp\left[-\int\tilde{\sigma}_i(t)\left(-\frac{\sqrt{N(p-1)!}}{\Delta_p}\sum_{i_1<\cdots<i_p}\eta_{i\,i_1\ldots i_{p-1}}\sigma_{i_1}\ldots\sigma_{i_{p-1}}\right)dt\right]\right\}\times \\
\times\,\mathbb{E}_\eta\left\{\prod_i \exp\left[-\int\tilde{\sigma}_i(t)\left(-\frac{\sqrt{N}}{\Delta_2}\sum_j\eta_{ij}\sigma_j\right)dt\right]\right\}.
\end{aligned}
\tag{50}
$$

In averaging over the η we need to be careful in grouping all the permutations of i with $i_1,\ldots,i_{p-1}$. For instance the exponent of the term in p is gives by

$$
\begin{aligned}
&-\frac{\sqrt{N(p-1)!}}{\Delta_p}\sum_{i_1<\cdots<i_p}\int\eta_{i_1\ldots i_p}\left(\tilde{\sigma}_{i_1}(t)\sigma_{i_2}(t)\ldots\sigma_{i_p}(t)+\cdots+\sigma_{i_1}(t)\sigma_{i_2}(t)\ldots\tilde{\sigma}_{i_p}(t)\right)dt \\
&=\frac{N(p-1)!}{2\Delta_p}\sum_{i_1<\cdots<i_p}\iint\left(\tilde{\sigma}_{i_1}(t)\sigma_{i_2}(t)\ldots\sigma_{i_p}(t)+\text{perm.}\right)\left(\tilde{\sigma}_{i_1}(t')\sigma_{i_2}(t')\ldots\sigma_{i_p}(t')+\text{perm.}\right)dtdt' \\
&=\frac{N}{2\Delta_p}\iint\Big(\langle\tilde{\boldsymbol{\sigma}}(t),\tilde{\boldsymbol{\sigma}}(t')\rangle\langle\boldsymbol{\sigma}(t),\boldsymbol{\sigma}(t)\rangle^{p-1}+(p-1)\langle\tilde{\boldsymbol{\sigma}}(t),\boldsymbol{\sigma}(t')\rangle\langle\boldsymbol{\sigma}(t),\tilde{\boldsymbol{\sigma}}(t')\rangle\langle\boldsymbol{\sigma}(t),\boldsymbol{\sigma}(t)\rangle^{p-2}\Big)dtdt'.
\end{aligned}
$$

This gives an action $\mathcal{S}[\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}]$ defined by

$$
\begin{aligned}
1 = \overline{\mathcal{Z}} = \int\mathcal{D}[\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}]e^{\mathcal{S}[\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}]} = \\
=\int\mathcal{D}\left[\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}\right]\prod_i\exp\left[N\int\tilde{\sigma}_i(t)\left(\partial_t\sigma_i(t)+\mu(t)\sigma_i(t)-Q'(\langle\boldsymbol{\sigma}(t),\boldsymbol{\sigma}^*\rangle^{p-1})\sigma_i^*\right)dt\right]\times \\
\times\exp\left[\frac{N}{2\Delta_p}\iint\langle\tilde{\boldsymbol{\sigma}}(t),\tilde{\boldsymbol{\sigma}}(t')\rangle\langle\boldsymbol{\sigma}(t),\boldsymbol{\sigma}(t)\rangle^{p-1}dtdt'\right]\times \\
\times\exp\left[\frac{N}{2\Delta_p}\iint(p-1)\langle\tilde{\boldsymbol{\sigma}}(t),\boldsymbol{\sigma}(t')\rangle\langle\boldsymbol{\sigma}(t),\tilde{\boldsymbol{\sigma}}(t')\rangle\langle\boldsymbol{\sigma}(t),\boldsymbol{\sigma}(t)\rangle^{p-2}dtdt'\right]\times \\
\times\exp\left[\frac{N}{2\Delta_2}\iint\left(\langle\tilde{\boldsymbol{\sigma}}(t),\tilde{\boldsymbol{\sigma}}(t')\rangle\langle\boldsymbol{\sigma}(t),\boldsymbol{\sigma}(t)\rangle+\langle\tilde{\boldsymbol{\sigma}}(t),\boldsymbol{\sigma}(t')\rangle\langle\boldsymbol{\sigma}(t),\tilde{\boldsymbol{\sigma}}(t')\rangle\right)dtdt'\right].
\end{aligned}
\tag{51}
$$

A simple way to proceed once evaluated the action was proposed in [38] and consists in taking the expectation with respect to the path distribution and exploiting simple identities together with integration by part:

$$
\begin{aligned}
0 = -\left\langle\frac{\delta\sigma_i(t')}{\delta\tilde{\sigma}_i(t)}\right\rangle_{\mathcal{S}} = \left\langle\sigma_i(t')\frac{\delta\mathcal{S}}{\delta\tilde{\sigma}_i(t)}\right\rangle_{\mathcal{S}} = \\
=N\Big\langle\partial_t\sigma_i(t)\sigma_i(t')+\mu(t)\sigma_i(t)\sigma_i(t')-Q'(\langle\boldsymbol{\sigma}(t),\boldsymbol{\sigma}^*\rangle^{p-1})\sigma_i^*\sigma_i(t')+ \\
+\frac{1}{\Delta_p}\int\left[\langle\boldsymbol{\sigma}(t),\boldsymbol{\sigma}(t)\rangle^{p-1}\tilde{\sigma}_i(t'')+(p-1)\langle\tilde{\boldsymbol{\sigma}}(t),\boldsymbol{\sigma}(t'')\rangle\langle\boldsymbol{\sigma}(t),\boldsymbol{\sigma}(t)\rangle^{p-2}\sigma_i(t'')\right]dt''+ \\
+\frac{1}{\Delta_2}\int\left[\langle\boldsymbol{\sigma}(t),\boldsymbol{\sigma}(t)\rangle\tilde{\sigma}_i(t'')+\langle\tilde{\boldsymbol{\sigma}}(t),\boldsymbol{\sigma}(t'')\rangle\sigma_i(t'')\right]dt''\Big\rangle_{\mathcal{S}}.
\end{aligned}
\tag{52}
$$

Finally, summing over the index i and dividing by N we recover Eq. (8). The remaining CHSCK Eqs. (9-10)

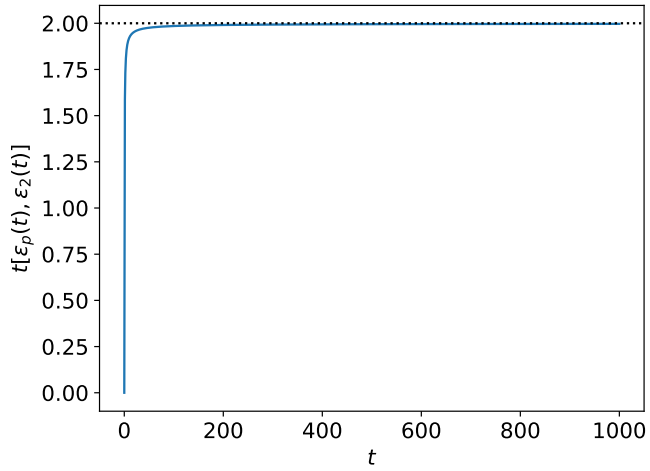


Figure 7: $t = -\left(p\epsilon_p(t) + 2\epsilon_2(t)\right)/\sqrt{Q''(1)}$ for $p = 3$, $\Delta_p = 1.0$ and $1/\Delta_2 = 1.9$ evaluated numerically from the CHSCK equations.

follow analogously from:

$$\delta(t-t') = \sum_i \left\langle \frac{\delta\tilde{\sigma}_i(t')}{\delta\tilde{\sigma}_i(t)} \right\rangle_{\mathcal{S}} = -\sum_i \left\langle \tilde{\sigma}_i(t') \frac{\delta S}{\delta\tilde{\sigma}_i(t)} \right\rangle_{\mathcal{S}} ; \tag{53}$$

$$0 = -\left\langle \frac{\delta\sigma_i^*}{\delta\tilde{\sigma}_i(t)} \right\rangle_{\mathcal{S}} = \left\langle \sigma_i^* \frac{\delta\mathcal{S}}{\delta\tilde{\sigma}_i(t)} \right\rangle_{\mathcal{S}} . \tag{54}$$

and Eq. (11) comes from imposing the spherical constrain, $C(t,t) = 1 \quad \forall t$, on Eq. (8).

In the following we are going to perform the analysis proposed by [29] in the present model. We need to consider Langevin dynamics instead of gradient-flow dynamics

$$\dot{\sigma}_i(t) = -\mu(t)\sigma(t) - \frac{\partial \ell(\boldsymbol{\sigma}(t)|\boldsymbol{T},\boldsymbol{Y})}{\partial\sigma_i(t)} + \frac{1}{\sqrt{N}}\eta_i^{(L)}(t)\,, \tag{55}$$

where the last term represents the Langevin noise, which is white Gaussian noise with moments: $\mathbb{E}_L[\eta_i^{(L)}(t)] = 0$ and $\mathbb{E}_L[\eta_i^{(L)}(t)\eta_j^{(L)}(t')] = 2T\delta_{ij}\delta(t-t')$ with T that has the physical meaning of temperature. The CHSCK equations are slightly modified,

$$\begin{aligned}\frac{\partial}{\partial t}C(t,t') &= TR(t',t) - \mu(t)\,C(t,t') + Q'(m(t))m(t') + \\ &+ \int_0^t R(t,t'')Q''(C(t,t''))C(t',t'')dt'' + \int_0^{t'} R(t',t'')Q'(C(t,t''))dt''\,,\end{aligned} \tag{56}$$

$$\frac{\partial}{\partial t}R(t,t') = -\mu(t)\,R(t,t') + \int_{t'}^t R(t,t'')Q''(C(t,t''))R(t'',t')dt''\,, \tag{57}$$

$$\frac{\partial}{\partial t}m(t) = -\mu(t)\,m(t) + Q'(m(t)) + \int_0^t R(t,t'')m(t'')Q''(C(t,t''))dt''\,, \tag{58}$$

$$\mu(t) = T + Q'(m(t))m(t) + \int_0^t R(t,t'')\left[Q'(C(t,t'')) + Q''(C(t,t''))C(t,t'')\right]dt''\,. \tag{59}$$

B.2 CHSCK Equations Separation of Time-Scales

The theory of glassy dynamics [25] is quite involved. We have therefore decided to show first some numerical results that directly confirm assumptions made in the main text, and then show in full glory that these assumptions can be obtained analytically from the theory.

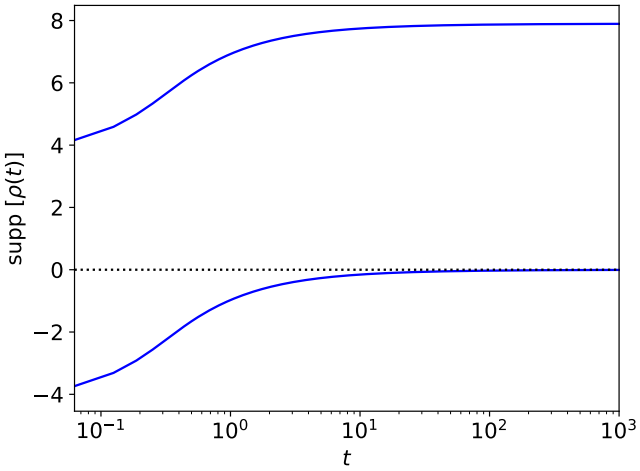


Figure 8: Support of the density of eigenvalues of the Hessian along the dynamics for $p = 3$, $\Delta_p = 1.0$ and $1/\Delta_2 = 1.9$. This figure illustrates the GF tends to the marginally stable minima for low signal-to-noise ratio.

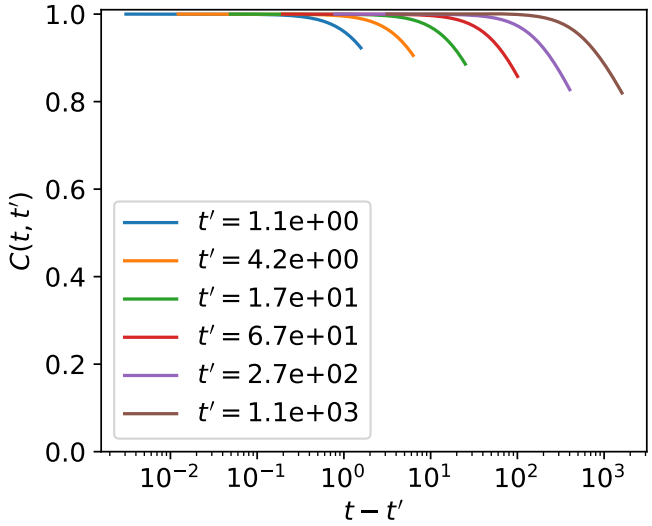


Figure 9: The correlation function $C(t, t')$ with $p = 3$, $\Delta_p = 1.0$ and $1/\Delta_2 = 1.5$ evaluated numerically from the CHSCK equations. The correlation is plotted as difference of the two times showing the as $t - t' \ll t, t'$ it remains close to 1. This shows as well that in this regime the correlation function shows time translational invariance.

B.2.1 Numerical tests

The first property that we wish to test is that for low signal-to-noise ratio GF is trapped in minima that are marginally stable. This can be checked computing from the CHSCK equations the evolution of $t = -\left(p\epsilon_p(t) + 2\epsilon_2(t)\right)/\sqrt{Q''(1)}$, the terms $\epsilon_p(t)$ and $\epsilon_2(t)$ can be expressed in terms of C, R, m as shown in Rmk. 6 of section B.2.6. In Fig. 7 we show that $t = -\left(p\epsilon_p(t) + 2\epsilon_2(t)\right)/\sqrt{Q''(1)}$ ($\epsilon_p(t)$ starts from zero at initial time and then converges to two. Thus the minima to which GF tends to at long times and for small signal-to-noise ratio are indeed the *marginally stable ones* characterized by a spectrum of the Hessian whose left edge touches zero. Actually, transferring the results obtained in the context of spin-glasses to our case [39], we know that as

long as m remains zero, i.e for small signal-to-noise ratio, the spectrum of the Hessian along the dynamics is a Wigner semi-circle with support $[\sqrt{Q''(1)}(-2+t), \sqrt{Q''(1)}(2+t)]$. We show the evolution of the support as a function of time in Fig. 8. This is another illustration of the fact that minima trapping GF are marginally stable.

The other point we wish to test is the assumption (i) made in the main text on $C(t,t')$, which we repeat here for convenience: $C(t,t')=1$ when $t-t'$ finite; $C(t,t')$ becomes less than one when $t-t'$ diverges with t and t'. We show in Fig. 9 the correlation function $C(t,t')$ as a function of $t-t'$ (in log-scale) for several values of t'. This is a good illustration of the "aging ansatz" defined in the main text.

Let us stress that these numerical tests where already done in the past on similar spin-glass models. We show them here in order to make the paper self-contained and so that the reader does not have to go back to physics literature. In the same vein, in the next sections we show the full theoretical analysis of the dynamical equations, which closely follows the theory of glassy dynamics developed in physics [25].

B.2.2 General aging ansatz

We study the behaviours of the dynamics at large times in the two-time regimes introduced in the main text (now generalized at finite temperature).

1. $t,t' \gg 1$ with $\frac{t-t'}{t} \to 0$, see Fig. 9. In this regime we have two important aspects: the two-times function depends only on the difference of the two times, $\tau = t-t'$, and we say that they respect time-translational invariance. Under this observation we redefine the two functions as $C(t,t') \to C_{\rm TTI}(\tau) \equiv C(t-t',0)$ and $R(t,t') \to R_{\rm TTI}(\tau) \equiv R(t-t',0)$. The second important aspect is the validity of the fluctuation-dissipation theorem (FDT) that links correlation and response function by the relation $R_{\rm TTI}(\tau) = -\frac{1}{T}\frac{dC_{\rm TTI}}{d\tau}(\tau)$ for τ positive.

2. $t,t' \gg 1$ with $\frac{t-t'}{t} = O(1)$. In this regime the relevant variable to consider is $\lambda = \frac{t'}{t}$. In reason of the "weak-long term memory" property it is useful to redefine rescale the response function and define $\mathcal{R}(\lambda) = tR(t,t')$. It is also convenient to consider the function $q\mathcal{C}(\lambda) = C(t,t')$ with $q = \lim_{\tau\to\infty} C_{\rm TTI}(\tau)$. Finally, in this regime a generalised version of the fluctuation-dissipation theorem holds $\mathcal{R}(\lambda) = \frac{1}{T} x q \frac{d\mathcal{C}(\lambda)}{d\lambda}$ where x is called *violation parameter*.

Under the (generalized) FDT the equations for correlation and response function that we obtain in the two-time regime collapse into a single equation. In the first regime we analyse only the correlation because of this link, while in the second regime we consider the two equations separately since we need to determine the violation parameter x.

In the analysis that follows we use massively the hypothesis of the two regimes to split the integrals and analyse them separately. We start analysing the single time equation for the Lagrange multiplier $\mu(t)$, then we proceed with the two-times function concentrating first on the time-translational invariant part and then on the aging part.

B.2.3 Langrange multiplier in the large time limit.

Starting from Eq. (59), we introduce the symbol $\clubsuit_p$ to isolate the two contribution of matrix and tensor to the integral. As the time tends to infinity m and μ tend to their asymptotic value, respectively m_∞ and μ_∞, Eq. (59) tends to

$$\mu_\infty = T + Q'(m_\infty)m_\infty + p\clubsuit_p + 2\clubsuit_{p=2}\,.$$

We can now use the idea of the separation in two-time regimes. Call $Q_p(x) = x^p/(p\Delta_p)$ the part related to p in $Q(x)$,

$$\begin{aligned}
\clubsuit_p &= \int_0^t Q_p'(C(t,t''))R(t,t'')dt'' = \int_{\text{FDT}} + \int_{\text{aging}} = \\
&= -\int_t^0 Q_p'(C(t,t-\tilde{t}))R(t,t-\tilde{t})d\tilde{t} + \int_0^1 \mathcal{R}(\lambda)Q_p'(q\mathcal{C}(\lambda))d\lambda = \\
&= +\int_0^\infty Q_p'(C_{\text{TTI}}(\tilde{t}))R_{\text{TTI}}(\tilde{t})d\tilde{t} + \int_0^1 \mathcal{R}(\lambda)Q_p'(q\mathcal{C}(\lambda))d\lambda = \\
&= -\int_0^\infty \frac{1}{T}\frac{d}{d\tilde{t}}Q_p(C_{\text{TTI}}(\tilde{t}))d\tilde{t} + \int_0^1 \mathcal{R}(\lambda)Q_p'(q\mathcal{C}(\lambda))d\lambda = \\
&= \frac{1-q^p}{T\Delta_p} + \int_0^1 \mathcal{R}(\lambda)Q_p'(q\mathcal{C}(\lambda))d\lambda
\end{aligned}$$

the resulting equation is

$$\begin{aligned}
\mu_\infty = T + Q'(m_\infty)m_\infty + \frac{1}{T}\left[Q'(1) - q\,Q'(q)\right] + \\
+ p\int_0^1 \mathcal{R}(\lambda)Q_p'(q\mathcal{C}(\lambda))d\lambda + 2\int_0^1 \mathcal{R}(\lambda)Q_2'(q\mathcal{C}(\lambda))d\lambda\,.
\end{aligned} \tag{60}$$

B.2.4 Regime 1: FDT.

We apply the same scheme of separating the times scale and applying FDT to the correlation function. All over the analysis we isolate terms in the equations using the symbols ♣ and ♠. Eq. (56) in the large time is

$$(\partial_\tau + \mu_\infty)C_{\text{TTI}}(\tau) = Q'(m_\infty)m_\infty + \spadesuit + \clubsuit\,, \tag{61}$$

with:

$$\begin{aligned}
\clubsuit &= \int_0^{t'} Q'(C(t,t''))R(t',t'')dt'' = \int_{\text{FDT}} + \int_{\text{aging}} = \\
&= -\int_{t'}^0 Q'(C(t,t'-\tilde{t}))R(t',t'-\tilde{t})d\tilde{t} + \int_0^1 Q'(q\mathcal{C}(\lambda))\mathcal{R}(\lambda)d\lambda = \\
&= -\frac{1}{T}\int_0^\infty Q'(C_{\text{TTI}}(\tau+\tilde{t}))\frac{d}{d\tilde{t}}C_{\text{TTI}}(\tilde{t})d\tilde{t} + \int_0^1 Q'(q\mathcal{C}(\lambda))\mathcal{R}(\lambda)d\lambda
\end{aligned}$$

and

$$\begin{aligned}
\spadesuit &= \int_0^t Q''(C(t,t''))R(t,t'')C(t',t'')dt'' = \int_{t'}^t + \int_0^{t'} = \int_{t'}^t + \int_{\text{FDT}} + \int_{\text{aging}} = \\
&= \frac{1}{T}\left[Q'(1)C_{\text{TTI}}(\tau) - Q'(q)q\right] - \frac{1}{T}\int_0^\tau Q'(C_{\text{TTI}}(\tau-\tilde{t}))\frac{d}{d\tilde{t}}C_{\text{TTI}}(\tilde{t})d\tilde{t} \\
&+ \frac{1}{T}\int_0^\infty Q'(C_{\text{TTI}}(\tau+\tilde{t}))\frac{d}{d\tilde{t}}C_{\text{TTI}}(\tilde{t})d\tilde{t} + \int_0^1 \mathcal{R}(\lambda)Q''(q\mathcal{C}(\lambda))\,q\mathcal{C}(\lambda)d\lambda\,.
\end{aligned}$$

Substituting these expressions and using Eq. (60) in Eq. (61) we have the first equation, which characterises the first regime

$$\partial_\tau C_{\text{TTI}}(\tau) + \left(\frac{1}{T}Q'(1) - \mu_\infty\right)\left[1 - C_{\text{TTI}}(\tau)\right] + T = -\frac{1}{T}\int_0^\tau Q'(C_{\text{TTI}}(\tau-\tau''))\frac{d}{d\tau''}C_{\text{TTI}}(\tau'')d\tau''\,. \tag{62}$$

An important limit that is used later on in the computation is when $\tau \to \infty$ and the variations of C_{TTI} becomes irrelevant. This gives:

$$\mu_\infty = \frac{T}{1-q} + \frac{Q'(1) - Q'(q)}{T}\,. \tag{63}$$

B.2.5 Regime 2: aging.

Starting from the evolution of the response function (57), in this regime the time derivative is negligible.

$$0 = -\mu_\infty \frac{\mathcal{R}(\lambda)}{t} + \clubsuit$$

with $\clubsuit$ that can be separated into three terms $\clubsuit^{(1)}$, $\clubsuit^{(2)}$ and $\clubsuit^{(3)}$

$$\clubsuit = \int_{t'}^{t} R(t,t'')Q''(C(t,t''))R(t'',t')dt'' = \int_{t''\lesssim t} + \int_{t''\gtrsim t'} + \int_{\text{aging}} = \clubsuit^{(1)} + \clubsuit^{(2)} + \clubsuit^{(3)}\,.$$

In the first two integrals we can apply FDT respectively for t'' close to t and for t'' close to t':

$$\begin{aligned}
\clubsuit^{(1)} &= \int_0^\infty R(t,t-\tilde{t})Q''(C(t,t-\tilde{t}))\frac{\mathcal{R}(\lambda)}{t}d\tilde{t} = -\frac{\mathcal{R}(\lambda)}{t}\frac{1}{T}\int_0^\infty \frac{d}{d\tilde{t}}Q'(C_{\text{TTI}}(\tilde{t}))d\tilde{t} = \\
&= \frac{1}{T}\frac{\mathcal{R}(\lambda)}{t}\left(Q'(1) - Q'(q)\right)\,, \\
\clubsuit^{(2)} &= \int_0^\infty \frac{1}{t}\mathcal{R}(\lambda)Q''(q\mathcal{C}(\lambda))R(t'+\tilde{t},t')d\tilde{t} = -\frac{\mathcal{R}(\lambda)}{t}\,Q''(q\mathcal{C}(\lambda))\frac{1}{T}\int_0^\infty \frac{d}{d\tilde{t}}C_{\text{TTI}}(\tilde{t})d\tilde{t} = \\
&= \frac{1}{T}\frac{\mathcal{R}(\lambda)}{t}\,Q''(q\mathcal{C}(\lambda))(1-q)\,.
\end{aligned}$$

The last terms displays aging:

$$\clubsuit^{(3)} = \int_{t'}^{t} \frac{\mathcal{R}(\frac{t''}{t})}{t}\frac{\mathcal{R}(\frac{t'}{t''})}{t''}Q''\left(q\mathcal{C}\left(\frac{t''}{t}\right)\right)dt'' = \frac{1}{t}\int_\lambda^1 \frac{\mathcal{R}(\lambda'')}{\lambda''}\mathcal{R}\left(\frac{\lambda}{\lambda''}\right)Q''\left(q\mathcal{C}(\lambda'')\right)d\lambda''\,.$$

Combining these pieces together and using (63) we obtain an expression for the aging function of the response:

$$0 = \left[-\frac{T}{1-q} + \frac{Q''(q\mathcal{C}(\lambda))(1-q)}{T}\right]\mathcal{R}(\lambda) + \int_\lambda^1 \frac{\mathcal{R}(\lambda'')}{\lambda''}\mathcal{R}\left(\frac{\lambda}{\lambda''}\right)Q''\left(q\mathcal{C}(\lambda'')\right)d\lambda''\,. \tag{64}$$

Following the same steps in Eq. (56) and using again (63) we obtain the expression for the correlation:

$$\begin{aligned}
0 = &-\left[\frac{T}{1-q} + \frac{Q'(q\mathcal{C}(\lambda))(1-q)}{q\,\mathcal{C}(\lambda)\,T}\right]q\,\mathcal{C}(\lambda) + Q'(m_\infty)m_\infty + \\
&+ \int_0^\lambda Q'(q\mathcal{C}(\lambda''))\mathcal{R}\left(\frac{\lambda''}{\lambda}\right)\frac{d\lambda''}{\lambda} + q\int_0^1 \mathcal{R}(\lambda'')Q''(q\mathcal{C}(\lambda''))\mathcal{C}\left[\left(\frac{\lambda''}{\lambda}\right)^{\text{sign}(\lambda-\lambda'')}\right]d\lambda''
\end{aligned}\,. \tag{65}$$

Remark 3 (generalized-FDT) *In the derivation we never used generalized-FDT ansatz, $\mathcal{R}(\lambda) = \frac{1}{T}xq\frac{d\mathcal{C}(\lambda)}{d\lambda}$. A posteriori we can observe its validity as Eq. (64) and Eq. (65) collapse to a single equation as Eq. (65) is derived by λ and generalized-FDT is used.*

B.2.6 Characterisation of the unknown parameters.

In order to fully characterized the FDT Eq. (62) and the aging Eqs. (64-65), we need to determine the parameters m_∞, μ_∞, q, the FDT violation index x and $q_0 = q\mathcal{C}(0)$. We do not determine all of them, we consider only the few that are used in the analysis, but for sake of completeness we say how the five equations can be determined: Eq. (58) taking $t \to \infty$, Eq. (60) plugging the generalized FDT ansatz, Eq. (62) in the large τ limit, Eq. (64) in the limit $\lambda \to 1$, Eq. (65) in the limit $\lambda \to 0$.

In particular the $\lim \lambda \to 1$ of Eq. (64) gives

$$\frac{T^2}{(1-q)^2} = Q''(q)\,. \tag{66}$$

From Eq. (65), in the limit $\lambda \to 1$ and $\lambda \to 0$ we obtain

$$0 = \left[\frac{Tq}{1-q} - \frac{Q'(q)(1-q)}{T}\right] + Q'(m_\infty)m_\infty + \frac{x}{T}[q\,Q'(q) - q_0\,Q'(q_0)]\,, \tag{67}$$

$$0 = \left[\frac{Tq_0}{1-q} - \frac{Q'(q_0)(1-q)}{T}\right] + Q'(m_\infty)m_\infty + q_0\frac{x}{T}[Q'(q) - Q'(q_0)]\,. \tag{68}$$

In the regime where the system does not find a good overlap with the signal thus $m_\infty = 0$, the second equation gives the solution $q_0 = 0$. As T tends to 0 (and q tends to 1)

$$\frac{x}{T} = \frac{1}{q\,Q'(q)}\left[\frac{T}{1-q} - \frac{Q'(1)(1-q)}{T}\right] = \frac{1}{q\,Q'(q)}\left[\sqrt{Q''(q)} - \frac{Q'(1)}{\sqrt{Q''(q)}}\right]. \tag{69}$$

Remark 4 ($\overline{R}$) *In the large time limit, and using FDT, we have*

$$\overline{R} = \int_0^\infty R_{TTI}(\tau'')d\tau'' = -\frac{1}{T}\int_0^\infty C'_{TTI}(\tau'')d\tau'' = \frac{1-q}{T}\,, \tag{70}$$

using Eq. (66) as $T \to 0$ we the result reported in the main text

$$\overline{R} = \frac{1}{\sqrt{Q''(1)}}\,. \tag{71}$$

Remark 5 (marginal states) *Combining Eq. (66) with Eq. (63), we obtain:*

$$\mu_\infty = \sqrt{Q''(q)} + \frac{1}{T}\left[Q'(1) - Q'(q)\right] \tag{72}$$

expanding $q \lessapprox 1$, and using again Eq. (66),

$$\mu_\infty = 2\sqrt{Q''(1)}\,. \tag{73}$$

As we explained in the main text, the distribution of the Hessian is associated to a semicircle of radius $2\sqrt{Q''(1)}$ and centred in μ. This equation tells that asymptotically, if aging does not stops – as it happens if it jumps to the solution – the systems tends to the marginal states. We have shown a numerical confirmation of this property in Fig. 7.

Remark 6 (threshold loss) *As we show in the main text the Lagrange multiplier μ depends on the two losses as $\mu = -p\epsilon_p - 2\epsilon_2$ (or $\mu = T - p\epsilon_p - 2\epsilon_2$ for arbitrary temperature). Observing that the equation holds for any Δ_p and Δ_2, in particular when they tend to infinity and therefore their contribution to the total loss becomes irrelevant, it follows from Eq. (11) (respectively Eq. (59)),*

$$\epsilon_p(t) = -\frac{1}{p}\left[Q'_p(m(t))m(t) + \int_0^t R(t,t'')\left[Q'_p(C(t,t'')) + Q''_p(C(t,t''))C(t,t'')\right]dt''\right] \tag{74}$$

and analogously $\epsilon_2(t)$. We then write the expression for the total loss

$$\begin{aligned}\epsilon(t) = &-\frac{1}{p}Q'_p(m(t))m(t) - \frac{1}{2}\left[Q'_2(m(t))m(t)\right] + \\ &+ \int_0^t R(t,t'')\left[Q'(C(t,t'')) + Q''(C(t,t''))C(t,t'')\right]dt''\,.\end{aligned} \tag{75}$$

From the equation we established above and using the aging ansatz, one can obtain the asymptotic value of the loss for low signal-to-noise ratio, i.e. the loss of the minima trapping the dynamics.

Remark 7 (threshold energy) *The large time limit of Eq. (75) gives two threshold states. Applying the same scheme used in Eq. (60), i.e. integrating Eq. (74) for $t, t' \gg 1$ considering the two time-regimes gives*

$$\epsilon_{p,\text{th}}^{\text{dyn}} = -\frac{1}{p}\left[Q_p'(m_\infty)m_\infty + \frac{1}{T}\left[Q_p'(1) - q\,Q_p'(q)\right] + p\int_0^1 \mathcal{R}(\lambda)Q_p'(q\mathcal{C}(\lambda))d\lambda\right]. \tag{76}$$

Applying the generalized fluctuation dissipation ansatz and Eq. (69) in the integral, and finally taking $T \to 0$ ($q \to 1$)

$$\epsilon_{p,\text{th}}^{\text{dyn}} = -\frac{1}{p}\frac{Q_p'(1) + Q_p''(1)}{\sqrt{Q''(1)}} - \frac{x}{T}Q_p(1)\,. \tag{77}$$

The threshold energy will be given by the some of two contributions, giving

$$\epsilon_{\text{th}}^{\text{dyn}} = -\frac{Q'(1)}{\sqrt{Q''(1)}} - \frac{Q(1)\,(Q''(1) - Q'(1))}{\sqrt{Q''(1)}Q'(1)}\,. \tag{78}$$

Remark 8 (threshold energies equivalence) *After some manipulation we notice that the threshold energy evaluated using Kac-Rice formula, Eq. (44), and the threshold energy evaluated using the dynamical equations, Eq. (77), are equivalent and therefor the two methods are coherent.*

C Numerical Simulations of Gradient-Flow

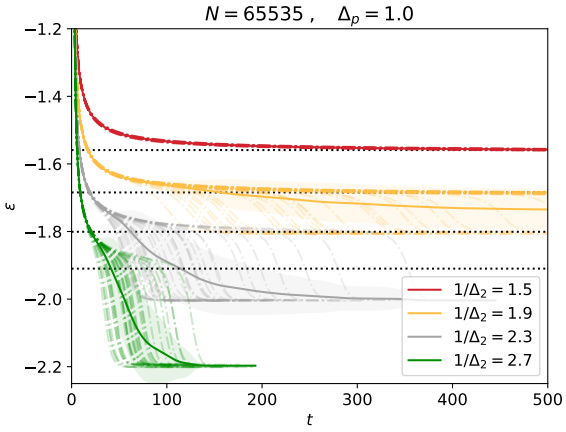


Figure 10: Evolution of the loss in time from numerical simulations realised over 100 instances of disorder and noise, for the spiked matrix-tensor model with $p = 3$. The simulations has been done with systems of size $N = 2^{16} - 1 = 65535$ with parameters $1/\Delta_2 \in \{1.5, 1.9, 2.3, 2.7\}$ and $\Delta_p = 1.0$.

In order to evaluate Gradient flow dynamics we discretized time and evaluated Eq. (4) numerically using effectively gradient descent

$$\sigma_i^{t+1} = -\mu^t \sigma_i^t - \frac{\partial \ell(\boldsymbol{\sigma}^t | \boldsymbol{T}, \boldsymbol{Y})}{\partial \sigma_i^t}\,. \tag{79}$$

In our experiments we run the dynamics on numerous realisations of the problem for different values of the parameters at $p = 3$. Given a signal $\boldsymbol{\sigma}^* \in \mathbb{S}^{N-1}$, the number of computations per interaction scales as N^3, which makes the system hard to simulate for large N. In order to increase the size of the system, we considered a diluted system, as proposed in [40], instead of the real system, such that the first and the second moment of the loss, in the leading order on N. In the original system the (hyper)-graph of interaction is fully connected and counts $N^3/3!$ (hyper)-edges for the tensor and $N^2/2$ edges for the matrix. In the diluted systems we replace the (hyper)-graphs by graphs less connected in particular we take N^2 (hyper)-edges for the tensor and $N\sqrt{N}$

edges for the matrix. In systems with spherical variables there is a known problem [41] associated with reducing too much the number of interaction. In general given a tensor of order p if the number of interaction becomes less then N^{p-1} the system tends to favour a finite number of (hyper)-edges and aligns completely with them. The dynamics then converges to a final configuration where $O(p)$ spins have value of order $O(1)$ and the rest is of order $O(1/\sqrt{N})$. In order to have the same averages for the observables — such as overlap with the signal and loss — called $\#(\cdot)$ that counts the number of interactions, we multiplied the variances of the noise by $N^{3/2}/(3!\#(\boldsymbol{T}))$ and $N/(2\#(\boldsymbol{Y}))$ respectively the tensor noise and the matrix noise.

Using this observation in the code we obtain a simple algorithm that given a dt approximate gradient flow by computing a gradient descent dynamics, with $dt = 1.0$ in the simulations. This value was chosen observing that in the runs the algorithm always descends in terms of loss and not appreciable difference appeared reducing it further. The code is made available and attached to this paper. Using this code we were able to simulate systems of the size $N = 2^{16} - 1 = 65535$ and reduce finite size effects. Fig. 10 shows the average over different initialisation and realisation of the noise for the parameters presented in the paper $\Delta_p = 1.0$ and $1/\Delta_2 \in \{1.5, , 1.9, , 2.3, , 2.7, \}$. In the figure we use a continuous line surrounded by a shadow to represent mean and standard deviation under a Gaussian hypothesis, individual simulations are represented using dashed-dotted lines. For $p = 3$ and $\Delta_p = 1.0$ the critical threshold for gradient flow occurs at $1/\Delta_2^{\mathrm{GF}} = 2.0$ and in fact we observe that the green line ($1/\Delta_2 = 1.9$) shows finite size effects and some simulations find good overlap with the ground truth. To conclude the figure shows a very good agreement with the averaged value evaluated using CHSCK equations, see Fig. 4-b. In particular is evident how all the dynamics tends to the threshold states, whose corresponding losses are drown with horizontal dotted lines, before eventually find the good direction and then the signal.