

High-Dimensional Non-Convex Landscapes and Gradient Descent Dynamics

Tony Bonnaire

Laboratoire de Physique de l'École normale supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université Paris Cité, F-75005 Paris, France.

E-mail: tony.bonnaire@ens.fr

Davide Ghoio

Ecole Polytechnique Fédérale de Lausanne (EPFL). IdePHICS Laboratory

E-mail: davide.ghio@epfl.ch

Kamesh Krishnamurthy

Joseph Henry Laboratories of Physics & PNI, Princeton University

E-mail: kameshk@princeton.edu

Francesca Mignacco

Joseph Henry Laboratories of Physics, Princeton University & Initiative for Theoretical Sciences, Graduate Center, CUNY.

E-mail: fmignacco@princeton.edu

Atsushi Yamamura

Department of Applied Physics, Stanford University

E-mail: atsushi3@stanford.edu

Giulio Biroli

Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université Paris-Diderot, Sorbonne Paris Cité, Paris, France

E-mail: giulio.biroli@ens.fr

August 2017

Abstract

In these lecture notes we present different methods and concepts developed in statistical physics to analyze gradient descent dynamics in high-dimensional non-convex landscapes. Our aim is to show how approaches developed in physics, mainly

statistical physics of disordered systems, can be used to tackle open questions on high-dimensional dynamics in Machine Learning.

1. Introduction

Analyzing high-dimensional dynamics is a central problem in modern science. It appears in very disparate fields: physics, biology, social science and machine learning. What is challenging is that many tools developed for studying the dynamics of few degrees of freedom do not transfer to the high-dimensional case. Several phenomena taking place in high-dimensional dynamics defy low-dimensional intuition. Developing a theory thus requires new paradigms, new methods and new concepts.

Modern machine learning problems, the main context we focus on here, are at the center of this challenge. In fact, the number of data, their dimension, and the number of parameters used in machine learning algorithms are huge and are increasing in a steady manner over the years. Questions such as characterizing the loss landscape and the training dynamics are therefore central examples of the high-dimensional dynamics problem we have described above.

In this set of notes we present methods and concepts developed in statistical physics, mainly statistical physics of disordered systems, which have proven to be useful in tackling high-dimensional dynamics. Needless to say, there is still much to do, to discover and to understand, which makes the research on these topics exciting and open to many new potential contributions.

2. A crash course in random matrix theory

Large random matrices are ubiquitous in many domains ranging from physics to finance and biology [1, 2, 3]. They also play an important role in many modern problems of machine learning, mainly because of the large number of data and their high dimensionality [4]. In this section, we shall recall some key properties of random matrices. We will then present a method – the Dyson Brownian Motion – which triggered substantial progress in recent years.

2.1. The Gaussian orthogonal ensemble

One of the most emblematic sets of random matrices is the Gaussian orthogonal ensemble (GOE). Matrices from this ensemble are symmetric and defined as

$$\mathbf{M} = \frac{1}{\sqrt{2N}}(\mathbf{J} + \mathbf{J}^T), \quad (1)$$

with $j_{ij} \sim \mathcal{N}(0, \sigma^2)$. In other terms, the off-diagonal elements of \mathbf{M} , m_{ij} with $i > j$ are independent Gaussian random variables with zero mean and variance σ^2/N . Diagonal

elements m_{ii} , on the other hand, have twice this variance. The probability distribution associated with the GOE can hence be written

$$p(\mathbf{M}) = \frac{1}{Z} \prod_{i < j} \exp \left\{ -\frac{N}{\sigma^2} \frac{m_{ij}^2}{2} \right\} \prod_i \exp \left\{ -\frac{N}{\sigma^2} \frac{m_{ii}^2}{4} \right\}, \quad (2)$$

$$= \frac{1}{Z} \exp \left\{ \sum_{i,j} -\frac{N}{\sigma^2} \frac{m_{ij}^2}{4} \right\}, \quad (3)$$

$$= \frac{1}{Z} \exp \left\{ -\frac{N}{4\sigma^2} \text{Tr } \mathbf{M}^2 \right\}, \quad (4)$$

where Z is the normalization constant of the probability measure. The measure defined by Eq. (4) is invariant under orthogonal transformations of the matrix $\mathbf{M} \rightarrow \mathbf{O}^\top \mathbf{M} \mathbf{O}$ with $\mathbf{O} \in \mathbb{R}^{N \times N}$ an orthogonal matrix. This invariance explains the name *orthogonal* given to the ensemble. In the following, we consider the case $\sigma = 1$.

2.2. Eigenvector and eigenvalue distributions of the GOE

In many problems, one is actually interested in the spectrum of the matrix instead of its raw elements. Our first aim is hence to study the eigenvalues and eigenvectors of a matrix \mathbf{M} drawn from the GOE. Both these quantities are random variables, and we denote \mathbf{v}^α and λ_α with $\alpha \in \{1, \dots, N\}$ the eigenvectors and eigenvalues of \mathbf{M} respectively. We also order the sets such that λ_1 is the largest eigenvalue and λ_N is the smallest.

Let us discuss the probability distribution of eigenvectors first. The symmetry under orthogonal transformation established in Eq. (4) implies rotational invariance. In consequence, the probability of any given eigenvector \mathbf{v}^α is uniform on the sphere of radius \sqrt{N} and the elements of \mathbf{v}^α have variance $1/N$. By concentration of the Gaussian measure in high dimensions, the probability distribution on the sphere can be approximated by a Gaussian for each component $p(v_i^\alpha) \propto \exp \left\{ -N (v_i^\alpha)^2 / 2 \right\}$. This result can also be seen as a consequence of the the Gaussian annulus theorem‡ or, physically, as the equivalence between micro-canonical and canonical measures. The bottom line is that eigenvectors have a quite simple statistics, and do not present much interest for this ensemble of random matrices.

We now focus on the eigenvalue distribution of \mathbf{M} . To do so, we first build a matrix stochastic process defined by

$$\mathbf{M}(t + dt) = \frac{\mathbf{M}(t) + \mathbf{G}(t)}{\sqrt{1 + dt}}, \quad (5)$$

where $\mathbf{G}(t)$ is a matrix from the GOE with $g_{ij} \sim \mathcal{N}(0, dt/N)$. From the independence of $\mathbf{M}(t)$ and $\mathbf{G}(t)$, the resulting matrix $\mathbf{M}(t + dt)$ is also belonging to the GOE and has, at any time t , the same variance as $\mathbf{M}(t)$ thanks to the normalization term. Such a matrix stochastic process is called a *Dyson-Brownian motion* (DBM) and its invariant measure

‡ Stating that $p(\|\mathbf{v}^\alpha\|_2 - \sqrt{N} \leq t) \geq 2 \exp\{-ct^2\}$.

gives the GOE by construction. Note also the different orders with respect to dt in the terms of Eq. (5); elements of the matrix $\mathbf{M}(t)$ are of order one while those of $\mathbf{G}(t)$ are of order \sqrt{dt} by definition. Assuming dt small, one can write

$$\mathbf{M}(t + dt) \approx \frac{\mathbf{M}(t) + \mathbf{G}(t)}{1 + \frac{dt}{2}}, \quad (6)$$

$$\approx \mathbf{M}(t) - \frac{1}{2} \mathbf{M}(t) dt + \mathbf{G}(t) + \mathcal{O}((dt)^{3/2}), \quad (7)$$

$$\approx \mathbf{M}(t) + \delta\mathbf{M}, \quad (8)$$

where $\delta\mathbf{M} = -\frac{1}{2}\mathbf{M}(t)dt + \mathbf{G}(t)$ is a small perturbation of order $dt + (dt)^{1/2}$ of the matrix $\mathbf{M}(t)$. This can be equivalently written in terms of the eigenvalues of $\mathbf{M}(t)$, for all $\alpha \in \{1, \dots, N\}$,

$$\lambda_\alpha(t + dt) = \lambda_\alpha(t) + \delta\lambda_\alpha, \quad (9)$$

where $\delta\lambda_\alpha$ denotes the small perturbation associated to the eigenvalue λ_α of $\mathbf{M}(t)$. Using time-independent perturbation theory for matrices (or operators) and the notations from quantum mechanics, we have[§]

$$\lambda_\alpha(t + dt) = \lambda_\alpha(t) + \langle \alpha | \delta\mathbf{M} | \alpha \rangle + \sum_{\beta \neq \alpha} \frac{|\langle \alpha | \delta\mathbf{M} | \beta \rangle|^2}{\lambda_\alpha(t) - \lambda_\beta(t)} + \mathcal{O}((dt)^{3/2}), \quad (10)$$

where we used the bra-ket notation for the eigenvectors \mathbf{v}^α of $\mathbf{M}(t)$. For instance, $\langle \alpha | \delta\mathbf{M} | \alpha \rangle$ is the $\alpha\alpha$ element of $\delta\mathbf{M}$ in the basis diagonalizing $\mathbf{M}(t)$. The previous equation can be rewritten as

$$\lambda_\alpha(t + dt) = \lambda_\alpha(t) - \frac{1}{2} dt \lambda_\alpha(t) + g_{\alpha\alpha} + \sum_{\beta \neq \alpha} \frac{g_{\alpha\beta}^2}{\lambda_\alpha(t) - \lambda_\beta(t)} + \mathcal{O}((dt)^{3/2}), \quad (11)$$

where, using the properties of the GOE, $g_{\alpha\alpha}$ and $g_{\alpha\beta}$ are uncorrelated gaussian variables (also in the basis of \mathbf{M} because of rotational invariance). The previous equation is a discretized version of a stochastic equation. In order to consider its continuum limit, it is important to assess the order of magnitude of the different terms. The second term of the right-hand side is deterministic and of order dt , whereas the third term is Gaussian and of order \sqrt{dt} . These are indeed the usual scalings for stochastic equations. The fourth term can be written as

$$\sum_{\beta \neq \alpha} \frac{g_{\alpha\beta}^2}{\lambda_\alpha(t) - \lambda_\beta(t)} = \mathbb{E}_g \left[\sum_{\beta \neq \alpha} \frac{g_{\alpha\beta}^2}{\lambda_\alpha(t) - \lambda_\beta(t)} \right] + \text{fluctuations.}$$

Since the fluctuations are of order dt , they can be neglected in the continuum limit (they give a sub-leading term with respect to $g_{\alpha\alpha}$). Therefore, one can replace the fourth term

[§] This equation can be obtained by computing the eigenvalues of the power series development of Eq. (8) for small perturbations and keeping only the two first terms.

by its average $\frac{dt}{N} \sum_{\beta \neq \alpha} \frac{1}{\lambda_\alpha(t) - \lambda_\beta(t)}$. One therefore obtains the continuum limit stochastic equation on eigenvalues

$$\frac{d\lambda_\alpha}{dt} = -\frac{\lambda_\alpha(t)}{2} + \frac{1}{N} \sum_{\beta \neq \alpha}^N \frac{1}{\lambda_\alpha(t) - \lambda_\beta(t)} + \eta_\alpha(t), \quad (12)$$

where $\eta_\alpha(t)$ is a white noise such that $\mathbb{E}[\eta_\alpha(t)\eta_\beta(t)] = 2\delta_{\alpha,\beta}\delta(t-t')/N$. In a more “mathematically friendly” way, we can write

$$d\lambda_\alpha = \left(-\frac{\lambda_\alpha(t)}{2} + \frac{1}{N} \sum_{\beta \neq \alpha}^N \frac{1}{\lambda_\alpha(t) - \lambda_\beta(t)} \right) dt + dB_\alpha \sqrt{\frac{2}{N}}, \quad (13)$$

where dB_α denotes the Brownian increment. Note this equation is valid for any N , although we will often consider the high-dimensional limit $N \rightarrow \infty$ in what follows. We also remark that Eq. (13) is a Langevin equation for N interacting particles, that can be expressed as

$$\frac{d\lambda_\alpha}{dt} = -\frac{\partial V}{\partial \lambda_\alpha} + \eta_\alpha(t), \quad (14)$$

with η a Gaussian noise term with second moment

$$\mathbb{E}(\eta_\alpha(t)\eta_\alpha(t')) = 2T\delta(t-t')\delta_{\alpha,\beta}, \quad (15)$$

where $T = 1/N$ is the temperature. The potential V is defined by

$$-\frac{\partial V}{\partial \lambda_\alpha} = -\frac{\lambda_\alpha(t)}{2} + \frac{1}{N} \sum_{\beta \neq \alpha}^N \frac{1}{\lambda_\alpha(t) - \lambda_\beta(t)}, \quad (16)$$

leading to

$$V = \sum_\alpha \frac{\lambda_\alpha(t)^2}{4} - \frac{1}{N} \sum_{\alpha < \beta} \ln |\lambda_\alpha(t) - \lambda_\beta(t)|. \quad (17)$$

We recognize two competing terms in V . The first one is a quadratic trapping potential forcing the eigenvalues to be close to zero. The second term is a repulsion term with the form of a two-dimensional Coulomb gas potential forcing the eigenvalues to space apart from each other. It implies in particular that all eigenvalues are correlated, making the system very complex to analyze with standard tools.

The stationary distribution of the process defined by Eq. (14) is known to be given by the Boltzmann distribution

$$p(\lambda_1, \dots, \lambda_N) = \frac{1}{Z} \exp\left\{-\frac{V}{T}\right\}, \quad (18)$$

$$= \frac{1}{Z} \exp\left\{-N \sum_\alpha \frac{(\lambda_\alpha)^2}{4}\right\} \prod_{\alpha < \beta} |\lambda_\alpha - \lambda_\beta|, \quad (19)$$

hence describing the joint probability distribution of the eigenvalues of matrices belonging to the GOE, and where Z provides the normalization constant.

2.3. Density of eigenvalues

The previous subsection tackled the problem of computing $p(\lambda_1, \dots, \lambda_N)$, the joint probability distribution of eigenvalues. We may however wonder what is the typical density distribution of eigenvalues $\rho(\lambda)$ of a matrix taken from the GOE in the large N limit. Injecting the empirical measure

$$\rho_N(\lambda) := \frac{1}{N} \sum_{\alpha} \delta(\lambda - \lambda_{\alpha}) \quad (20)$$

in the joint probability distribution, Eq. (19) reads

$$p(\lambda_1, \dots, \lambda_N) \propto \exp \left[-N^2 \int d\lambda \frac{\lambda^2}{4} \rho_N(\lambda) + \frac{N^2}{2} \int d\lambda d\lambda' \rho_N(\lambda) \rho_N(\lambda') \tilde{\ln} |\lambda - \lambda'| - N \ln \epsilon \right], \quad (21)$$

where $\tilde{\ln}(|x|) = \ln(|x| + \epsilon)$ ^{||}. The probability measure over the function $\rho_N(\lambda)$ can hence be written as a marginal distribution over the joint distribution of eigenvalues as

$$p(\rho_N(\lambda)) = \int \prod_{\alpha=1}^N d\lambda_{\alpha} \delta[\rho_N(\lambda) - \rho(\lambda)] p(\lambda_1, \dots, \lambda_N), \quad (22)$$

$$\propto \exp \{-N^2 F(\rho_N(\lambda)) - N \ln \epsilon\} \int \prod_{\alpha=1}^N d\lambda_{\alpha} \delta[\rho_N(\lambda) - \rho(\lambda)], \quad (23)$$

where $\delta[\cdot]$ denotes the functional delta function, and F reads

$$F(\rho_N(\lambda)) = \int d\lambda \frac{\lambda^2}{4} \rho_N(\lambda) - \frac{1}{2} \int d\lambda d\lambda' \rho_N(\lambda) \rho_N(\lambda') \ln |\lambda - \lambda'|. \quad (24)$$

The first term of Eq. (23) is an energetic contribution in which we can neglect the $N \ln \epsilon$ term in the large N limit, as the leading contribution is of order N^2 [¶]. The second term is an entropic factor, that we will denote $S(\rho_N)$, accounting for the number of ways one can have N particles giving a density $\rho_N(\lambda)$. It can also be seen as the Jacobian of the transformation going from an eigenvalue probability distribution to a functional of the density ρ_N . To compute this entropic factor, let us first replace the delta distribution by its Fourier representation introducing an auxiliary (imaginary) function g leading to

$$S(\rho_N) \propto \int \prod_{\alpha=1}^N d\lambda_{\alpha} \int Dg \exp \left\{ N \int d\lambda' g(\lambda') \rho_N(\lambda') - \sum_{\alpha=1}^N g(\lambda_{\alpha}) \right\}, \quad (25)$$

^{||} We have added this regularizer to take care of the term corresponding to $\lambda_i = \lambda_j$. At the end we will take the $\epsilon \rightarrow 0$ limit. As we shall see, it will be possible to neglect the additional term $-N \ln \epsilon$ since it is subleading with respect to the ones in N^2 .

[¶] This treatment of the ϵ term could look suspicious. The main point is that we are interested in the probability of the "macroscopic" density of eigenvalues, i.e. on scales of order one with respect to N .

allowing to decouple the integral over all λ_α 's. Indeed, one can now rewrite the exponential containing the λ_α 's as a product and then perform independent integrals over the λ_α 's:

$$S(\rho_N) \propto \int Dg \exp \left\{ N \int d\lambda' g(\lambda') \rho_N(\lambda') \right\} \int \prod_{\alpha=1}^N d\lambda_\alpha \exp \left\{ - \sum_{\alpha=1}^N g(\lambda_\alpha) \right\}, \quad (26)$$

$$\propto \int Dg \exp \left\{ N \int d\lambda' g(\lambda') \rho_N(\lambda') \right\} \left(\int d\lambda' \exp \{-g(\lambda')\} \right)^N, \quad (27)$$

finally leading to

$$S(\rho_N) \propto \int Dg \exp \left\{ N \left[\int d\lambda' g(\lambda') \rho_N(\lambda') + \log \int d\lambda' \exp \{-g(\lambda')\} \right] \right\}. \quad (28)$$

Performing a saddle-point on $S(\rho_N)$ requires to compute the functional derivative of the exponent in the exponential, which reads

$$\int d\lambda' \rho_N(\lambda') \delta(\lambda - \lambda') - \frac{\int d\lambda' \delta(\lambda - \lambda') \exp \{-g(\lambda')\}}{\int d\lambda' \exp \{-g(\lambda')\}} = 0, \quad (29)$$

hence leading to

$$g(\lambda) = -\log \rho_N(\lambda) - \log Z, \quad (30)$$

with $Z = \int d\lambda' \exp \{-g(\lambda')\}$. Substituting it back into Eq. (28) finally gives

$$S(\rho_N) \propto \exp \left\{ -N \int d\lambda \rho_N(\lambda) \log \rho_N(\lambda) \right\}. \quad (31)$$

This entropic factor is consequently of order N in the exponential meaning that it can be dropped out compared to the energetic contribution in N^2 when $N \rightarrow +\infty$. All in all, we have found that the probability density of ρ_N is

$$p(\rho_N(\lambda)) \propto \exp \{-N^2 F(\rho_N(\lambda))\}. \quad (32)$$

This has the form of a large-deviation principle – a natural thermodynamic result as ρ_N is a macroscopic observable. It is interesting to remark that, contrary to the usual thermodynamic expression in which there is a N in front of the intensive free-energy, we have got an N^2 . The reason is that the temperature of the associated physical system is very small $T = 1/N$.

By concentration of measure arguments when $N \rightarrow \infty$, and a Laplace method on the set of plausible functions, we obtain that the density is a non-fluctuating quantity given by

$$\rho_N(\lambda) = \rho^*(\lambda) + \mathcal{O}\left(\frac{1}{N}\right), \quad (33)$$

where the average $\rho^*(\lambda) = \min_{\rho(\lambda)} F(\rho(\lambda))$, under the constraint that $\rho^*(\lambda)$ normalizes to one. To solve this variational problem, one needs to take the functional derivative of Eq. (24) and set it to zero which yields

$$\frac{\lambda^2}{4} - \int d\lambda' \rho^*(\lambda') \ln |\lambda - \lambda'| = 0. \quad (34)$$

From there, one trick consists in taking the derivative of this latter expression with respect to λ (taking care of the singularity in the logarithm term) giving

$$\lim_{\Delta \rightarrow 0} \left[\int_{-\infty}^{\lambda - \Delta} d\lambda' \frac{\rho^*(\lambda')}{|\lambda - \lambda'|} + \int_{\lambda + \Delta}^{+\infty} d\lambda' \frac{\rho^*(\lambda')}{|\lambda - \lambda'|} \right] = \frac{\lambda}{2}. \quad (35)$$

The left-hand side of this equation is called the Cauchy principal value (noted Pr) of the integral $\int d\lambda' \rho^*(\lambda')/|\lambda - \lambda'|$. Therefore, the problem now boils down to finding ρ^* solution to

$$\text{Pr} \left[\int d\lambda' \frac{\rho^*(\lambda')}{|\lambda - \lambda'|} \right] = \frac{\lambda}{2}. \quad (36)$$

Actually, such an equation can be solved by Tricomi's theorem [5], see Ref. [6] for the derivation. The solution is

$$\rho^*(\lambda) = \frac{\sqrt{4 - \lambda^2}}{2\pi}. \quad (37)$$

For completeness, we also reinsert σ , whose dependence can be trivially deduced from the $\sigma = 1$ result, to obtain the famous Wigner semi-circle law

$$\rho^*(\lambda) = \frac{\sqrt{4\sigma^2 - \lambda^2}}{2\pi\sigma^2}, \quad (38)$$

illustrated on Fig. 1 as the solid red line together with an empirical distribution of eigenvalues obtained from an $N = 2000$ and $\sigma^2 = 1$ simulation. Let us now comment on the finite but large N case. When N is finite, the typical spacing of the eigenvalues in the bulk are of the order $1/N$, while they are of order $(1/N)^{2/3}$ at the edges of the spectrum located at $\pm 2\sigma$ [2]. We also see from Eq. (32) that the fluctuations of the full distribution $\rho(\lambda)$ are scaling with $\exp\{-N^2\}$, meaning the convergence toward the global shape of the semi-circle law is very fast. However, for one eigenvalue to move away from the expected $\pm 2\sigma$ edge, Eq. (19) teaches us that the cost is only exponential in N , as each eigenvalue contributes with a term of order N to produce the overall N^2 contribution.

So far, we have seen how to study random matrices belonging to the Gaussian orthogonal ensemble. Another interesting case, in particular for data sciences, is the Wishart ensemble with matrices of the form

$$W = \frac{1}{T} \sum_{\mu=1}^T \xi_\mu \xi_\mu^\top, \quad (39)$$

where $\xi_{\mu,i} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Such matrices are naturally arising as covariances of data sets but also as Hessian matrices of some single-layered neural networks. In this case, the

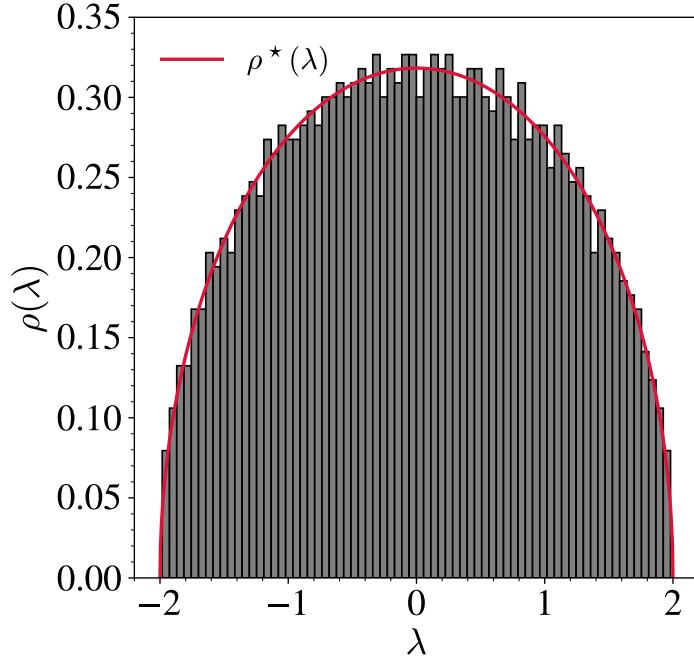


Figure 1. Density of eigenvalues for a matrix of the GOE. The solid red line displays the Wigner semi-circle law from Eq. (37) while the histogram shows the empirical distribution of eigenvalues obtained for a matrix with $N = 2000$ and $\sigma^2 = 1$.

eigenvalue density is known to follow the Marcenko-Pastur distribution [7] when $N \rightarrow \infty$ and $T \rightarrow \infty$ simultaneously such that $q = N/T$ is finite, and

$$\rho^*(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi q \lambda}, \quad (40)$$

where $\lambda_{\pm} = (1 \pm \sqrt{q})^2$. The result above holds for $q < 1$. For $q > 1$ there is also a delta contribution located in the origin and with weight $1 - 1/q$ (the matrix W has rank T , and hence $N - T$ zero eigenvalues for $N > T$). To obtain this result, and others, one can apply the previous procedure based on DBM and similarly build a stochastic process for which the joint probability of eigenvalues is a stationary state

$$\xi_{\mu,i}(t + dt) = \frac{\xi_{\mu,i} + g_{\mu,i}}{\sqrt{1 + dt}}, \quad (41)$$

with $g_{\mu,i} \sim \mathcal{N}(0, dt)$. There obviously exists other methods allowing the derivation of the quantities of interest from this section, for instance relying on the moment method, as done in the seminal paper by Wigner [8], free probabilities [9] or on super-symmetry arguments [10]. However, the Dyson-Brownian motion method bridges well with the topic of this lecture on physics-inspired dynamics. It has also been shown particularly useful for the study of eigenvectors and how they evolve in other ensembles of random matrices, which also explains the recent resurgence of this particular approach in the literature [11].

2.4. Signal to noise ratio transition in random matrices

We now apply DBM to derive a well-known signal-to-noise transition involving random matrices called the BBP phase transition [12] (the transition was first studied by the replica method in [13]). We are interested in recovering a rank-one matrix (the signal) planted in a noisy background. The model is described as follows. Imagine that we are given a matrix \tilde{M} :

$$\tilde{M}_{ij} = M_{ij} + \rho v_i v_j \quad , \quad |\mathbf{v}| = 1 , \quad \mathbf{M} \sim \text{GOE} \quad (42)$$

Here, \tilde{M}_{ij} , $1 \leq i, j \leq N$ are our measurements, \mathbf{v} is the signal, M_{ij} corresponds to the background noise, and the parameter ρ controls the signal-to-noise. The matrix \mathbf{M} is from the GOE, and its off-diagonal entries have a variance $1/N$. We are interested in the values of ρ for which we can recover the signal \mathbf{v} , in the large N limit. This is a variant of the Principal Component Analysis problem, also called sometimes matrix PCA.

To study this problem, we construct the DBM as follows: let $\tilde{M}(t=0) = \rho \mathbf{v} \mathbf{v}^T$, and

$$\tilde{M}(t+dt) = \tilde{M}(t) + \mathbf{g}(t) \quad (43)$$

where $\mathbf{g}(t)$ is a matrix from the GOE with a variance dt/N . By construction, $\tilde{M}(t=1) = \tilde{M} = \mathbf{M} + \rho \mathbf{v} \mathbf{v}^T$, we have therefore to study this stochastic matrix process from $t = 0$ to $t = 1$. At $t = 0$, there is only one non-zero eigenvalue $\lambda_1 = \rho$; at a small but finite time, $\lambda_1 = \rho$ and the remaining eigenvalues are clustered around zero. As before, we can write a stochastic equation describing the evolution of the eigenvalues

$$\frac{d\lambda_\alpha}{dt} = \frac{1}{N} \sum_{\beta \neq \alpha} \frac{1}{\lambda_\alpha(t) - \lambda_\beta(t)} + \eta_\alpha(t). \quad (44)$$

For $\alpha > 1$, it can be rewritten as

$$\frac{d\lambda_\alpha}{dt} = \frac{1}{N} \sum_{\beta \neq \alpha, 1} \frac{1}{\lambda_\alpha(t) - \lambda_\beta(t)} + \frac{1}{N} \frac{1}{\lambda_\alpha(t) - \lambda_1(t)} + \eta_\alpha(t). \quad (45)$$

The first term is $O(1)$, the second term is $O(1/N)$ and the noise term is $O(1/\sqrt{N})$. In the large N limit, we can thus ignore the second term, and the time evolution of λ_α for $\alpha > 1$ is the same than for GOE: at time t the $\{\lambda_\alpha\}$ s will be distributed according to the semi-circle law with edges at $\pm 2\sqrt{t}$. Note that the reason the semi-circle has edges growing with t is because we did not renormalize the variance after each infinitesimal step like before.

For the top eigenvalue, the equation reads

$$\frac{d\lambda_1}{dt} = \frac{1}{N} \sum_{\beta \neq 1} \frac{1}{\lambda_1(t) - \lambda_\beta(t)} + \eta_1(t) \quad (46)$$

The first term is $O(1)$, and the noise term is $O(1/\sqrt{N})$. In the large N limit we can make two more approximations: (i) neglect the noise term, and (ii) replace the sum in the first term with an integral over the distribution $\rho(\lambda_{\alpha>2})$ – the semi-circle law with edges at

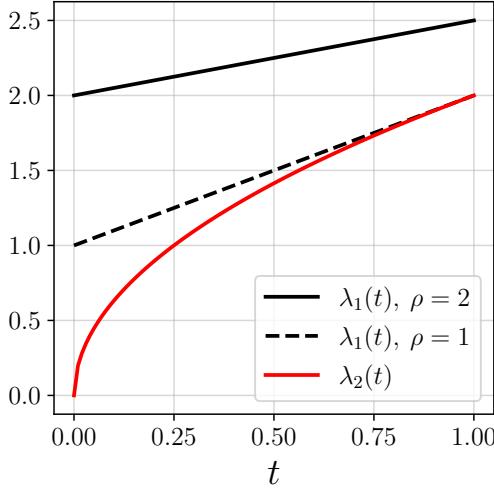


Figure 2. Dyson-Brownian motion for the BBP transition. Evolution of the outlier eigenvalue λ_1 (black lines) for two different values of ρ and the edge of the semi-circle λ_2 (red line). When $\rho > 1$, the outlier is clearly separated from the bulk and the signal can be distinguished from the noise background. When $\rho = 1$, there is a transition, and the outlier and the edge coincide for $t = 1$. For $\rho < 1$, there is no outlier, and the signal is lost in the noise background.

$\pm 2\sqrt{t}$. The assumption (ii) holds as long as λ_1 is at a finite distance from all the other eigenvalues (finite means not vanishing when $N \rightarrow \infty$). With these approximations, we get

$$\frac{d\lambda_1}{dt} = \int d\lambda \rho(\lambda) \frac{1}{\lambda_1 - \lambda} = \frac{\lambda_1(t) - \sqrt{\lambda_1^2(t) - 2t}}{2t}. \quad (47)$$

The solution to this equation turns out to be simply given by

$$\lambda_1(t) = \rho + \frac{t}{\rho}. \quad (48)$$

It is important to ensure that the solution is consistent with the assumptions we made previously. In particular, we relied on a gap existing between λ_1 and the rest of the eigenvalues, i.e. $\lambda_1 - \lambda_2 > \epsilon > 0$, where $\lambda_2(t) = 2\sqrt{t}$ is the right edge of the semi-circle. Fig. 2 shows the behavior of the two eigenvalues for two different values of ρ .

As explained before, we need to focus on $t = 1$ to study the statistics of \tilde{M} . We find that for $\rho > 1$ the signal is strong and the matrix \tilde{M} has one eigenvalue (a spike associated to the original signal) out of the Wigner semi-circle. For $\rho = 1$, the two largest eigenvalues meet at $t = 1$ and we have a transition; for $\rho < 1$ reaches the edge of the semi-circle before $t = 1$. In this case, one can show that for larger times λ_1 remains at the right edge of the semi-circle and the eigenvalue corresponding to the signal is buried within the bulk.

Actually, to recover information on the signal, one has to focus on the eigenvector v_1 corresponding to the largest eigenvalue. When $\rho > 1$, it can be shown that the leading

eigenvector has a finite overlap with the signal direction \mathbf{v} [12]. Specifically,

$$\mathbf{v}_1 = \left(\sqrt{1 - \frac{1}{\rho^2}} \right) \mathbf{v} + \frac{1}{\rho} \mathbf{v}_{\perp} \quad (49)$$

where \mathbf{v}_{\perp} is orthogonal to \mathbf{v} . However, when $\rho < 1$, the overlap of the leading eigenvector with the signal direction is vanishingly small : $\langle \mathbf{v}_1, \mathbf{v} \rangle \sim O(1/\sqrt{N})$. This can be understood studying the DBM for the eigenvectors (due to the small denominators in the perturbation theory the leading eigenvector hybridizes with all the other ones). Hence, $\rho = 1$ corresponds to a transition from a regime in which the eigenvector associated to the largest eigenvalue has a finite component in the direction of the signal to a regime in which it does not.

3. Gradient flow in matrix PCA and spherical spin-glasses

We now use the signal+noise problem exposed in the previous section as a model to study how the dynamics of optimization in high-dimensions can lead to surprising phenomena. To do this, we formulate the problem of recovering the signal matrix in the noise background as an optimization problem and use gradient flow to find the optimum. Note that the purpose of doing this is to gain an insight on the dynamics and it does not imply that this is an efficient method to find the solution. We shall see that even in this simple problem one can gain interesting insights. In particular, we shall show that gradient flow has an algorithmic transition. Close to the transition, the system first converges toward uninformative saddles and then escape via a direction correlated with the signal. This is the simplest example of a more general mechanism that we will discuss later.

Let us change slightly the notation of the signal+noise problem: our measurements M_{ij} are given by

$$M_{ij} = v_i v_j + \frac{N}{\rho} J_{ij}, \quad (50)$$

where \mathbf{v} is a vector such that $\|\mathbf{v}\|^2 = N$, ρ is the signal-to-noise ratio and J , which corresponds to the noise, is from the GOE. We aim at recovering \mathbf{v} using gradient flow on the sphere. To do this we construct an energy function,

$$E(\mathbf{x}) = \frac{\rho}{4N} \sum_{lm} (M_{lm} - x_l x_m)^2 ; \quad \sum_i x_i^2 = N, \quad (51)$$

and perform gradient flow on the sphere with random initial conditions. The x -dependent term of the energy function can be rewritten as

$$-\frac{\rho}{2N} \sum_{lm} M_{lm} x_l x_m, \quad (52)$$

$$= \underbrace{-\frac{1}{2} \sum_{lm} x_l J_{lm} x_m}_{\text{spin-glass term}} - \underbrace{\frac{\rho}{2} N \left(\sum_l \frac{x_l v_l}{N} \right)^2}_{\text{deterministic term}}. \quad (53)$$

The first term has the form of an energy of a spherical spin glass. The presence of the spin-glass term makes the landscape non-convex and leads to non-trivial dynamics (although a simple one due to the spherical constraint).

We begin by analyzing the critical points of this energy function on the sphere. To do so, we form the Lagrangian given by

$$\mathcal{L}(\mathbf{x}, \lambda) = E(\mathbf{x}) + \frac{\lambda}{2} \left(\sum_i x_i^2 - N \right). \quad (54)$$

where λ is a Lagrange multiplier enforcing the spherical constraint. The extrema of the Lagrangian on the sphere satisfy

$$\frac{\partial E}{\partial x_i} + \lambda x_i = 0 \quad ; \quad \sum_i x_i^2 = N. \quad (55)$$

Let us define $\tilde{\mathbf{M}} = \rho \mathbf{M}/N = \mathbf{J} + \rho \mathbf{v}\mathbf{v}^\top$. Then, $E = -1/2 \sum_{l,m} x_l \tilde{M}_{lm} x_m$ and the equation for the extremum reads

$$-\sum_m \tilde{M}_{l,m} x_m + \lambda x_l = 0. \quad (56)$$

This is nothing but an eigenvalue equation with the corresponding eigenvalues $\{\lambda^\alpha\}$ and eigenvectors \mathbf{v}^α as solutions. Therefore, this energy landscape has $2N$ critical points $\mathbf{x}^\alpha = \pm \sqrt{N} \mathbf{v}^\alpha$, each couple associated to an eigenvector and eigenvalue.

The Hessian (evaluated on the sphere) of the critical point with index α is given by

$$H_{lm}^\alpha = -\tilde{M}_{lm} + \lambda^\alpha \delta_{lm}, \quad (57)$$

and the eigenvalues of the Hessian are given by

$$\epsilon^\beta = -\lambda^\beta + \lambda^\alpha; \quad (\beta \neq \alpha). \quad (58)$$

Note there are only $N-1$ eigenvalues since the Hessian is calculated on the sphere. When $\alpha = 1$, all the eigenvalues ϵ^β are positive, and this corresponds to two global minima, whereas when $\alpha = N$, all the eigenvalues are negative, leading to two global maxima. All intermediate values of α have at least one unstable direction making them saddles. In this case, the landscape is hence simply composed of two maxima, two minima and $2N-4$ saddles. We can now use the previous results obtained for the matrix $\tilde{\mathbf{M}}$ by the DBM method. Fig. 3 shows the spectrum of the Hessian for the global minima when ρ is less/greater than 1. When $\rho > 1$ we have a stable global minimum as there is a finite gap between the largest eigenvalue of $\tilde{\mathbf{M}}$ and the second largest (hence leading to $\epsilon^1 > 0$); this is not the case for $\rho < 1$, in which there is no finite gap, $\epsilon^1 \rightarrow 0$ for $N \rightarrow 0$, hence leading to a marginally stable global minimum with many flat directions (or more precisely a Hessian characterized by arbitrary small positive eigenvalues). Moreover, when $\rho > 1$ the leading eigenvector \mathbf{v}^1 , which points in the direction of the global minimum, has a finite overlap with the signal direction \mathbf{v} , whereas when $\rho < 1$, \mathbf{v}^1 is unaligned with \mathbf{v} .

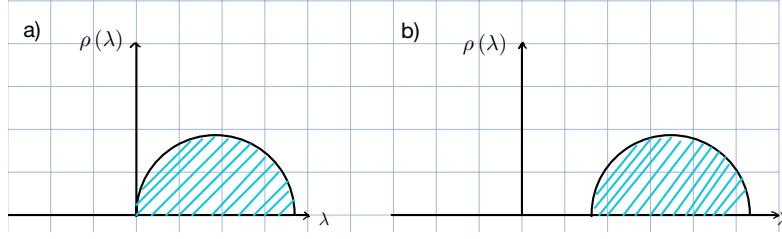


Figure 3. Hessian spectra of the global minimum on the sphere. a) when $\rho < 1$, the global minima are marginally stable. b) when $\rho > 1$, we have a stable global minimum

The dynamics of the gradient flow on the sphere is given by

$$\frac{dx_i}{dt} = \sum_m \tilde{M}_{i,m} x_m(t) - \lambda(t) x_i(t), \quad (59)$$

where the Lagrange multiplier $\lambda(t)$ enforces the spherical constraint on \mathbf{x} . Note that the Lagrange multiplier can be related to the energy as $\lambda(t) = -2E(t)/N$. We consider initial conditions sampled uniformly randomly on the sphere.

Transforming to the eigenbasis of $\tilde{\mathbf{M}}$, we get

$$\frac{dx_\alpha}{dt} = \lambda_\alpha x_\alpha(t) - \lambda(t) x_\alpha(t), \quad (60)$$

which implies

$$x_\alpha(t) = x_\alpha(0) \exp \left\{ \int_0^t dt' (\lambda_\alpha - \lambda(t')) \right\}. \quad (61)$$

Moreover, the normalization constraint implies,

$$N = \sum_\alpha x_\alpha(0)^2 \exp \left\{ 2 \int_0^t dt' (\lambda_\alpha - \lambda(t')) \right\}, \quad (62)$$

which fixes $\lambda(t)$. To proceed, we first take the limit $N \rightarrow \infty$ (before $t \rightarrow \infty$). For large N , we can use the fact that the eigenvectors are uncorrelated with the eigenvalues, to write

$$1 = \sum_\alpha \frac{x_\alpha(0)^2}{N} \exp \left\{ 2 \int_0^t dt' (\lambda_\alpha - \lambda(t')) \right\}, \quad (63)$$

$$= \sum_\alpha \exp \left\{ 2 \int_0^t dt' (\lambda_\alpha - \lambda(t')) \right\}, \quad (64)$$

$$= \int d\eta \rho^*(\eta) \exp \left\{ 2 \int_0^t dt' (\eta - \lambda(t')) \right\}, \quad (65)$$

$$= \int d\eta \frac{\sqrt{2-\eta^2}}{2\pi} \exp \left\{ 2 \int_0^t dt' (\eta - \lambda(t')) \right\}, \quad (66)$$

$$= \left(\int d\eta \frac{\sqrt{2-\eta^2}}{2\pi} \exp \{ 2\eta t \} \right) \exp \left\{ -2 \int_0^t dt' \lambda(t') \right\}. \quad (67)$$

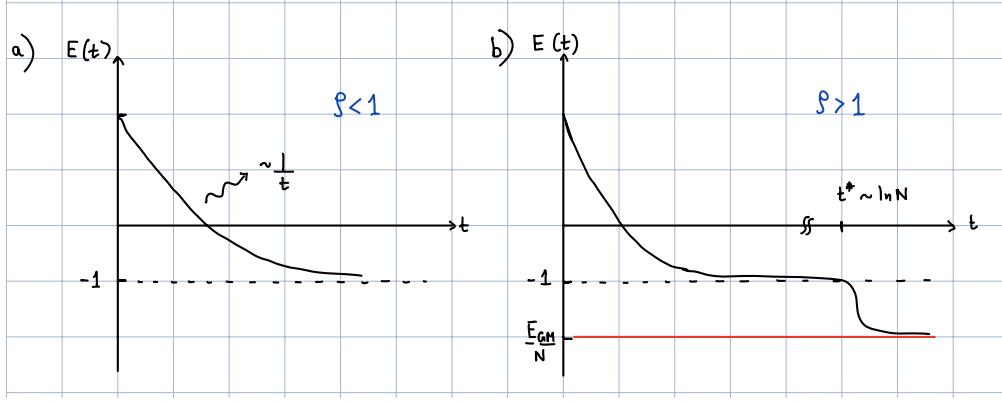


Figure 4. Energy dynamics: (a) When $\rho < 1$ the energy relaxes to -1 as $\sim 1/t$ – i.e. a power law. (b) When $\rho > 1$, there is a power law relaxation to -1 , but after a time that grows as $t^* \sim \ln N$, the system will converge exponentially fast to the global minimum $E_{GM}/N < -1$.

This relation allows us to study the behavior of $\lambda(t)$ – and thus the energy – in different regimes. For instance, for large t , the term in the parenthesis is dominated by the largest exponent and to leading order it scales as $\exp\{4t\}/t^{3/2}$, and this implies that for the RHS to be finite, we should have (for large t)

$$\lambda(t) = 2 - \frac{3}{4t} + O\left(\frac{1}{t^2}\right), \quad (68)$$

and hence

$$E(t) = -1 + \frac{3}{8t} + O\left(\frac{1}{t^2}\right). \quad (69)$$

We see that there is a power-law relaxation in the energy, and it reaches its lowest value of -1 slowly. How does this asymptotic value compare to the energy of the global minima E_{gm} ? Using that E_{gm} is given by $-1/2$ times the largest eigenvalue of \tilde{M} , one finds that when $\rho > 1$, $E_{gm} = -1/2(\rho + 1/\rho) < -1$ and when $\rho < 1$, $E_{gm} = -1$. Thus, we find that for $\rho < 1$ the system is able to reach the energy of the global minimum at large times. However, in this case the signal is not recovered as the global minimum does not point in the direction of the signal. On the contrary, for $\rho > 1$ the global minimum does point in the direction of the signal but the system never reaches it in final time and, instead, remains at higher energies.

In order to solve this puzzle, let us consider the behavior when N is large, but not infinite. When, $\rho > 1$, the condition that enforces the spherical constraint is given by

$$1 = \left(\sum_{\alpha \neq 1} \frac{x_\alpha(0)^2}{N} \exp\{2\lambda_\alpha t\} + \frac{x_1(0)^2}{N} \exp\{2\lambda_{max} t\} \right) \exp\left\{-2 \int_0^t dt' \lambda(t')\right\}. \quad (70)$$

This gives us a timescale t^* below which the first term dominates and, for $t \gg t^*$, the second term will dominate. The timescale is consequently given by

$$t^* \sim \frac{\ln N}{\lambda_{max} - 2} \quad (71)$$

Moreover, for $t \gg t^*$, $E(t) = -\lambda(t)/2 \approx -\lambda_{max}/2$ meaning that the gradient flow eventually finds the global minimum, but only after a “search phase” which takes $\sim t^*$ time. This behaviour is sketched in Fig. 4. Such a behavior of the dynamics separated in two phases: a *search* phase (for $t \ll t^*$) and a *convergence* phase (for $t \gg t^*$) as it is also observed in more complex models and discussed in [14].

In this section, we examined a quite simple model with a landscape that exhibits $2N$ critical points. The analysis of more complicated landscapes with exponential number of minimas is more involved and require more elaborated tools. This will be the topic of the next sections.

4. Critical points of high-dimensional landscapes

In many interesting problems, both in machine learning and in physics of disordered systems, one needs to deal with dynamics in various kinds of high-dimensional rough landscapes. Characterizing the properties of such landscape is generally a challenge. The analysis of mean-field spin-glasses provided a case in which such a challenge was faced for the first time [15]. In recent years, this topic received a lot of interest from the mathematical and physics community. One of the main progress was the development of a method, called Kac-Rice, which allows to study in full details, and to a large extent rigorously, high-dimensional rough landscapes.

This section is divided as follows: in Sect. 4.1 we define the kind of energy landscape we are going to study, motivating why it is important, and then in Sect. 4.2 is introduced the Kac-Rice method, which allows us to compute the complexity of critical points in such a framework. After going through the computation for a simplified case, we show the results for more general models, concluding with some references to applications of the Kac-Rice method beyond physics.

4.1. Random Gaussian energy functions in high-dimensions and generalized spin-glasses

The energy function we are going to consider is of the form

$$E(\mathbf{s}) = - \sum_{p'=1}^{\infty} c_{p'} \sum_{i_1, \dots, i_{p'}} J_{i_1 \dots i_{p'}} s_{i_1} \dots s_{i_{p'}} - rN \sum_{p''=1}^{\infty} \frac{b_{p''}}{p''} \left(\frac{1}{N} \sum_i s_i v_i \right)^{p''}, \quad (72)$$

where \mathbf{s} is a variable on the unitary sphere (i.e. $\sum_i s_i^2 = N$), $J_{i_1 \dots i_p} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{2N^{p-1}})$, r plays the role of the Signal-to-Noise-Ratio and \mathbf{v} is the signal that we want to retrieve. We notice that we have a first term which is random and a second term which instead is deterministic, and we will see how changing r , and thus changing the relative strength of the two, modifies the properties of the landscape. The energy function above can also be seen as the one of a generalized mean-field spin glass in the presence of a (generalized) field \mathbf{v} .

In order to show the generality of this model, let us introduce two examples that can be described with this kind of formalism:

- Tensor PCA [16]: $p' = p'' = p \geq 3$. This is nothing else than the generalization to tensors of what we studied in the previous section, such that for example for $p = 3$ one has $T_{i_1 i_2 i_3} = v_{i_1} v_{i_2} v_{i_3} + \lambda J_{i_1 i_2 i_3}$ and writing an energy function $E = \sum_{i_1, i_2, i_3} (T_{i_1 i_2 i_3} - s_{i_1} s_{i_2} s_{i_3})^2$ it is easy to see that we come back to the form (72) for $E(\mathbf{s})$, in which the parameter λ acts as an inverse SNR (i.e. a Noise-to-Signal-Ratio).
- Random Gaussian Functions on the sphere (in high-dimension): Taking just one element of the first sum in (72), we have $E(\mathbf{s}) = -\sum_{i_1, \dots, i_p} J_{i_1 \dots i_p} s_{i_1} \dots s_{i_p}$, which is a Gaussian random variable on the sphere, since \mathbf{J} is Gaussian and \mathbf{s} is defined on the sphere. Therefore, we can easily compute

$$\langle E(\mathbf{s}) \rangle_J = 0, \quad (73)$$

$$\begin{aligned} \langle E(\mathbf{s}) E(\mathbf{s}') \rangle_J &= \left\langle \sum_{i_1, \dots, i_p} J_{i_1 \dots i_p} s_{i_1} \dots s_{i_p} \sum_{i'_1, \dots, i'_p} J_{i'_1 \dots i'_p} s'_{i'_1} \dots s'_{i'_p} \right\rangle_J = \\ &= \frac{1}{2N^{p-1}} \sum_{i_1} s_{i_1} s'_{i_1} \dots \sum_{i_p} s_{i_p} s'_{i_p} = \frac{N}{2} \left(\frac{\sum_i s_i s'_i}{N} \right)^p = \frac{N}{2} q^p(\mathbf{s}, \mathbf{s}'), \end{aligned} \quad (74)$$

where we first used the independence of the elements of \mathbf{J} , then the value of their variance and finally we defined the *overlap* $q(\mathbf{s}, \mathbf{s}') = \frac{1}{N} \sum_i s_i s'_i$. Thus, we have showed that the first term of (72) is nothing else than the sum of Gaussian random functions of zero mean and covariance related to the overlap between two signals defined on the sphere. Putting back also the second (deterministic) term, which contributes only to the mean value, in general for (72) we have

$$\langle E(\mathbf{s}) \rangle_J = -r f_D(q(\mathbf{s}, \mathbf{v})) N, \quad \text{where } f_D(x) = \sum_{p''=1}^{\infty} \frac{b_{p''}}{p''} x^{p''}, \quad (75)$$

$$\langle E(\mathbf{s}) E(\mathbf{s}') \rangle_J - \langle E(\mathbf{s}) \rangle_J \langle E(\mathbf{s}') \rangle_J = \frac{N}{2} f_R(q(\mathbf{s}, \mathbf{s}')), \quad \text{where } f_R(x) = \sum_{p'=1}^{\infty} c_{p'} x^{p'}. \quad (76)$$

Thus, the average of our Gaussian random function is going to depend only on the overlap between \mathbf{s} and the signal \mathbf{v} through a deterministic function f_D , which is defined in such a way that the overlap needs to be high in order to minimize the energy function. In the meantime, the second equation tells us that there is also a random part (of zero mean) with covariance depending just on the overlap between \mathbf{s} and \mathbf{s}' .

We will now show a technique which allows us to study the structure of the critical points of this energy function and how it changes with r , f_D and f_R .

4.2. The Kac-Rice method

The aim is to compute the number of critical points of $E(\mathbf{s})$, namely the number of vectors on the sphere \mathbf{s}' such that $\nabla_{\perp} E(\mathbf{s}') = 0$, where the symbol \perp indicates that we are taking

the gradient along the sphere.

We know that in one dimension the number of zeros of a certain function $f(x)$ can be written as $\int dx \delta(f(x))|f'(x)|$. The generalization in N dimensions, for the zeros of the gradient, is then

$$\mathcal{N}(E)dE = \int ds \delta(\nabla_{\perp}E) |\det \nabla_{\perp}^2 E(s)| \delta(E(s) - E) dE, \quad (77)$$

where we defined $\mathcal{N}(E)$ as the number of critical points with energy between E and $E+dE$. Since $E(s)$ is a Gaussian random function, also $\mathcal{N}(E)$ is random, and in this section we are going to compute its average, following the pioneering work [17]. This average is called *annealed*, and we will write it as

$$\langle \mathcal{N}(E) \rangle \sim e^{N\Sigma(E/N)}, \quad (78)$$

where Σ is called *complexity* (of critical points). In principle, one would be interested to go beyond just this average and use the fact that

$$\mathcal{N}(E) \sim e^{\langle \ln \mathcal{N}(E) \rangle}, \quad (79)$$

but computing the *quenched average* $\langle \ln \mathcal{N}(E) \rangle$ is in general a much more complicated task and can be done by combining the Kac-Rice and replica methods [18].

4.2.1. Purely random case Let us consider the case $f_D(x) = 0$, in which we can solve the problem in three, rather easy, steps. The computation we are going to show was first presented in [17].

First, we rewrite

$$\langle \mathcal{N}(E) \rangle = \int ds P(s), \quad \text{where } P(s) = \langle \delta(\nabla_{\perp}E) |\det \nabla_{\perp}^2 E(s)| \delta(E(s) - E) \rangle. \quad (80)$$

Then, since we have $\langle E(s) \rangle_J = 0$ and $\langle E(s)E(s') \rangle_J = \frac{N}{2}f_R(q(s,s'))$, the probability in (80) depends just on the overlaps and as a consequence we have rotational invariance of $P(s)$, which thus needs to be uniform and we can write $P(s) = P(\mathbf{1}) \forall s$. This means we can write $\langle \mathcal{N}(E) \rangle = P(\mathbf{1}) \int ds = P(\mathbf{1})S_N(\sqrt{N})$, with $S_N(\sqrt{N})$ the surface of the N -dimensional sphere of radius \sqrt{N} , which can be computed in closed form and for N large can be approximated as $S_N(\sqrt{N}) \approx (2\pi e)^{N/2}$.

Thus, the second step is to compute

$$P(\mathbf{1}) = \langle \delta(\nabla_{\perp}E(\mathbf{1})) |\det \nabla_{\perp}^2 E(\mathbf{1})| \delta(E(\mathbf{1}) - E) \rangle. \quad (81)$$

Considering again just one term of the first sum in (72) by fixing $p' = p$, we see that $E(\mathbf{1}) = -\sum_{i_1, \dots, i_p} J_{i_1 \dots i_p}$ is clearly a Gaussian variable. But then $\nabla_{\perp}E(\mathbf{1})$ will be a Gaussian vector and $\nabla_{\perp}^2 E(\mathbf{1})$ a Gaussian matrix, so that (81) is nothing else than a Gaussian average of Gaussian variables, and thus we just need to compute their individual averages and

covariances to solve (81). First, since we neglected the deterministic part, we know that $\langle E(\mathbf{1}) \rangle = \langle \nabla_{\perp} E(\mathbf{1}) \rangle = \langle \nabla_{\perp}^2 E(\mathbf{1}) \rangle = 0$ and we can also easily see that $\langle E(\mathbf{1})E(\mathbf{1}) \rangle = \frac{N}{2}f_R(1)$. Concerning the other covariances, we first use the fact that

$$\langle (\nabla E(s))E(s') \rangle = \nabla \langle E(s)E(s') \rangle = \frac{N}{2}f'_R(q(s,s'))\frac{s'}{N} \quad (82)$$

to get $\langle (\nabla E(\mathbf{1}))E(\mathbf{1}) \rangle = \frac{f'_R(1)}{2}\mathbf{1}$. Then, in order to get $\langle (\nabla_{\perp} E(\mathbf{1}))E(\mathbf{1}) \rangle$ we project the LHS on the plane orthogonal to $\mathbf{1}$. However, repeating the same for the RHS, which is proportional to $\mathbf{1}$, we get zero hence establishing that $\langle (\nabla_{\perp} E(\mathbf{1}))E(\mathbf{1}) \rangle = 0$. This means that that $\nabla_{\perp} E(\mathbf{1})$ is not correlated with $E(\mathbf{1})$. Similarly, one can obtain

$$\langle \nabla_{\perp}^{\alpha} E(\mathbf{1}) \nabla_{\perp}^{\beta} E(\mathbf{1}) \rangle = \nabla_{\perp}^{\alpha} \nabla_{\perp}^{\beta} \langle E(\mathbf{1})E(\mathbf{1}) \rangle = \delta_{\alpha\beta} \frac{f'_R(1)}{2} + \mathcal{O}(1/N) \quad (83)$$

and rewrite the Hessian as

$$[\nabla_{\perp}^2 E(\mathbf{1})]_{\alpha\beta} = G_{\alpha\beta} - f'_R(1) \frac{E(\mathbf{1})}{N} \delta_{\alpha\beta}, \quad (84)$$

where G is a GOE matrix of covariance $\langle G_{\alpha\beta}^2 \rangle = \frac{f''_R(1)}{2N}$. From (84) we see that the Hessian is correlated with the energy, and in particular one has

$$\langle E(\mathbf{1}) [\nabla_{\perp}^2 E(\mathbf{1})]_{\alpha\beta} \rangle = -f'_R(1) \frac{\langle E(\mathbf{1})E(\mathbf{1}) \rangle}{N} \delta_{\alpha\beta} = -\frac{f_R(1)f'_R(1)}{2} \delta_{\alpha\beta} \quad (85)$$

but not with the gradient, since it is easy to see that $\langle \nabla_{\perp}^2 E(\mathbf{1}) \nabla_{\perp} E(\mathbf{1}) \rangle = 0$.

Now we have all we need to compute $P(\mathbf{1})$. Defining $f_{\alpha} \equiv \nabla_{\perp}^{\alpha} E(\mathbf{1})$ and $\bar{e} \equiv E(\mathbf{1})/N$ we can write

$$\begin{aligned} P(\mathbf{1}) &= \int \prod_{\alpha=1}^{N-1} df_{\alpha} \frac{e^{-f_{\alpha}^2/f'_R(1)}}{\sqrt{\pi f'_R(1)}} d\bar{e} \frac{e^{-N\bar{e}^2/f_R(1)}}{\sqrt{\pi f_R(1)/N}} \prod_{i,j} dG_{ij} \frac{e^{-G_{ij}^2/f''_R(1)}}{\sqrt{\pi f''_R(1)}} \prod_{\alpha} \delta(f_{\alpha}) \cdot \\ &\quad \cdot |\det(G_{\alpha\beta} - f'_R(1)\bar{e}\delta_{\alpha\beta})| \delta(e - \bar{e}) = \\ &= \left(\frac{1}{\sqrt{\pi f'_R(1)}} \right)^{N-1} \frac{e^{-N\bar{e}^2/f_R(1)}}{\sqrt{\pi f_R(1)/N}} \langle |\det(G - f'_R(1)e\mathbb{I})| \rangle_{\text{GOE}}, \end{aligned} \quad (86)$$

where we used the delta distributions and we wrote the integral over the G_{ij} implicitly. The third and final step is to compute explicitly this GOE average, using the techniques we described in Section 2. This could be done in general for every N , but what we are going to do is to take the limit $N \rightarrow \infty$. Then, calling λ_{α} the eigenvalues of G we can rewrite the average as

$$\left\langle \prod_{\alpha} |\lambda_{\alpha} - f'_R(1)e| \right\rangle = \left\langle \exp \sum_{\alpha} \ln |\lambda_{\alpha} - f'_R(1)e| \right\rangle = \left\langle \exp \left(N \int d\lambda \rho(\lambda) \ln |\lambda - f'_R(1)e| \right) \right\rangle \quad (87)$$

where we used the definition of the density of eigenvalues (20). Now in general this average is still complicated to compute, but in the large N limit we can exploit the

properties of GOEs that we reported in Section 2; namely that $\rho(\lambda)$ converges to the semicircle law $\rho^*(\lambda)$ as e^{-N^2} , which allows us to erase the average in (87) just by substituting the semicircle law to the density $\rho(\lambda)$, since all the fluctuations become exponentially small for $N \rightarrow \infty$. Then writing explicitly $\rho^*(\lambda)$, we get

$$\exp\left(N \int d\lambda \frac{\sqrt{2f_R''(1)-\lambda^2}}{\pi f_R''(1)} \ln |\lambda - f_R'(1)e|\right). \quad (88)$$

Putting everything together, we finally get

$$\begin{aligned} \langle \mathcal{N}(E) \rangle = & \exp\left(N \left[\left(\frac{1}{2} \ln \pi - \frac{1}{2} \ln \frac{1}{2} + \frac{1}{2} \right) - \frac{1}{2} \ln \pi f_R'(1) - \frac{e^2}{f_R(1)} + \right. \right. \\ & \left. \left. + \int d\lambda \frac{\sqrt{2f_R''(1)-\lambda^2}}{\pi f_R''(1)} \ln |\lambda - f_R'(1)e| \right] \right) \equiv \exp(N\Sigma(e)), \end{aligned} \quad (89)$$

where we used that $S_N(\sqrt{N}) =_{N \gg 1} (2\pi e)^{N/2}$ in the first step, the second and third terms come from the second step, and the last one from the third step.

We can now use this expression to see how the complexity behaves for some particular examples, starting with the case in which $f_D(x) = 0$ and $f_R(x) = x^p$, such that $E(s) = -\sum_{i_1, \dots, i_p} J_{i_1 \dots i_p} s_{i_1} \dots s_{i_p}$, which is the well-known spherical p -spin model. For such a problem it was proven, first with arguments from physics in [15] and then rigorously in [19, 20, 21], that the annealed and the quenched averages coincide; i.e. $\Sigma_A(e) = \Sigma_Q(e)$. In such a case one can also compute the complexities of specific kinds of critical points; namely separating them depending on their index value l , describing the number of negative directions along which the function increases. For example, $\Sigma_{l=0}$ is the complexity of the minima, while $\Sigma_{l=1}$ is the complexity of the saddle points with just one direction going down and all the others going up, and so on. What one finds, depicted in Fig. 5, is that there exists a threshold e_{thr} under which the minima dominate ($\Sigma_A = \Sigma_0$), and the eigen-spectrum of the Hessian is a shifted semi-circle that does not touch zero, and over which there are no minima with probability 1 but only saddles with a finite fraction of directions which go down, and the Hessian's eigen-spectrum includes negative values. This implies that the energy landscape shows a sub-exponential number of minima for $e = e_{GS}$, then for $e_{GS} < e < e_{thr}$ an exponential number of minima appear, until at $e = e_{thr}$ they become marginally stable and after this point they disappear and the landscape becomes dominated by saddles with a finite fraction of directions going down.

4.2.2. Results for the general case Let us now move back to the more general case in which we consider also a deterministic part f_D , which favors configurations in the direction of the signal v . Without proving them, we now show the results of the Kac-Rice method previously introduced, in three interesting cases.

First, let us discuss the case in which $f_D'(0) > 0$, which has as simplest example $f_D(x) = x$, corresponding to the spherical p -spin model in an external magnetic field of strength r . The behaviour with r of the energy landscape in this case is displayed in Fig. 6. For $r = 0$

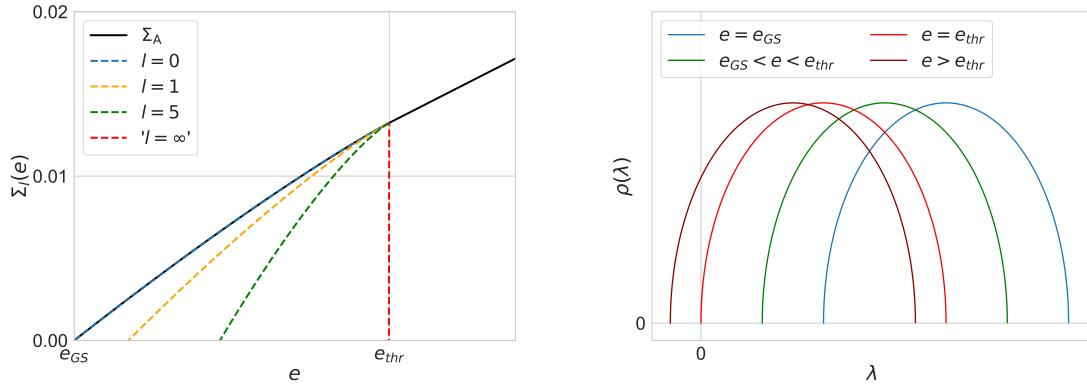


Figure 5. Left Panel: Behavior of the complexity of critical points for the spherical p -spin model with $p = 3$. The plot was inspired from [22].

Right Panel: Qualitative behaviour of the density of eigenvalues of the Hessians, given by shifted semi-circles, for different values of the energy e .

(the first sphere on the left), there is an exponential number of minima around the equator and the deepest ones are located exactly at the equator, which is also the parallel where the most numerous minima are located. When increasing r , the strip containing all the minima moves toward the north pole and starts shrinking, while the deepest minima are on a parallel closer to the north pole as soon as $r > 0$ and thus the most numerous ones are on a different parallel with smaller latitude. By increasing r , the landscape becomes

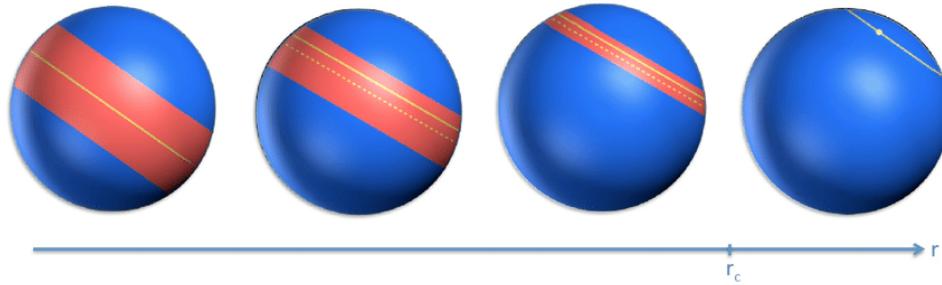


Figure 6. Behaviour of the energy landscape in the case in which $f'_D(0) > 0$. This drawing, taken from [23], illustrates the evolution of the energy landscape due to the increase of r . The red strip denotes the region on the sphere where minima lie in an exponential number. The continuous yellow line corresponds to the parallel where the deepest minima are located. The dashed yellow line corresponds to the parallel where the most numerous minima are located. At r_c , the energy landscape has a transition: For $r < r_c$, it is rough and full of minima; for $r > r_c$, it is smooth and contains only one minimum (represented by the yellow dot in the figure).

smoother due to a larger deterministic term, and, accordingly, the number of minima and the strip where they are located shrinks until reaching a value r_c , called *trivialization point*, above which only one minimum remains and all the other critical points disappear. In this case, the random contribution due to the first term in the Hamiltonian is no longer

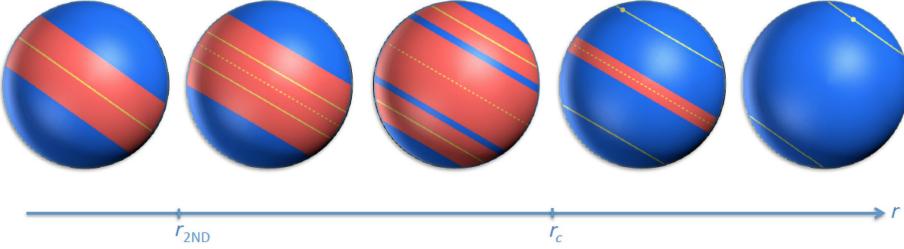


Figure 7. Behaviour of the energy landscape in the case in which $f'_D(0) = 0$ and $f''_D(0) > 0$. This drawing, taken from [23], illustrates the evolution of the energy landscape due to the increase of r . The red strip denotes the region on the sphere where minima lie in an exponential number. The continuous (dashed) yellow line corresponds to the parallel where the deepest (most numerous) minima are located. The energy landscape has several transitions. At r_{2ND} , the deepest minima are no longer on the equator and move toward the poles. Afterwards, the band containing the exponential number of minima fractures into three parts, one around the equator and two symmetric ones closer to the poles. At r_c , the bands closer to the pole implode and are replaced by two isolated global minima (the one on the south hemisphere is not visible, since it is on the back of the sphere), but the band at the equator persists. Finally, for even larger values of r , the landscape becomes completely smooth with only two symmetric minima.

strong enough to create a rugged landscape but still deforms it sufficiently to move the global minimum at a finite overlap with the signal v .

The second interesting case is the one in which $f_D(x)$ has a vanishing derivative in $x = 0$ but a finite second derivative, monotonically increasing from $x = 0$ to $x = 1$. As a simple example, in Fig. 7 we consider the case $f_D(x) = x^2/2$, which corresponds to a p -spin spherical model with an extra ferromagnetic interaction among spins, with r playing the role of the coupling. The landscape for $r = 0$ is the same as before, but by increasing r , the strip containing all the minima widens, and the deepest ones and the most numerous ones remain stuck on the equator. This situation persists until $r = r_{2ND}$, at which a *second order phase transition* takes place at the bottom of the landscape; i.e. by increasing r above r_{2ND} , the deepest minima continuously detach from the equator. For higher values of r , the strip separates into three bands, two closer to the north and south poles, respectively, to which the deepest minima belong, and one around the equator where the most numerous ones are located. At $r = r_c$, the trivialization transition happens, such that the two bands closer to the north and south poles containing an exponential number of minima shrink to zero and are replaced by an isolated global minimum per hemisphere. Finally, at even larger values of r , all minima around the equator disappear, and a final transition toward a fully smooth landscape characterized by only two minima takes place.

Finally, the last case we discuss is the one in which $f'_D(0) = f''_D(0) = 0$, and the simplest example of such a function is $f_D(x) = x^k$ with $k \geq 3$. In Fig. 8, we focus on this case taking $k = p = 3$, corresponding to the spiked-tensor model [24, 25]. The particularity of this case is that the critical points on the equator are not affected at all by the deterministic perturbation: they remain stable and unperturbed for any finite value of r . Therefore,

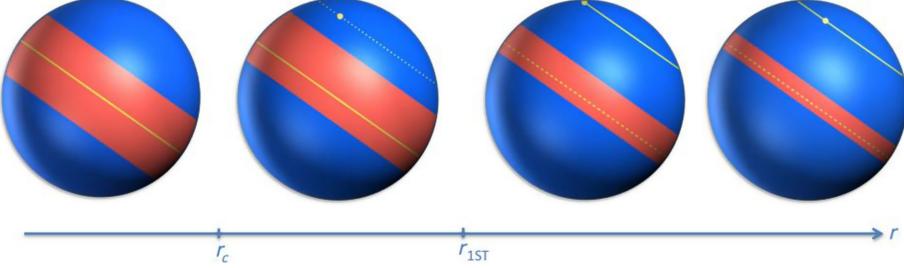


Figure 8. Behaviour of the energy landscape in the case in which $f_D(x) = x^3/3$ and $p = 3$. This drawing, taken from [23], illustrates the evolution of the energy landscape due to the increase of r . The red strips denote the regions on the sphere where minima lie in an exponential number. The continuous yellow line corresponds to the parallel where the global minimum is located. At r_C , an isolated local minimum appears. The dotted yellow line denotes that it is not yet the global one. At $r_{1ST} > r_C$, the deepest minimum is no longer on the equator and switches discontinuously to the isolated one close to the north pole. For larger values of r , the global minimum approaches the north pole, and the band around the equator shrinks but does not disappear for any finite r . The most numerous states, denoted by a dashed line, are always located on the equator.

there is always a strip of minima around the equator. One can see that, starting from the same landscape at $r = 0$, a band of minima, growing with r , is found around the equator. At a value r_C , an isolated minimum detaches from the top of the band, and for larger values of r it moves to higher latitudes, while the rest of the band shrinks around the equator. The deepest minima are located on the equator and are the ones of the original (unperturbed) p -spin model until a value of r , that we call r_{1ST} , is reached. When r reaches this value, the global minimum switches from the equator to the single minimum outside the band and close to the north pole. Increasing r further, the isolated global minimum approaches the north pole, and the band around the equator shrinks but never disappears for any finite r .

4.2.3. Further works and applications The Kac-Rice method has a vast range of applicability. It has been used not only in physics, but also for example in ecology [26, 27] to study the number of equilibria in large complex system. Recent reviews can be found in [28, 29], while the quenched computation using Kac-Rice plus replica is in [23]. Regarding neural networks, one of the first application of the Kac-Rice method was to compute the number of fixed points for random recurrent neural networks [30]. Recently, in [31] a teacher-student network was considered, with a loss

$$\mathcal{L}_{T-S}(\mathbf{w}) = \sum_{\mu} \ell(\xi_{\mu}^{\top} \mathbf{w}^*, \xi_{\mu}^{\top} \mathbf{w}), \quad (90)$$

where $\xi_{\mu}^i \sim \mathcal{N}(0, 1)$. This was a methodological advance since the the loss is non-Gaussian. The Kac-Rice method was generalized to analyse this case.

5. Dynamical mean-field theory for the perceptron model

In this lecture, we introduce dynamical mean-field theory (DMFT). This technique has a long history in statistical physics, where it has been used to analyse the high-dimensional dynamics of strongly correlated disordered systems [32, 33, 34]. DMFT has also been widely employed in condensed matter theory to describe strongly correlated electrons [35, 36] resulting in significant advances in both theory and applications.

This method has a great potential to study the dynamics of high-dimensional problem in machine learning. We introduce it in a simple setting, corresponding to the perceptron problem. We focus on gradient-flow dynamics in a prototypical learning problem, namely the teacher-student perceptron in dimension N . The training dataset $\mathcal{D} = \{(\xi_\mu, y_\mu)\}_{\mu=1}^M$ of size $M = \alpha N$, $\alpha \sim \mathcal{O}(1)$, is made of N -dimensional i.i.d. Gaussian samples $\xi_\mu^i \sim \mathcal{N}(0, 1/N)$, $\forall i = 1, \dots, N$, $\forall \mu = 1, \dots, M$, and teacher-generated labels: $y_\mu = \phi(\xi_\mu^\top \mathbf{w}^*)$. The teacher vector \mathbf{w}^* is drawn uniformly at random on the hypersphere of radius \sqrt{N} : $\mathbf{w}^* \in \mathcal{S}^{N-1}(\sqrt{N})$. The optimization is performed via gradient descent on the empirical risk:

$$\mathcal{L}(\mathbf{w}) = \sum_{\mu=1}^M \ell(\xi_\mu^\top \mathbf{w}^*, \xi_\mu^\top \mathbf{w}), \quad (91)$$

where the weight vector \mathbf{w} is constrained on the hypersphere $\mathbf{w} \in \mathcal{S}^{N-1}(\sqrt{N})$ at each step of the dynamics. For simplicity, we have incorporated the dependence on the activation function ϕ in the loss function ℓ . This formulation encompasses different widely-studied settings, some celebrated examples being *binary teacher-student classification*: $\phi(\cdot) = \text{sign}(\cdot)$ [37], and the *sign retrieval problem*: $\phi(\cdot) = (\cdot)^2$, i.e., *phase retrieval* in real space (see [38] and references therein).

The initial condition is drawn at random $\mathbf{w}(0) \sim P_0$, where the initial distribution P_0 does not depend on the dataset. The gradient flow dynamics is defined as follows:

$$\frac{d}{dt} \mathbf{w}(t) = -\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}(t)) - \lambda(t) \mathbf{w}(t) \quad (92)$$

where $\lambda(t)$ is a Lagrange multiplier enforcing the spherical constraint $\sum_{i=1}^N w_i(t)^2 = N$ at all training times. We aim at describing the system in the infinite-dimensional (a.k.a. *thermodynamic*) limit $N \rightarrow \infty$. To this end, we look for low dimensional order parameters evolving according to a self-consistent equation that effectively characterizes the dynamics of the high-dimensional system.

5.1. The dynamical cavity method

In order to achieve this effective characterization, we employ the *dynamical cavity method* [39, 40, 41]. In particular, we follow the derivation introduced in [41] in the context of high-dimensional interacting particle systems. We start by identifying the three reference directions that are relevant in our problem:

- The “lab frame”, a fixed direction \mathbf{v}_i uncorrelated with the examples $\{\xi_\mu\}_{\mu=1}^M$, such that $\mathbf{w}^\top \mathbf{v}_i := w_i$, $\|\mathbf{v}_i\| = 1$;
- The special directions given by the examples ξ_μ , that enter in the loss, such that $\mathbf{w}^\top \xi_\mu := w_\mu$, where $\|\xi_\mu\| \rightarrow 1$ in the infinite dimensional limit;
- The signal (or teacher) \mathbf{w}^* : we call the teacher-student alignment $m := \frac{\mathbf{w}^\top \mathbf{w}^*}{\sqrt{N} \|\mathbf{w}^*\|}$ the *magnetization*, in line with the physics terminology.

We now proceed to write a simple self-consistent stochastic process along these key directions. For simplicity, we focus on the random case, where no teacher is present and hence there is no correlation between the random data and labels. At the end, we will discuss how the result modifies if the teacher is introduced and the related generalization properties. In the random case, the loss is only a function of the scalar product $\xi_\mu^\top \mathbf{w}$. We start by writing the dynamics along the first (lab-frame) direction:

$$\frac{d}{dt} w_i(t) = - \sum_{\mu=1}^M \ell'(\xi_\mu^\top \mathbf{w}(t)) \xi_\mu^i - \lambda(t) w_i(t), \quad \text{where} \quad \xi_\mu^i = \xi_\mu^\top \mathbf{v}_i. \quad (93)$$

The remaining degrees of freedom $\mathbf{w}_\perp = \mathbf{w} - w_i \mathbf{v}_i$, orthogonal to \mathbf{v}_i , follow the dynamics

$$\frac{d}{dt} \mathbf{w}_\perp(t) = - \sum_{\mu=1}^M \ell'(\xi_\mu^\top \mathbf{w}(t)) \xi_\mu^\perp - \lambda(t) \mathbf{w}_\perp(t), \quad \text{where} \quad \xi_\mu^\perp = \xi_\mu - \xi_\mu^i \mathbf{v}_i. \quad (94)$$

Similarly as when deriving the Langevin equation from Newton equations, we can solve the equations for the remaining degrees of freedom \mathbf{w}_\perp (akin to the “environment”) at fixed w_i , and then plug this solution in Eq. (93) in order to get a closed equation on w_i . We remind that $\xi_\mu^i \sim \mathcal{N}(0, 1/N)$, therefore at large N , by the central limit theorem: $\xi_\mu^i w_i \sim \mathcal{O}(1/\sqrt{N})$ and $\mathbf{w}_\perp^\top \xi_\mu^\perp \sim \mathcal{O}(1)$. It follows that $\xi_\mu^\top \mathbf{w} = \mathbf{w}_\perp^\top \xi_\mu^\perp + \mathcal{O}(1/\sqrt{N})$ and we can solve for \mathbf{w}_\perp by using perturbation theory up to linear order. The zeroth-order term in perturbation theory reads

$$\frac{d}{dt} \mathbf{w}_\perp^0 = - \sum_{\mu=1}^M \ell'((\xi_\mu^\perp)^\top \mathbf{w}_\perp^0(t)) \xi_\mu^\perp - \lambda(t) \mathbf{w}_\perp^0(t) \quad (95)$$

with random initial condition on $\mathbf{w}_\perp^0 \in \mathbb{R}^{N-1}$. Note that Eq. (95) has a well-defined solution for \mathbf{w}_\perp^0 . We can now compute the solution for \mathbf{w}_\perp as a linear order perturbation to \mathbf{w}_\perp^0 . To this end, we add the infinitesimal field $h_\mu(t) = \xi_\mu^i w_i(t)$ to the argument of the loss function: $\ell((\xi_\mu^\perp)^\top \mathbf{w}_\perp^0(t)) \leftarrow \ell((\xi_\mu^\perp)^\top \mathbf{w}_\perp^0(t) + h_\mu(t))$. We find

$$\mathbf{w}_\perp(t) = \mathbf{w}_\perp^0(t) + \sum_{\mu=1}^M \int_0^t dt' \frac{\delta \mathbf{w}_\perp^0(t)}{\delta h_\mu(t')} \Big|_{h_\mu=0} \xi_\mu^i w_i(t'), \quad (96)$$

and we neglect higher order terms in the perturbation. At this point, we can plug the solution for \mathbf{w}_\perp given by Eq. (96) into Eq. (93) to obtain a closed equation for the dynamics

of w_i :

$$\frac{d}{dt} w_i(t) = - \sum_{\mu=1}^M \ell' \left((\xi_\mu^\perp)^\top \mathbf{w}_\perp(t) + \xi_\mu^i w_i(t) \right) \xi_\mu^i - \lambda(t) w_i(t) \quad (97)$$

$$= \underbrace{- \sum_{\mu=1}^M \ell' \left((\xi_\mu^\perp)^\top \mathbf{w}_\perp^0(t) \right) \xi_\mu^i}_{\text{(I): Random force}} - \underbrace{\sum_{\mu, \mu'=1}^M \int_0^t dt' \frac{\delta \ell' \left((\xi_\mu^\perp)^\top \mathbf{w}_\perp^0(t') \right)}{\delta h_{\mu'}(t')} \Big|_{h_{\mu'}=0} \xi_\mu^i \xi_{\mu'}^i w_i(t')}_{\text{(II): Retarded friction}} \quad (98)$$

$$- \underbrace{\sum_{\mu=1}^M \ell'' \left((\xi_\mu^\perp)^\top \mathbf{w}_\perp^0(t) \right) (\xi_\mu^i)^2 w_i(t) - \lambda(t) w_i(t)}_{\text{(III): Dynamic renormalization of regularization}} \quad (99)$$

that is correct up to linear order in the perturbation. We now briefly comment on the physical meaning of the different terms appearing in Eq. (99).

(I) The zeroth-order term in the perturbation is usually called *random force* in physics:

$$F_i(t) = - \sum_{\mu=1}^M \ell' \left((\xi_\mu^\perp)^\top \mathbf{w}_\perp^0(t) \right) \xi_\mu^i.$$

Notice that $(\xi_\mu^\perp)^\top \mathbf{w}_\perp^0$ is uncorrelated from ξ_μ^i , therefore $F_i(t)$ is a Gaussian function in the thermodynamic limit, with zero mean $\langle F_i(t) \rangle = 0$ and covariance

$$\begin{aligned} \langle F_i(t) F_i(t') \rangle &= \left\langle \sum_{\mu=1}^M (\xi_\mu^i)^2 \ell' \left((\xi_\mu^\perp)^\top \mathbf{w}_\perp^0(t) \right) \ell' \left((\xi_\mu^\perp)^\top \mathbf{w}_\perp^0(t') \right) \right\rangle_{\mathbf{w}_\perp, \xi_\mu} \\ &= \frac{1}{N} \sum_{\mu=1}^M \left\langle \ell' \left((\xi_\mu^\perp)^\top \mathbf{w}_\perp^0(t) \right) \ell' \left((\xi_\mu^\perp)^\top \mathbf{w}_\perp^0(t') \right) \right\rangle_{\mathbf{w}_\perp, \xi_\mu^\perp} \\ &= \alpha \left\langle \ell' \left(\xi_\mu^\top \mathbf{w}(t) \right) \ell' \left(\xi_\mu^\top \mathbf{w}(t') \right) \right\rangle_{\mathbf{w}, \xi_\mu} := M(t, t'). \end{aligned}$$

The last equality is obtained by observing that all the examples denoted by $\mu \in \{1, \dots, M\}$ are statistically equivalent, and putting back in the loss argument the $\mathcal{O}(1/\sqrt{N})$ correction that is negligible in the average.

(II) The second term contains the correction to linear order in the perturbation. The contribution coming from $\mu = \mu'$ concentrates in the high-dimensional limit:

$$\begin{aligned} &- \sum_{\mu=1}^M \int_0^t dt' \frac{\delta \ell' \left((\xi_\mu^\perp)^\top \mathbf{w}_\perp^0(t) \right)}{\delta h_\mu(t')} \Big|_{h_\mu=0} (\xi_\mu^i)^2 w_i(t') \\ &\xrightarrow{N \rightarrow \infty} -\alpha \int_0^t dt' \left\langle \frac{\delta \ell' \left(\xi_\mu^\top \mathbf{w}(t) \right)}{\delta h_\mu(t')} \Big|_{h_\mu=0} \right\rangle_{\mathbf{w}, \xi_\mu} w_i(t') := -\alpha \int_0^t dt' R(t, t') w_i(t'), \end{aligned}$$

where we have reintroduced the perturbation in the loss argument without changing the result, similarly as before. We have denoted the above average by $R(t, t')$ since

it is a response function. Notice that the terms in $\mu \neq \mu'$ vanish. Overall the term (II) plays the role of a *dissipation* or *retarded friction*. This contribution is also called *Onsager reaction term*.

(III) The third term concentrates in a similar way

$$-\sum_{\mu=1}^M \ell''((\xi_\mu^\perp)^\top \mathbf{w}_\perp^0(t)) (\xi_\mu^i)^2 w_i(t) \xrightarrow{N \rightarrow \infty} -\alpha \langle \ell''(\xi_\mu^\top \mathbf{w}(t)) \rangle_{w, \xi_\mu} w_i(t) := -\alpha \nu(t) w_i(t),$$

and results in a dynamical renormalization of the regularization term.

Finally, we can regroup all the terms above and write the stochastic equation for the dynamics of w_i :

$$\frac{d}{dt} w_i(t) = -\alpha \int_0^t dt' R(t, t') w_i(t') + F_i(t) - (\lambda(t) + \alpha \nu(t)) w_i(t), \quad (100)$$

with random Gaussian initial condition $w_i(0) \sim \mathcal{N}(0, 1)$. The noise comes from the Gaussian force $F_i(t)$, with zero mean and covariance $\langle F_i(t) F_i(t') \rangle = M(t, t')$.

At this point, we need one last ingredient to close the equations. Indeed, the argument of the loss function depends on the projection of the weights \mathbf{w} onto the direction of the examples ξ_μ . Therefore, we need to derive a dynamical equation for this order parameter. The dynamics along the direction ξ_μ is given by

$$\frac{d}{dt} w_\mu = -\sum_{\mu'=1}^M \ell'(\xi_{\mu'}^\top \mathbf{w}(t)) \xi_\mu^\top \xi_{\mu'} - \lambda(t) w_\mu(t) \quad (101)$$

$$= -\sum_{\mu'(\neq \mu)} \ell'(\xi_{\mu'}^\top \mathbf{w}(t)) \xi_\mu^\top \xi_{\mu'} - \lambda(t) w_\mu(t) - \underbrace{\ell'(\xi_\mu^\top \mathbf{w}(t)) \|\xi_\mu\|^2}_{=1}. \quad (102)$$

The above Eq. (102) is now formally identical to the one for w_i (Eq. (93)), with only two differences: the sum in this second case runs over $M-1$ examples ($\mu \neq \mu'$) and there is an additional term $-\ell'(\xi_\mu^\top \mathbf{w}(t))$. The first difference is negligible in the high-dimensional limit. Moreover, the examples $\xi_{\mu'}$ with $\mu' \neq \mu$ are uncorrelated from ξ_μ . Therefore, if there was no extra term in Eq. (102), the effective stochastic process for the dynamics of w_μ would be the same as in Eq. (100). Including the extra term, we obtain

$$\frac{d}{dt} w_\mu(t) = -\alpha \int_0^t dt' R(t, t') w_\mu(t') + F_\mu(t) - (\lambda(t) + \alpha \nu(t)) w_\mu(t) - \ell'(w_\mu(t)), \quad (103)$$

with Gaussian initial condition on $w_\mu(0)$ and the same definitions as in Eq. (100) for $R(t, t')$, $M(t, t') = \langle F_\mu(t) F_\mu(t') \rangle$ and $\nu(t)$:

$$R(t, t') := \left. \frac{\delta \langle \ell'(w_\mu(t)) \rangle_{w_\mu}}{\delta h_\mu(t')} \right|_{h_\mu=0}, \quad (104)$$

$$M(t, t') := \alpha \langle \ell'(w_\mu(t)) \ell'(w_\mu(t')) \rangle_{w_\mu}, \quad (105)$$

$$\nu(t) := \langle \ell''(w_\mu(t)) \rangle_{w_\mu}. \quad (106)$$

Finally, we need to specify an equation to compute the Lagrange multiplier $\lambda(t)$ enforcing the spherical constraint $\sum_{i=1}^N w_i(t)^2 = N$, $\forall t \geq 0$. By taking the derivative with respect to t on both sides and using Eq. (93), we find

$$\sum_{i=1}^N w_i(t) \frac{d}{dt} w_i(t) = - \sum_{\mu=1}^M \ell'(\xi_\mu^\top \mathbf{w}(t)) \xi_\mu^\top \mathbf{w}(t) - \lambda(t) \underbrace{\sum_{i=1}^N w_i(t)^2}_{=N} = 0. \quad (107)$$

Dividing both sides of the above Eq. (107) by N , we obtain that in the high-dimensional limit

$$\lambda(t) = -\alpha \langle \ell'(w_\mu) w_\mu \rangle_{w_\mu}. \quad (108)$$

It is important to remark that the dynamical mean-field equations derived above are expressed in terms of a self-consistent stochastic process. Indeed, the equations depend on the kernels R and M and the auxiliary function ν , that are in turn obtained as averages over the same stochastic process. By causality, it can be shown that the solution of the DMFT system is unique. This circular structure highlights the “mean-field” nature of these equations, similarly as the celebrated equation for the magnetization: $m = \text{th}(\beta m)$ for the Ising model at inverse temperature β .

Interestingly, the above equations – derived here with an heuristic method – have been put on rigorous ground in some cases [42, 43, 44, 45, 46].

Armed with the effective description of the high-dimensional gradient-descent dynamics provided by DMFT, we can now proceed to analyze the case where the dataset includes labels generated by a teacher vector \mathbf{w}^* , representing a prototype supervised learning problem. As previously anticipated, this modification introduces another important order parameter, i.e., the *teacher-student overlap* or *magnetization* $m(t) = \mathbf{w}(t)^\top \mathbf{w}^*$. One way to obtain an effective equation for $m(t)$ is to notice that $\sum_{i=1}^N w_i(t)^\top w_i^*/N \xrightarrow{N \rightarrow \infty} \langle w_i^* w_i(t) \rangle_{w_i, w_i^*}$. By multiplying Eq. (100) by w_i^* and taking the average, we find the following ODE:

$$\frac{d}{dt} m(t) = -\alpha \int_0^t dt' R(t, t') m(t') - (\lambda(t) + \alpha \nu(t)) m(t). \quad (109)$$

The above equations allow us to study the learning curves of the problem, for instance:

- The dynamical evolution of the average loss function:

$$\frac{1}{M} \mathcal{L}(\mathbf{w}(t)) = \frac{1}{M} \sum_{\mu=1}^M \ell(\xi_\mu^\top \mathbf{w}^*; \xi_\mu^\top \mathbf{w}(t)) \xrightarrow{N \rightarrow \infty} \langle \ell(w_\mu^*, w_\mu(t)) \rangle_{w_\mu};$$

- The dynamical evolution of the magnetization,⁺ that we can use to investigate signal recovery and its time scales;

⁺ In this spherical case, the generalization error: $\mathbb{E}[\mathbf{1}(\phi(\xi^\top \mathbf{w}) \neq y)]$, with $\mathbf{1}(\cdot)$ denoting the indicator function, is monotonic decreasing in the magnetization. Therefore, the magnetization captures all the relevant information on the performance.

- The properties of the correlation function: $C(t, t') = \sum_{i=1}^N w_i(t)w_i(t')/N \xrightarrow{N \rightarrow \infty} \langle w_i(t)w_i(t') \rangle$.

It is important to remark that we have started with a *deterministic* dynamics in high dimensions, and we have ended up with an effective dynamics involving a *random* force and a dissipation term. This is the result of isolating a representative variable while integrating out all the remaining degrees of freedom, as it happens when studying the dynamics of many physical systems (e.g., a molecule in a liquid, a spin in a magnetic material).

Another important point to underline is that the infinite-dimensional limit $N \rightarrow \infty$ is taken at *fixed* time window $[0, t]$. Therefore, this method cannot address timescales that diverge with the system size N .

5.1.1. Special case: the spherical spin glass It is instructive to mention a special case where the DMFT equations considerably simplify. This is the spherical p -spin glass model [47], described by the disordered long-range p -body Hamiltonian

$$H(\mathbf{s}) = - \sum_{i_1, i_2, \dots, i_p} J_{i_1, i_2, \dots, i_p} s_{i_1} s_{i_2} \dots s_{i_p}, \quad \text{with} \quad \sum_{i=1}^N s_i^2 = N, \quad (110)$$

where we have denoted the degrees of freedom by $\mathbf{s} \in \mathcal{S}^{N-1}(\sqrt{N})$ for historical reasons. The symmetric tensor J_{i_1, i_2, \dots, i_p} of rank p can be either drawn from a standard i.i.d. Gaussian distribution or generated by a teacher vector. We can implement the same procedure as above and obtain an effective equation for the i^{th} spin:

$$\frac{d}{dt} s_i(t) = \underbrace{F(t)}_{\text{random force}} + \frac{p(p-1)}{2} \int_0^t dt' R(t, t') C^{p-2}(t, t') s_i(t') - \lambda(t) s_i(t), \quad (111)$$

where

$$\langle F(t)F(t') \rangle := C(t, t') = \langle s_i(t)s_i(t') \rangle, \quad (112)$$

$$R(t, t') := \left. \frac{\delta \langle s_i(t) \rangle}{\delta h_i(t')} \right|_{h_i=0}. \quad (113)$$

The above equations are much simpler than the perceptron case since the effective stochastic process is a linear equation with additive noise and can be solved to write closed equations on the correlation and response functions C and R .

In summary, DMFT equations are a very powerful dimensional reduction tool to describe the dynamics of disordered systems at fixed time windows. Interesting extensions of the solution derived here are possible, for instance studying discrete variables with a Monte Carlo approach [48], Langevin noise* [49, 40], momentum-based accelerated methods [50], stochastic gradient descent [51].

* Notice that in the presence of Langevin noise in the dynamics particular care must be taken when computing derivatives, as in Eq. (107) where Itô's rule must be used.

A current bottleneck and important future development of the method regards the numerical implementation of the solution. The simplest strategy is to discretize the DMFT system, start by an initial guess for the kernels/auxiliary functions, use this guess to generate multiple realizations of the stochastic process, compute the averages and iteratively update the kernels $M_{\text{new}}(t, t') = (1 - \gamma)M_{\text{old}}(t, t') + \gamma M_{\text{new}}(t, t')$ until convergence, where γ is an appropriately chosen damping factor. This procedure was first implemented in the context of theoretical ecology [52]. A key challenge for future research is to improve these numerical solvers, similarly as what has already been done in the case of quantum problems, leading to great theoretical and application advances.

6. Energy barriers, entropic barriers, and signal recovery

In the following, we want to spell out some important features of gradient dynamics in high-dimensional non-convex landscapes. In the next paragraphs, we analyze the gradient flow algorithm for a simple toy model of signal recovery problems. First, we discuss how the gradient flow can be trapped by spurious local minima in a complex landscape, causing failure of signal recovery. Next, we consider a stochastic version of gradient flow dynamics and discuss the so-called "entropic barriers" coming from the dynamics' randomness as another cause of failure in recovery. At the end of this lecture, we discuss the connection with dynamics in deep neural networks.

6.1. Good and bad minima for gradient flow dynamics in high-dimensions

In this section, we discuss whether the gradient flow (GF) dynamics on non-convex energy landscapes can recover a signal hidden within a random noise following the works [49, 53, 54]. Within this setting we discuss the phenomenon of "bad and good minima" that is attracting a lot of attention in machine learning [55].

The model we consider here is the so-called spiked Matrix-Tensor model. Our task is to retrieve a signal on $(N - 1)$ -dimensional sphere $\sigma^* \in \mathbb{S}^{N-1} = \{\nu \in \mathbb{R}^N : \|\nu\| = 1\}$ by observing a matrix \mathbf{Y} and a tensor \mathbf{T} given by

$$\begin{aligned} T_{i_1 \dots i_p} &= \eta_{i_1 \dots i_p} + \sqrt{N(p-1)!} \sigma_{i_1}^* \dots \sigma_{i_p}^* \\ Y_{ij} &= \eta_{ij} + \sqrt{N} \sigma_i^* \sigma_j^*, \end{aligned} \tag{114}$$

where $\eta_{i_1 \dots i_p}$ and η_{ij} are independent centered Gaussian random variables with variance Δ_p and Δ_2 respectively. We here assume that $p > 2$. The maximum likelihood estimation of the signal σ^* corresponds to the minimization of an energy function $\mathcal{H} = \mathcal{H}_s + \mathcal{H}_p$, where the signal part \mathcal{H}_s and the noisy part \mathcal{H}_p are given by

$$\begin{aligned} \mathcal{H}_s &= -\frac{1}{\Delta_2 \sqrt{N}} \sum_{i < j} Y_{ij} \sigma_i \sigma_j \\ \mathcal{H}_p &= -\frac{\sqrt{(p-1)!}}{\Delta_p \sqrt{N}} \sum_{i_1 < i_2 < \dots < i_p} T_{i_1 i_2 \dots i_p} \sigma_{i_1} \dots \sigma_{i_p}. \end{aligned} \tag{115}$$

This minimization is performed by GF on the sphere S^{N-1} , initializing from a uniform random point on the sphere. To understand the performance of GF, we can apply the dynamical mean-field theory and the Kac-Rice formula [53]. These analyses result in the phase diagram Fig.9. In the bottom part of this phase diagram, it is impossible to recover the signal due to the absence of minima associated with the signal. In the left part of the figure with $1/\Delta_2 < 1.0$, there is a hard phase, where there exists a minimum with a finite correlation with the signal, but it is difficult to find it by the Approximate Message Passing algorithm (a very good algorithm for this problem). While the Approximate Message Passing algorithm succeeds in finding the signal above $1/\Delta_2 > 1.0$, GF fails below the dashed line. Interestingly, this line is well below the landscape trivialization threshold (the dotted line) computed by the Kac-Rice formula. This shows a fact that can be astonishing at first sight: GF can find the minimum associated with the signal while the energy landscape still has exponentially many spurious minima (grey-colored region). The mechanism behind this phenomenon can be understood by dynamical mean-field theory and the Kac-Rice method, as we shall explain below.

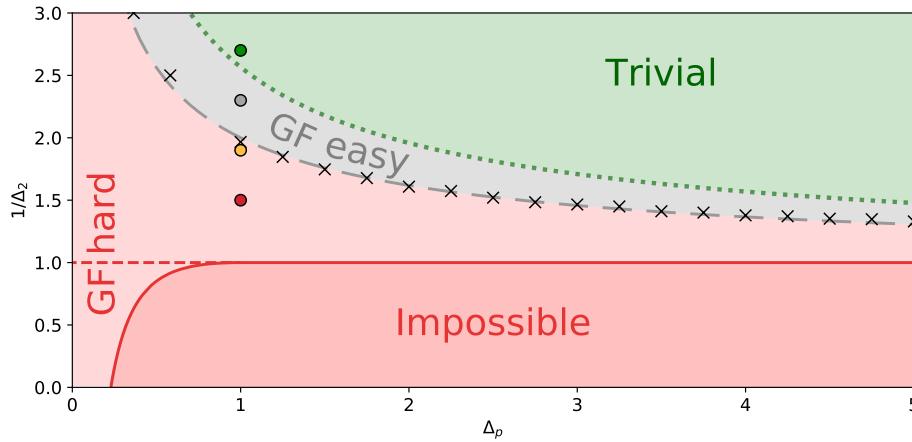


Figure 9. Phase diagram of the gradient flow for the spiked Matrix Tensor model. In the dark red region, it is information-theoretically impossible to obtain correlation with the signal[49]. In the light red region, while there exist minima correlated to the signal, the gradient flow typically cannot find them. In the green region, the energy landscape does not have spurious local minima. The gradient descent works well in the grey region between these two regions despite many spurious local minima. This figure is taken from S. Sarao Mannelli *et al.*, 2019 [53].

In the region of the phase diagram where GF fails, the numerical integration of the dynamical mean-field equation shows that GF relaxes into a certain energy level, well above the lowest energy (the red and yellow curves in the center panel of Fig.10). As the left panel of Fig.10 shows, the auto-correlation function $C(t, t')$ decreases as $t - t'$ increases but its relaxation time scale goes larger as the age of the system t' increases. This phenomenon is called aging in the physics literature [34]. We will discuss more in detail later. In this aging regime, we observe that the state after a long time $t \rightarrow \infty$ is marginally stable, i.e., the eigen-spectrum of its Hessian is a shifted-semicircle whose

left edge touches zero (the left panel of Fig.10). The Kac-Rice analysis reveals that the asymptotic energy (the dotted lines in the center panel of Fig.10) is the threshold energy where the left edge of the Hessian spectrum of typical critical points touches zero. These minima in this energy level are the most numerous and very flat; therefore, we can expect that they have large basins of attraction, which is why GF tends to converge to those points in this regime. When Δ_2 is large enough, similarly, the energy trajectory first converges to the threshold energy. Interestingly, however, it then suddenly drops within a finite time (the grey and green curves in the center panel of Fig.10), and eventually the state reaches a minimum with a finite overlap with the signal. This successful escape from the threshold states is due to their BBP transition, which can be seen by the Kac-Rice analysis. As the left panel of Fig.10 shows, when the signal-to-noise ratio Δ_2 is large enough, an isolated eigenvalue pops out from the bulk of Hessian's eigen-spectrum of the threshold states, and the corresponding eigenvector has a finite overlap with the signal σ^* . This isolated eigenvalue is strictly negative on the threshold energy, and hence the threshold states are typically unstable. This unstable direction has a finite overlap with the signal. By following it the system approaches the (good) minimum correlated with the signal. These two dynamical regimes (until the threshold and then toward the signal) are analogous to the search and convergence phase already discussed in Sec. 3.

Fig.11 summarizes the phase transitions as we increase the SNR. If the SNR is small, GF is typically trapped by the threshold states, which have large basins of attraction but no correlation with the signal. As we increase the SNR, however, the threshold states become unstable due to the BBP transition, and the unstable direction navigates the dynamics toward a minimum correlated to the signal. Once the SNR gets large enough, all the other spurious minima become unstable, and thus the energy landscape becomes trivial.

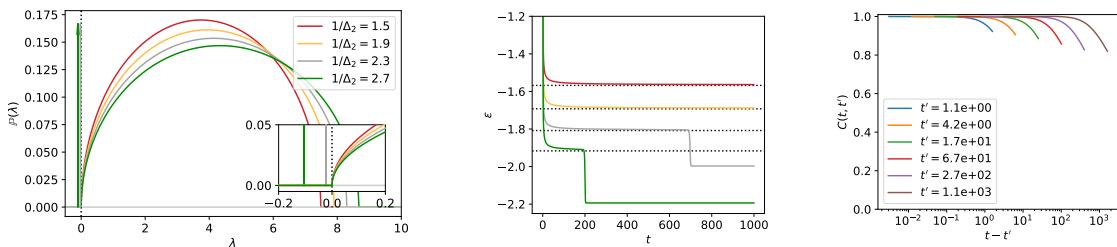


Figure 10. Left Panel: The hessian spectrum of threshold states with parameter values corresponding to the four points of the same color in Fig.9. When Δ_2 is large enough, a single isolated eigenvalue appears depicted as an arrow, destabilizing the threshold states. This destabilization allows the gradient flow dynamics suddenly go further down in the energy landscape, as we can observe in the center panel.

Center panel: The energy trajectories of the gradient flow dynamics. Each trajectory corresponds to the point on the phase diagram Fig.9 of the same color. The dotted line is the energy of threshold states computed by dynamical mean-field analysis and the Kac-Rice analysis.

Right panel: The correlation function $C(t, t')$ with $p = 3, \Delta_p = 1.0$ and $1/\Delta_2 = 1.5$ numerically obtained from the dynamical mean-field analysis. These figures are taken from S. Sarao Mannelli *et al.*, 2019 [53].

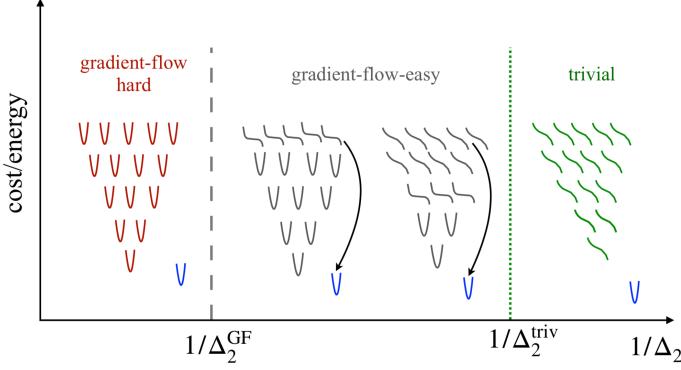


Figure 11. Cartoon of the phase transitions of the energy landscape of the spiked Matrix-Tensor model. This figure is taken from S. Sarao Mannelli *et al.*, 2019 [53].

6.2. Entropic barriers

This section discusses another mechanism hampering gradient descent dynamics: the so-called entropic barriers [56, 57, 58, 59]. We consider a very simple setting to discuss the main mechanism at play [56]: the online Stochastic Gradient dynamics with the Tensor-PCA model. At each time step, the Gaussian random tensor is independently chosen, i.e., the covariance between the tensor J^t sampled at time t and $J^{t'}$ sampled at time t' is given as follows

$$\langle J_{i_1, i_2, \dots, i_p}^t, J_{i_1, i_2, \dots, i_p}^{t'} \rangle = \frac{(p-1)!}{N^{p-1}} \delta(t - t'). \quad (116)$$

The gradient is given by

$$\frac{\partial E(\sigma | J^t)}{\partial \sigma_i} = - \sum_{i_1 < i_2 < \dots < i_{p-1}} \left(J_{i_1 \dots i_{p-1}}^t + J_{i_1 i_2 \dots i_{p-1}}^t + \dots + J_{i_1 \dots i_{p-1} i}^t \right) \sigma_{i_1} \dots \sigma_{i_{p-1}} - r \left(\sum_j \sigma_j \sigma_j^* \right)^{p-1} \sigma_i^*, \quad (117)$$

where r here is the signal-to-noise ratio. For simplicity, we here analyze the gradient flow, a continuous limit of discretized gradient descent. Let $\xi_i(t)$ denote the first term (without the negative sign) on the right-hand side. Then, the time derivative of σ_i is obtained as

$$\frac{d\sigma_i}{dt} = - \frac{\partial E(\sigma | J^t)}{\partial \sigma_i} = rm^{p-1}\sigma_i^* + \xi_i(t) - \lambda(t)\sigma_i, \quad (118)$$

where $m := N^{-1} \sum_j \sigma_j \sigma_j^*$, and $\lambda(t)$ is a Lagrange multiplier to constrain σ on the sphere S^{N-1} .

It is easy to see that $\xi_i(t)$ is a Gaussian noise with zero mean $\langle \xi_i(t) \rangle = 0$. The variance

$T := \langle \tilde{\xi}_i^2(t) \rangle$ has the order of $O(1)$, as is shown by the following calculation.

$$\begin{aligned} \langle \tilde{\xi}_i(t) \tilde{\xi}_i(t') \rangle &= \sum_{\substack{i_1 < i_2 < \dots < i_{p-1} \\ j_1 < j_2 < \dots < j_{p-1}}} \sigma_{i_1} \cdots \sigma_{i_{p-1}} \sigma_{j_1} \cdots \sigma_{j_{p-1}} \\ &\quad \times \langle (J_{i_1 \dots i_{p-1}}^t + \dots + J_{i_1 \dots i_{p-1} i}^t) (J_{j_1 \dots j_{p-1}}^{t'} + \dots + J_{j_1 \dots j_{p-1} i}^{t'}) \rangle \\ &= \frac{p!}{N^{p-1}} \sum_{i_1 < i_2 < \dots < i_{p-1}} \sigma_{i_1}^2 \cdots \sigma_{i_{p-1}}^2 \delta(t - t') \\ &= p \delta(t - t'), \end{aligned} \tag{119}$$

where we have neglected sub-leading terms in N (due to the large N limit).

The time derivative of m can be obtained from Eq.(118) by multiplying σ_i^*/N and taking the summation over index i , that is

$$\frac{dm}{dt} = rm^{p-1} - \lambda(t)m + A(t). \tag{120}$$

Here $A(t) := N^{-1} \sum_i \sigma_i^* \tilde{\xi}_i$ is of the order of $O(\frac{1}{\sqrt{N}})$. Next, we identify the Lagrange multiplier λ by imposing the condition of $\sum_i \sigma_i^2 = N$, which means $\frac{d}{dt} \sum_i \sigma_i^2 = 0$. Exploiting the Ito's formula,

$$\begin{aligned} \frac{1}{2N} \frac{d}{dt} \sum_i \sigma_i^2 &= \frac{1}{2N} \sum_i \frac{d\sigma_i^2}{d\sigma_i} \frac{d\sigma_i}{dt} + \frac{T}{2N} \sum_i \frac{d^2 \sigma_i^2}{d\sigma_i^2} \\ &= \frac{1}{N} \sum_i \sigma_i \frac{d\sigma_i}{dt} + T \\ &= rm^p - \lambda(t) + B(t) + T, \end{aligned} \tag{121}$$

where $B(t) := N^{-1} \sum_i \sigma_i \tilde{\xi}_i = O(1/\sqrt{N})$ is the stochastic part. Hence,

$$\lambda(t) = rm^p + B(t) + T. \tag{122}$$

Substituting this to Eq.(120),

$$\begin{aligned} \frac{dm}{dt} &= -(T + rm^p) + rm^{p-1} + A(t) + B(t) \\ &= -\frac{\partial V}{\partial m} + \text{noise}. \end{aligned} \tag{123}$$

where the effective potential is $V = Tm - r(1 - m^2)m^{p-1}$, drawn in Fig.12, and the noise is very weak (of order $1/\sqrt{N}$). When r is large enough, one finds two minima in the potential; one is at the origin, the other at large m and there is a barrier of order one separating them. Since $m(0) = O(1/\sqrt{N})$ with the random initialization, the dynamics start around the local minimum at the origin. Since the barrier height is much larger than the noise amplitude, one needs to wait an exponentially long time in N to climb up the barrier to find the global minimum at large m , i.e. the signal cannot be recovered in polynomial time in N . In consequence, the noise in the equation on m is completely ineffective and we can neglect it to understand the dynamical behavior. The barrier in the

potential is due to the first term proportional to the noise amplitude T . In fact, because of the noise, the system is kicked randomly at each time step by the drawn sample of the tensor, which typically brings the state back to the equator (i.e., the region of $m \approx 0$) since the measure on the sphere concentrates around its equator. Hence, this barrier is due to the large entropy of configurations around the equator, which we call an "Entropic Barrier". Note that when r is large enough $r/T \gg N^{(p-2)/2}$, the width of the entropic barrier is much smaller than $1/\sqrt{N}$, and therefore the dynamics start from the right side of the entropic barrier, in which case the global minimum is reachable.

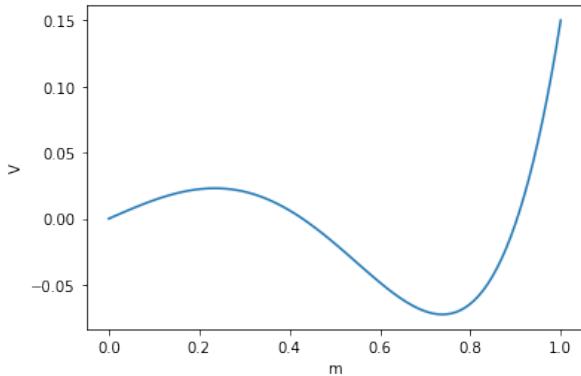


Figure 12. The effective potential function $V(m) = Tm - r(1-m^2)m^{p-1}$ shows the entropic barrier. The parameter values are chosen as follows: $T = 0.2, r = 1.0, p = 4$.

The main conclusion of this simple analysis is that noise is not always beneficial. In low dimensions it helps navigating the landscape and escaping bad minima, however in high dimensions it can bring the system in high-entropy configurations not correlated with the signal, whereas in other cases it can help if these regions have good generalization properties [60].

6.3. Numerical experiments in deep neural networks

In the previous sections, we analyzed simple theoretical models with techniques from glass physics. In this section, we discuss the empirical observation of deep neural networks comparing it with glassy aging dynamics, based on M. Baity-Jesi *et al.*, 2018 [61]. The aim of this work was to investigate whether the training dynamics is glassy or to what extent out of equilibrium using tools and observables developed in statistical physics. To this aim, we first briefly revisit the glassy dynamics of the spherical 3-spin model – an archetypical model of glasses. Its energy reads

$$E = - \sum_{\langle i_1, i_2, i_3 \rangle} J_{i_1, i_2, i_3} \sigma_{i_1} \sigma_{i_2} \sigma_{i_3}. \quad (124)$$

Here the summation goes over all the possible triplets of indexes running from 1 to N , and the coupling J_{i_1, i_2, i_3} are i.i.d. centered Gaussian random variables with variance $3/N^2$. The spin configuration σ is a N -dimensional vector on the sphere of radius \sqrt{N} . The plots in

Fig.13 correspond to stochastic Langevin dynamics under a quench from high temperature $T_i = \infty$ to low temperature $T_f = 0.5$. As we can observe in the left panel, the relaxation to the asymptotic energy is slower than exponential decay, which is a characterization of the aging phenomenon in glassy systems. Another characteristic of aging can be observed in the mean-square displacement, defined as

$$\Delta(t_w, t_w + t) = \frac{1}{N} \sum_{i=1}^N (\sigma_i(t_w) - \sigma_i(t_w + t))^2. \quad (125)$$

The right panel of Fig.13 shows the trajectories of the mean-square displacement against t with various fixed values of t_w . It clearly shows that as the age of the system t_w gets larger, it takes more time to decorrelate the system. This is another attribute of the aging phenomenon. Note that these phenomena are quite general, and are displayed by many physical glassy systems [62].

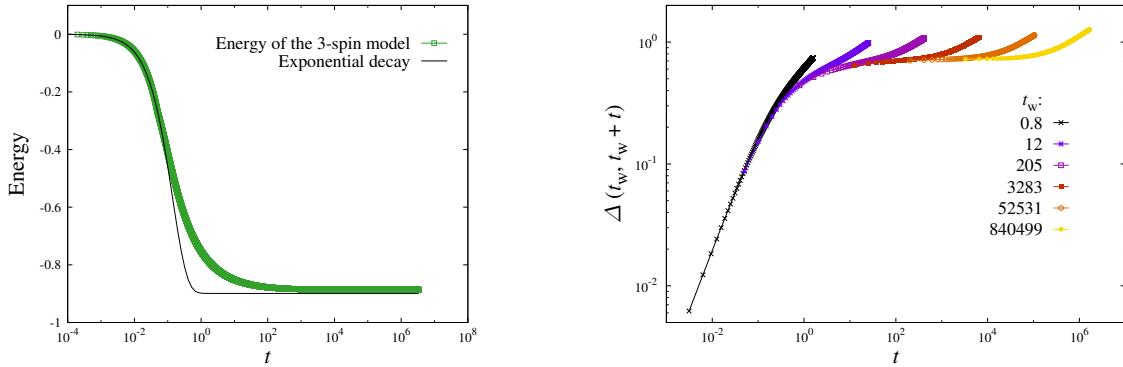


Figure 13. The trajectories of quenched dynamics of the 3-spin model from a high temperature $T_i = \infty$ to a low temperature $T_f = 0.5$. In the left panel, we plot the energy trajectory in comparison with an exponentially decaying curve. In the right panel, the mean square displacement is displayed for several values of t_w . These figures are taken from M. Baity-Jesi *et al*, 2018 [61].

Now we discuss the learning dynamics of the deep neural networks comparing it with the glassy dynamics. Fig.14 shows the trajectories of loss as a function of time steps when we train neural networks for CIFAR-10. In the left panel, the model is over-parameterized. This case shows three regimes during the training process. At first, it explores in the high-loss configurations up to $t = t_1$. Next, the loss decreases approximately linearly in $\log(t)$, and the accuracy increases similarly. This regime is up to time $t = t_2$, where the training loss goes to zero. In the last regime, the training loss always stays around zero. This behavior resembles the one discussed in the previous sections and consisting in a search and a convergence phase. For comparison, we show the under-parameterized case in the right panel of Fig.14, which shows different behavior from the one with over-parameterization. In this case, the dynamics resembles the one of glassy landscapes in which the system converges to bad minima and non-zero training loss.

Next, we discuss the mean-squared displacement of the neural networks' learning dynamics, shown in Fig.15. In the left panel, we show the case of the over-parameterized

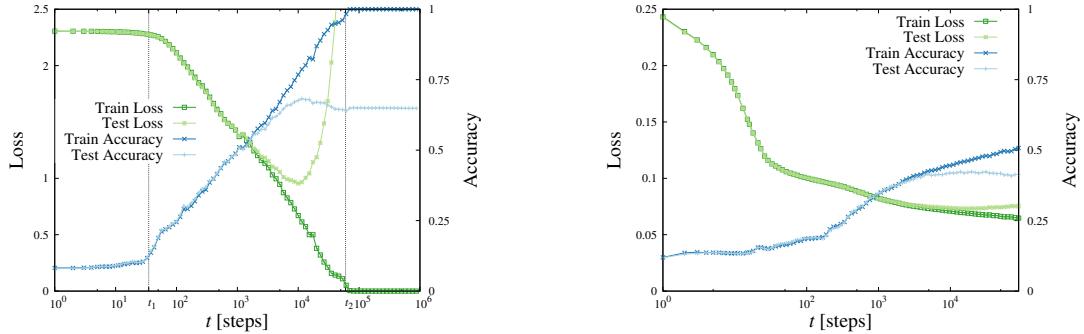


Figure 14. Train/test loss and accuracy of neural networks trained for CIFAR-10 as a function of $\log(t)$. The left and right panel displays the models with over-parameterization and under-parameterization, respectively. These figures are taken from M. Baity-Jesi *et al.*, 2018 [61].

networks trained for MNIST. The three different colors roughly correspond to the three regimes we discussed above. The intriguing observation here is that at the final stage of the dynamics (the yellow curves), up to the re-scaling of the noise amplitude $D(t_w)$, the mean-squared displacement is almost always independent of the age of the system t_w . On the other hand, the under-parameterized regime shown in the right panel of Fig.15 shows the aging phenomenon similar to glassy systems.

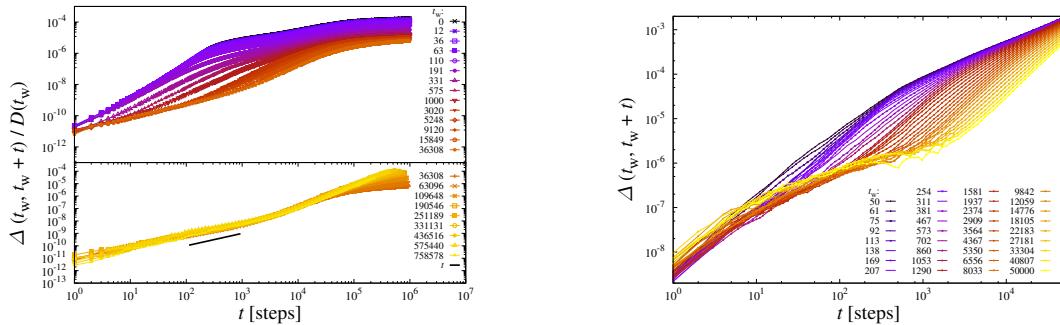


Figure 15. Mean-squared displacement of the learning trajectories of neural networks as a function of $\log(t)$. The left panel displays an over-parameterized model trained for MNIST, and the right panel displays an under-parameterized model trained for CIFAR-10. These figures are taken from M. Baity-Jesi *et al.*, 2018 [61].

The training dynamics of over-parameterized neural networks displays interesting phenomena. In the under-parametrized regime, one finds aging dynamics and slow convergence to bad minima, whereas in the over-parametrized one the dynamics it has a search phase and a convergence phase. At long times, it becomes stationary if one renormalizes the unit of time, corresponding to diffusion over the zero (or very small) training loss manifold. The theoretical understanding of the transitions between aging and non-aging dynamics and the three regimes during the training process has still to be completed, as shown here insights from physics can be helpful on this endeavour. Filling the gap between the theoretical toy models we understand so far and the empirical observations in the learning dynamics of neural networks is an important open problem.

Acknowledgments

We thank F. Krzakala and L. Zdeborová for organizing the summer school "Statistical Physics and Machine Learning".

Author contributions These lecture notes are based on a series of lectures give by G. Biroli at the Les Houches Summer School "Statistical Physics and Machine Learning". TB, DG, KK, FM, AY contributed equally by preparing these lecture notes. GB revised them.

Funding information GB acknowledge funding from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and from the Simons Foundation collaboration "Cracking the Glass Problem" (No. 454935 to G. Biroli). FM and KK were supported in part by the National Science Foundation, through the Center for the Physics of Biological Function (PHY-1734030). KK was also supported by a C.V. Starr Fellowship.

- [1] Eugene P. Wigner. Random matrices in physics. *SIAM Review*, 9(1):1–23, 1967.
- [2] Thomas Guhr, Axel Müller-Groeling, and Hans A Weidenmüller. Random-matrix theories in quantum physics: common concepts. *Physics Reports*, 299(4-6):189–425, 1998.
- [3] J. P. Bouchaud and M. Potters. Financial applications of random matrix theory: a short review, 2009.
- [4] Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022.
- [5] F. G. Tricomi. *Integral Equations*. Pure Appl. Math. V, Interscience, London, 1957.
- [6] Celine Nadal, Satya N Majumdar, and Massimo Vergassola. Phase transitions in the distribution of bipartite entanglement of a random pure state. *Physical review letters*, 104(11):110501, 2010.
- [7] V.A. Marcenko and L.A. Pastur. Distribution of Eigenvalues for Some Sets of Random Matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.
- [8] Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564, 1955.
- [9] Dan Voiculescu. Limit laws for random matrices and free products. *Inventiones mathematicae*, 104:201–220, 1991.
- [10] E. Brézin. Grassmann variables and supersymmetry in the theory of disordered systems. *Applications of Field Theory to Statistical Mechanics. Lecture Notes in Physics.*, 216:115–123, 1985.
- [11] László Erdős. Universality of wigner random matrices: a survey of recent results. *Russian Mathematical Surveys*, 66(3):507, 2011.
- [12] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.
- [13] Samuel F Edwards and Raymund C Jones. The eigenvalue spectrum of a large symmetric random matrix. *Journal of Physics A: Mathematical and General*, 9(10):1595, 1976.
- [14] Stéphane d’Ascoli, Maria Refinetti, and Giulio Biroli. Optimal learning rate schedules in high-dimensional non-convex optimization problems, 2022.
- [15] Andrea Cavagna, Irene Giardina, and Giorgio Parisi. Stationary points of the thouless-anderson-palmer free energy. *Phys. Rev. B*, 57:11251–11257, May 1998.
- [16] Emile Richard and Andrea Montanari. A statistical model for tensor pca. *Advances in neural information processing systems*, 27, 2014.
- [17] Yan V. Fyodorov. Complexity of random energy landscapes, glass transition, and absolute value of the spectral determinant of random matrices. *Phys. Rev. Lett.*, 92:240601, Jun 2004.
- [18] Valentina Ros, Giulio Biroli, and Chiara Cammarota. Complexity of energy barriers in mean-field glassy systems. *EPL (Europhysics Letters)*, 126(2):20003, 2019.
- [19] 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.
- [20] Eliran Subag. The complexity of spherical p -spin models—a second moment approach. *The Annals of Probability*, 45(5):3385–3450, 2017.
- [21] Eliran Subag and Ofer Zeitouni. Concentration of the complexity of spherical pure p -spin models at arbitrary energies. *Journal of mathematical physics*, 62(12):123301, 2021.
- [22] Antonio Auffinger and Julian Gold. The number of saddles of the spherical p -spin model, 2020.
- [23] 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. *Phys. Rev. X*, 9:011003, Jan 2019.
- [24] Gérard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica. The landscape of the spiked tensor model. *Communications on Pure and Applied Mathematics*, 72(11):2282–2330, 2019.
- [25] Thibault Lesieur, Leo Miolane, Marc Lelarge, Florent Krzakala, and Lenka Zdeborova. Statistical and computational phase transitions in spiked tensor estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017.
- [26] Gérard Ben Arous, Yan V Fyodorov, and Boris A Khoruzhenko. Counting equilibria of large complex systems by instability index. *Proceedings of the National Academy of Sciences*, 118(34):e2023719118, 2021.

- [27] Valentina Ros, Felix Roy, Giulio Biroli, Guy Bunin, and Ari M Turner. Generalized lotka-volterra equations with random, nonreciprocal interactions: The typical number of equilibria. *Physical Review Letters*, 130(25):257401, 2023.
- [28] Antonio Auffinger, Andrea Montanari, and Eliran Subag. Optimization of random high-dimensional functions: Structure and algorithms, 2022.
- [29] Valentina Ros and Yan V Fyodorov. The high-d landscapes paradigm: spin-glasses, and beyond. *arXiv preprint arXiv:2209.07975*, 2022.
- [30] Gilles Wainrib and Jonathan Touboul. Topological and dynamical complexity of random neural networks. *Physical review letters*, 110(11):118101, 2013.
- [31] Antoine Maillard, Gérard Ben Arous, and Giulio Biroli. Landscape complexity for the empirical risk of generalized linear models. In Jianfeng Lu and Rachel Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 287–327. PMLR, 20–24 Jul 2020.
- [32] H. Sompolinsky and Annette Zippelius. Dynamic theory of the spin-glass phase. *Phys. Rev. Lett.*, 47:359–362, Aug 1981.
- [33] H. Sompolinsky and Annette Zippelius. Relaxational dynamics of the edwards-anderson model and the mean-field theory of spin-glasses. *Phys. Rev. B*, 25:6860–6875, Jun 1982.
- [34] L. F Cugliandolo and J. Kurchan. Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model. *Phys. Rev. Lett.*, 71:173–176, Jul 1993.
- [35] Walter Metzner and Dieter Vollhardt. Correlated lattice fermions in $d = \infty$ dimensions. *Phys. Rev. Lett.*, 62:324–327, Jan 1989.
- [36] Antoine Georges, Gabriel Kotliar, Werner Krauth, and Marcelo J Rozenberg. Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions. *Reviews of Modern Physics*, 68(1):13, 1996.
- [37] Elizabeth Gardner and Bernard Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.
- [38] Jonathan Dong, Lorenzo Valzania, Antoine Maillard, Thanh-an Pham, Sylvain Gigan, and Michael Unser. Phase retrieval: From computational imaging to machine learning. *arXiv preprint arXiv:2204.03554*, 2022.
- [39] Marc Mézard, Giorgio Parisi, and Miguel A. Virasoro. *Spin glass theory and beyond*. World Scientific, Singapore, 1987.
- [40] Elisabeth Agoritsas, Giulio Biroli, Pierfrancesco Urbani, and Francesco Zamponi. Out-of-equilibrium dynamical mean-field equations for the perceptron model. *Journal of Physics A: Mathematical and Theoretical*, 51(8):085002, 2018.
- [41] Chen Liu, Giulio Biroli, David R Reichman, and Grzegorz Szamel. Dynamics of liquids in the large-dimensional limit. *Physical Review E*, 104(5):054606, 2021.
- [42] G Ben Arous and Alice Guionnet. Symmetric langevin spin glass dynamics. *The Annals of Probability*, 25(3):1367–1422, 1997.
- [43] G Ben Arous, Amir Dembo, and Alice Guionnet. Aging of spherical spin glasses. *Probability theory and related fields*, 120(1):1–67, 2001.
- [44] Gérard 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.
- [45] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.
- [46] Cedric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborova. Rigorous dynamical mean field theory for stochastic gradient descent methods. *arXiv preprint arXiv:2210.06591*, 2022.
- [47] Andrea Crisanti and H-J Sommers. The spherical-spin interaction spin glass model: the statics. *Zeitschrift für Physik B Condensed Matter*, 87(3):341–354, 1992.
- [48] H Eissfeller and M Opper. Mean-field monte carlo approach to the sherrington-kirkpatrick model with asymmetric couplings. *Physical Review E*, 50(2):709, 1994.

- [49] 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. *Physical Review X*, 10(1):011057, 2020.
- [50] Stefano Sarao Mannelli and Pierfrancesco Urbani. Analytical study of momentum-based acceleration methods in paradigmatic high-dimensional non-convex problems. *Advances in Neural Information Processing Systems*, 34:187–199, 2021.
- [51] Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. *Advances in Neural Information Processing Systems*, 33:9540–9550, 2020.
- [52] Felix Roy, Giulio Biroli, Guy Bunin, and Chiara Cammarota. Numerical implementation of dynamical mean field theory for disordered systems: Application to the lotka–volterra model of ecosystems. *Journal of Physics A: Mathematical and Theoretical*, 52(48):484001, 2019.
- [53] Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, and Lenka Zdeborová. Who is afraid of big bad minima? analysis of gradient-flow in spiked matrix-tensor models. *Advances in Neural Information Processing Systems*, 32, 2019.
- [54] 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. PMLR, 2019.
- [55] Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. Bad global minima exist and sgd can reach them. *Advances in Neural Information Processing Systems*, 33:8543–8552, 2020.
- [56] Giulio Biroli and Chiara Cammarota. *unpublished*, 2019.
- [57] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor pca. *The Annals of Probability*, 48(4):2052–2087, 2020.
- [58] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. *arXiv preprint arXiv:2206.04030*, 2022.
- [59] Matteo Bellitti, Federico Ricci-Tersenghi, and Antonello Scardicchio. Entropic barriers as a reason for hardness in both classical and quantum algorithms. *Physical Review Research*, 3(4):043015, 2021.
- [60] Brandon Livio Annesi, Clarissa Lauditi, Carlo Lucibello, Enrico M Malatesta, Gabriele Perugini, Fabrizio Pittorino, and Luca Saglietti. The star-shaped space of solutions of the spherical negative perceptron. *arXiv preprint arXiv:2305.10623*, 2023.
- [61] Marco Baity-Jesi, Levent Sagun, Mario Geiger, Stefano Spigler, Gérard Ben Arous, Chiara Cammarota, Yann LeCun, Matthieu Wyart, and Giulio Biroli. Comparing dynamics: Deep neural networks versus glassy systems. In *International Conference on Machine Learning*, pages 314–323. PMLR, 2018.
- [62] Ludovic Berthier and Giulio Biroli. Theoretical perspective on the glass transition and amorphous materials. *Reviews of modern physics*, 83(2):587, 2011.