Replica method for computational problems with randomness: principles and illustrations

J. Steinberg[1], U. Adomaitytė[2], A. Fachechi[3], P. Mergny[4], D. Barbier[4] and R. Monasson[5]

1 Center for the physics of biological function, Princeton University, New Jersey, USA
2 The KCL Disordered Systems Group, King's college, London, UK
3 Department of Mathematics, Sapienza University, Rome, Italy
4 Idephics Lab, EPFL, Lausanne, Switzerland
5 Laboratoire de Physique de l'Ecole Normale Supérieure, PSL and CNRS UMR8023, Sorbonne Université, Paris, France

I. LECTURE 1. WHAT ARE REPLICAS?

Systems with Random Interactions in Statistical Physics: Historical Background

The study of disordered systems in physics is historically related to the discovery of spin glasses, a peculiar class of magnetic materials. In standard paramagnetic systems, the magnetic moments carried by the electronic spins align when an external field is applied, but the effect disappears when the field vanishes. In ferromagnetic systems, however, the magnetization does not vanish, provided the temperature is lower than some critical value, the Curie temperature. The simplest model of a ferromagnetic system considers magnetic moments as vectors $\boldsymbol{S}_i$ of unit norms, located at the nodes i of a lattice. Each moment interacts with its nearest neighbors on the lattice, leading to the following energy function

$$E = -\sum_{i<j} J_{ij} \boldsymbol{S}_i \boldsymbol{S}_j \tag{1}$$

where $J_{ij} = J$ for nearest neighbour sites i, j, and 0 otherwise, is the strength of the interaction. The probability density of a given configuration is then given by the Boltzmann distribution

$$p(\boldsymbol{S}_1, \ldots, \boldsymbol{S}_N) = \frac{1}{Z[\beta, \{J_{ij}\}]}\, e^{\beta \sum_{i<j} J_{ij} \boldsymbol{S}_i \cdot \boldsymbol{S}_j} \tag{2}$$

where $\beta = \frac{1}{T}$ is the inverse temperature, and Z is the partition function ensuring normalization of p.

On cubic lattices in D dimensions, when $J > 0$, this model undergoes a phase transition at a critical temperature T_c in which all the spins begin to align along a preferential direction, breaking the global rotational symmetry of the energy E. In the case $J < 0$, a phase transition will still occur, but neighboring moments will align along opposite directions, a phenomenon called antiferromagnetism.

For more general lattices and/or more general interactions J_{ij}, the system can exhibit a phenomenon called frustration. Frustration occurs when, for some pairs i, j, the sign of $J_{ij} \boldsymbol{S}_i \cdot \boldsymbol{S}_j$ is negative, hence the corresponding bond increases the total energy of the spin configuration. For example, on a simple triangular lattice antiferromagnet with $N = 3$ spins, once the first two spins (or moments) are anti-aligned, the third one can obviously not be anti-aligned with both of them. In other words, frustration occurs when the multiple competing interactions between moments cannot be simultaneously satisfied. In frustrated systems, a greedy steepest descent of the energy E does not necessarily find the best configuration of moments minimizing the energy.

Interesting phenomena occur when frustration is pervasive. This can be accomplished in metal alloys, where iron atoms for example are randomly mixed into a grid of copper atoms. Each iron atom contributes a magnetic moment that interacts with other moments at distance R via a Ruderman–Kittel–Kasuya–Yoshida (RKKY) interaction

$$J(R) \propto (2k_F R)^{-2} \sin(2k_F R)\,, \tag{3}$$

where k_F is the Fermi momentum. The sign of the interaction change quickly with the distance, and can therefore be either positive or negative, hence leading to massive frustration. Such materials are called spin glasses, to emphasize the analogy between the disorder (randomness) in magnetic interactions and the positional disorder of particles in a conventional structural glass.

As the dynamics of spins is much faster than the ones of atoms in the materials, it is a reasonable approximation to assume that the interactions are frozen (quenched). We therefore consider the interactions J_{ij} between spins to be drawn at random and fixed for a given sample. Then, the distribution of spin configurations is given by Eq. (2), and depends on the specific realization of all the interactions J_{ij}. Fortunately, many observables of interest do not depend on the details

of the set of interactions attached to a sample, but only on their statistics. Such quantities are called self-averaging: their values are well described by the average over the distribution of couplings.

A historically important model of a spin glass was proposed by Sherrington and Kirkpatrick in 1974. In the so-called SK model, the spins take values $s_i = \pm 1$ and the interactions J_{ij} are drawn independently and at random from the Gaussian distribution

$$p(J_{ij}) = \sqrt{\frac{N}{2\pi J^2}} \exp\left\{-\frac{N}{2J^2}\left(J_{ij} - \frac{J_0}{N}\right)^2\right\}. \tag{4}$$

The mean value J_0/N determines the mean ferromagnetic interaction, while the large fluctuations of the order of $J/\sqrt{N}$ around this average can be negative or positive. Notice that, with these scalings, the energy E in Eq. (1) is expected to scale linearly with the number N of spins.

Analogies between disordered systems and problems in machine learning

The systems described above are characterized by two sets of variables. Quenched variables, such as the interactions J_{ij}, random, and one realization of these variables define the disordered sample. Thermalized (fast) variables, such as the spins s_i, adapt to these disordered background, and their distribution coincides with the Gibbs measure p at fixed temperature. A similar duality between variables takes place in many machine-learning models.

For example, in supervised learning, the training data consisting of a set of inputs with their corresponding outputs

$$D = \{\boldsymbol{x}_\mu, \boldsymbol{y}_\mu\} \tag{5}$$

can be interpreted as quenched variables. We can use this training set to learn the parameters of a parametric model $\boldsymbol{y} = f(\boldsymbol{x}, \theta)$ by minimizing the loss

$$L(\theta, D) = \sum_\mu \big(\boldsymbol{y}_\mu - f(\boldsymbol{x}_\mu, \theta)\big)^2 , \tag{6}$$

which is similar to an energy. The distribution over the 'thermalized' parameters θ after training at fixed inverse 'temperature' β is then given by

$$p(\theta|D) = \frac{1}{Z[D]} e^{-\beta L(\theta, D)} . \tag{7}$$

As a second example, consider unsupervised learning of a generative model, where the data are the set of items $D = \{\boldsymbol{x}_\mu\}$. For a parametric model, the likelihood is given by $p(\boldsymbol{x}|\theta)$ and the prior by $p_{prior}(\theta)$. After training, the posterior distribution of model parameters $p_{post}(\theta|D)$ is given by

$$p_{post}(\theta|D) = \frac{1}{Z[D]} p_{prior}(\theta) \prod_\mu p(\boldsymbol{x}_\mu|\theta) . \tag{8}$$

Again the data D can be seen as quenched variables, and the parameters θ as thermalized variables.

The Replica Method

Let us go back to spin models in the presence of quenched interactions. Suppose we want to compute the thermal average of some observable $\mathcal{O}$ that depends on the spin configurations $S = \{s_i\}$. For a specific realization $J = \{J_{ij}\}$ of the interactions, this is defined as

$$\langle \mathcal{O} \rangle(J) = \sum_S \mathcal{O}(S)\, p(S|J) = \sum_s \mathcal{O}(S)\, \frac{e^{-\beta E[S,J]}}{Z[\beta, J]} . \tag{9}$$

In the general case of self-averaging observables, we would like to average the expression above over the quenched variables, with the result

$$\overline{\langle \mathcal{O} \rangle} = \sum_S \mathcal{O}(S) \overline{\left(\frac{e^{-\beta E[S,J]}}{Z[\beta, J]}\right)} \tag{10}$$

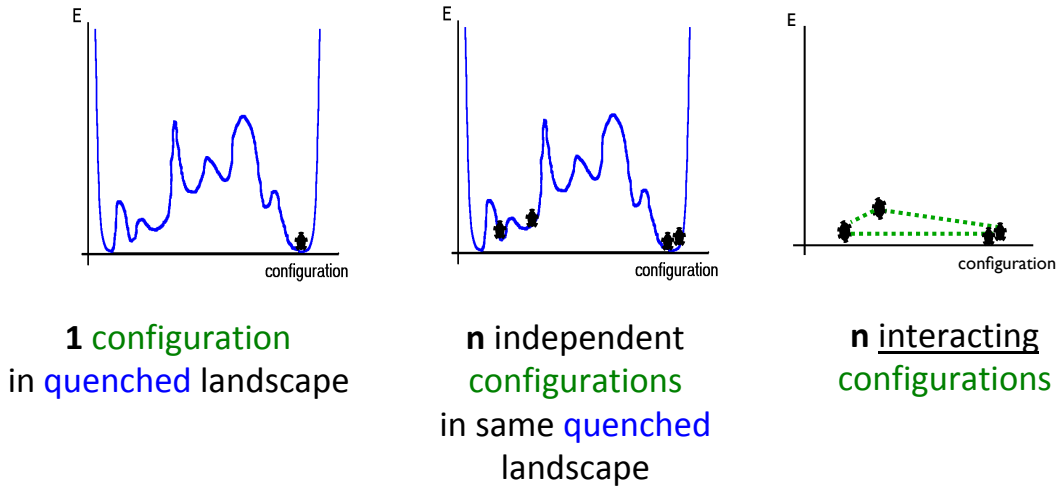

FIG. 1. Starting from one configuration in a quenched landscape, the replica trick creates n independent replicas lying in the same quenched landscape. Averaging over disorder produces n interacting replicas.

The average over J is generically very difficult to compute as the interactions appear both in the numerator and in the denominator. However, it turns out that this average can be performed using a technique called replica method. The replica method is a powerful tool that has been developed in the last decades to tackle disordered many-body problems and has provided solutions to theoretical problems in spin glass theory [30], [8], combinatorial optimization [6, 29], etc. Some of these solutions have been proven rigorously by mathematicians and mathematical physicists through alternative, probabilistic approaches [41].

The replica method considers a system consisting of n independent copies of the original system with the same realization of disorder

$$\frac{1}{Z[J]} = \lim_{n\to 0} Z[J]^{n-1} = \lim_{n\to 0} \sum_{S_2} \cdots \sum_{S_n} e^{-\beta \sum_{a=2}^{n} E[S_a,J]} \tag{11}$$

where the second equality holds for integer valued n (while the first one requires that $n \to 0$...). Thus, $\overline{\langle \mathcal{O} \rangle}$ formally becomes

$$\overline{\langle \mathcal{O} \rangle} = \lim_{n\to 0} \sum_{S_1} \cdots \sum_{S_n} O(S)\, \overline{e^{-\beta E[S,J] - \beta \sum_{a=2}^{n} E[S_a,J]}} = \lim_{n\to 0} \sum_{S_1} \cdots \sum_{S_n} O(S)\, \overline{e^{-E_{eff}[S_1,\ldots,S_n]}} \tag{12}$$

where the effective energy over the replica configurations

$$E_{eff}[S,\ldots,S_n] = -\log\left(\overline{e^{-\beta \sum_{a=1}^{n} E[S_a,J]}}\right) \tag{13}$$

is obtained after averaging over the quenched variables. Averaging over the disordered interactions transforms n independent replicas in the same quenched landscape into n interacting configurations. These induced interactions can be seen as a signature of the preexisting J-dependent landscape. Intuitively speaking, for a landscape with a deep attractive well, all replicas have a tendency to be similar (to the minimum of the potential), and their relative interactions should be very strong. On the contrary, for flat and smooth landscape, spin configurations can take very different values, and hence these replicas feel no interaction. This picture is sketched in Figure 1.

In general, for magnetic systems, the order parameter m measures the similarity between the average spin values and the ground state $\boldsymbol{g}$ (the configuration of spins minimizing the energy),

$$m = \frac{1}{N} \sum_{i}^{N} \langle s \rangle_i \, g_i \, . \tag{14}$$

For ferromagnetic systems, $g_i = +1$ (or -1) for all sites i, and this order parameter coincides with the magnetization. For an antiferromagnetic system on a cubic lattice, $g_i = +1$ and -1 alternatively on neighbour sites, and m is the staggered magnetization.

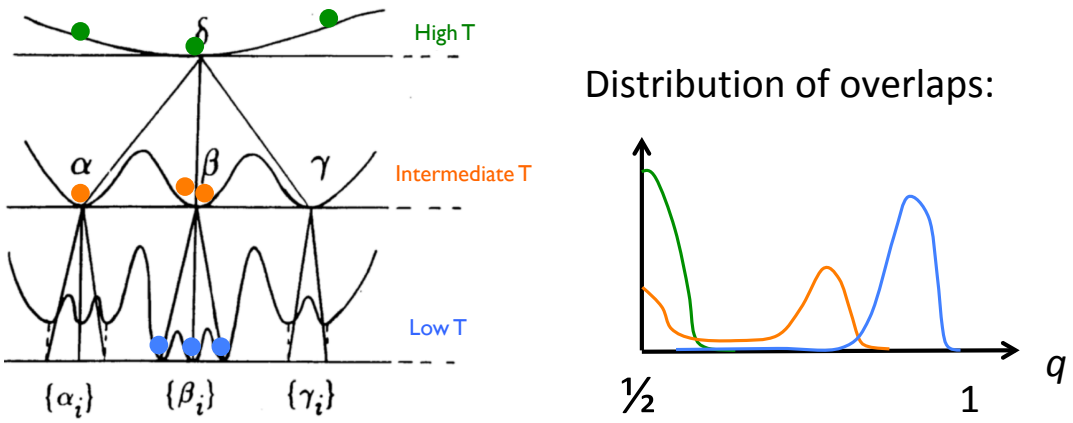

FIG. 2. Left panel: schematic changes in the free energy landscapes as the temperature is lowered. Right panel: distribution of overlaps q in each one of the left regime. At high temperatures, the distribution is peaked about $q = 0$ as replicas are statistically uncorrelated. At intermediate temperatures, the distribution is bimodal, indicating a breaking of the replica symmetry. This implies that the overlap between the two states can be quite different depending on where they are in the free energy landscape. At low temperatures, the distribution of overlaps is peaked at a value of q close to one. This implies that at low temperatures, all of the replicas tend to lie near the ground state.

However, in a spin glass, we do not know what the ground state looks like so we cannot define an order parameter through this standard procedure. Instead, we define $q(S,S')$ as a measure of the similarity between two different spin configurations as

$$q(S,S') = \frac{1}{N}\sum_{i=1}^{N} s_i s_i' \tag{15}$$

The thermal expectation of q is given by

$$\langle q(J)\rangle = \sum_{SS'} q(S,S')\,p(S|J)\,p(S'|J) = \sum_{SS'} q(S,S')\,\frac{e^{\beta E[S,J]}}{Z[\beta,J]}\,\frac{e^{\beta E[S',J]}}{Z[\beta,J]} \tag{16}$$

We introduce $n-2$ replicas to express the denominator in a tractable way,

$$\frac{1}{Z[J]^2} = \lim_{n\to 0}\sum_{S_3}\sum_{S_4}\cdots\sum_{S_n} e^{-\sum_{a=3}^{n} E[S_a,J]}\,, \tag{17}$$

and obtain the following expression for the mean overlap,

$$\overline{\langle q\rangle} = \lim_{n\to 0}\sum_{S_1}\cdots\sum_{S_n} q(S_1,S_2)\; e^{-E_{eff}[S_1,S_2,\ldots,S_n]}\,, \tag{18}$$

where the effective energy E_{eff} was defined in Eq. (13). As a conclusion, the order parameter in the replica methods is a measure of the similarity between those replicas resulting from how they interact.

The spherical spin glass model (without replicas)

We will illustrate the replica method on a specific model of spins, called spherical spin glass. This model can be solved without replicas, with basic knowledge in random matrix theory. It is therefore a very good testground to understand how replicas work.

In the spherical spin glass model the spin variables x_i, where $i = 1,\ldots,N$, are real valued. The measure over the N-dimensional spin configuration $\boldsymbol{x} = (x_1,\ldots,x_N)$ is defined as

$$\rho(\boldsymbol{x}|W) = \frac{1}{Z(W)}\,e^{\frac{1}{2}\sum_{ij} x_i W_{ij} x_j}\,\delta\left(\sum_i (x_i)^2 - N\right) \tag{19}$$

where W is an $N \times N$ symmetric random matrix from the Gaussian orthogonal ensemble, i.e.

$$W_{ii} \sim \mathcal{N}\left(0, \frac{2\sigma^2}{N}\right) \quad , \qquad W_{ij} \sim \mathcal{N}\left(0, \frac{\sigma^2}{N}\right) \tag{20}$$

where σ fixes the scale of W and plays the role of inverse temperature β. We consider that the interaction matrix W is quenched, *i.e.* the individual components of the matrix are all fixed. This model can be seen as a relaxed version of the standard SK model, in which $x_i = \pm 1$ and the Hamiltonian is given by

$$H = -\sum_{i<j} x_i W_{ij} x_j \tag{21}$$

The δ-function in Eq. (19) ensures that $\boldsymbol{x}$ lives on the N-dimensional hypersphere of radius $\sqrt{N}$. We start by relaxing this constraint by replacing the δ-function with a soft constraint

$$\sum_i \langle x_i^2 \rangle = N \tag{22}$$

implying that the spherical constraint is is satisfied on average. Later on, we will reintroduce the original, hard constraint on the norm of $\boldsymbol{x}$.

Enforcing the constraint in Eq. (22) with the help of a Lagrange multiplier μ, we get

$$\rho(\boldsymbol{x}|W) \propto e^{\frac{1}{2}\sum_{ij} x_i (\mu \mathbf{1} - W)_{ij} x_j} \, , \tag{23}$$

which is simply a multivariate Gaussian distribution with covariance matrix elements

$$C_{ij} = \langle x_i x_j \rangle = (\mu - W)^{-1}_{ij} \, . \tag{24}$$

Here μ is chosen so that $\mathrm{Tr}\,(C) = N$. We can expand this in the eigenbasis of W, i.e.

$$W \boldsymbol{e}_a = \lambda_a \boldsymbol{e}_a, \; \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \tag{25}$$

so that $\tilde{x}_a = \boldsymbol{x} \cdot \boldsymbol{e}_a$ where $\langle \tilde{x}_a \rangle = 0$ and

$$\langle \tilde{x}_a \tilde{x}_b \rangle = \frac{\delta_{ab}}{\mu - \lambda_a} \, . \tag{26}$$

The soft constraint then becomes

$$\sum_i \langle x_i^2 \rangle = N = \sum_a \langle \tilde{x}_a^2 \rangle = \sum_a \frac{1}{\mu - \lambda_a} \, . \tag{27}$$

We now define the function

$$F(\mu) = \frac{1}{N} \sum_a \frac{1}{\mu - \lambda_a} \tag{28}$$

It is easy to convince oneself that the equation $F(\mu) = 1$ has a unique root $\mu > \lambda_1$; this root is the value of the Lagrange multplier corresponding to our soft constraint on $\boldsymbol{x}$.

The condensation phase transition of the spherical model

Let us better characterize the value of μ. We define the density of eigenvalues $\rho(\lambda)$ of W as

$$\rho(\lambda) = \frac{1}{N} \sum_a \delta(\lambda - \lambda_a) \, . \tag{29}$$

In the limit $N \to \infty$, ρ is given by Wigner's semicircle,

$$\rho(\lambda) = \frac{1}{2\pi\sigma^2} \sqrt{4\sigma^2 - \lambda^2} \, . \tag{30}$$

$\rho(\lambda)$ has support in the interval $[-2\sigma, 2\sigma]$, and $\lambda_1 = 2\sigma$ and $\lambda_N = -2\sigma$. We then replace the sum in Eq. (28) with an integral, obtaining

$$1 = F(\mu) = \int_{-2\sigma}^{2\sigma} d\lambda \frac{\rho(\lambda)}{\mu - \lambda} = \frac{\mu - \sqrt{\mu^2 - 4\sigma^2}}{2\sigma^2} \,. \tag{31}$$

The root of this equation is given by

$$\mu = 1 + \sigma^2 \,. \tag{32}$$

This solution fulfills the condition $\mu > 2\sigma$ implicitly imposed by Eq. (31) as long as $\sigma < 1$. When σ approaches 1 both μ and λ_1 approach the same value 2. In reality, there must be a small gap between μ and λ_1.

For $\sigma > 1$, we therefore look for a solution where $\mu \approx \lambda_1 = 2\sigma$ plus a small gap. As in the case of Bose-Einstein condensation, we can determine this gap by considering the contribution of the term for λ_1 in a separate way:

$$F(\mu) = \frac{1}{N}\frac{1}{\mu - \lambda_1} + \frac{1}{N}\sum_{a \geq 2} \frac{1}{\mu - \lambda_a} \approx \frac{1}{N}\frac{1}{\mu - \lambda_1} + \int_{-2\sigma}^{2\sigma} d\lambda \frac{\rho(\lambda)}{\mu - \lambda} \tag{33}$$

Evaluating Eq. (33), we get the following expression for the gap[32]:

$$\mu - \lambda_1 \approx \frac{1}{N}\frac{\sigma}{\sigma - 1} \,. \tag{34}$$

The calculation above is correct since there is no need to pull out more isolated contributions from the integral. Consider for instance the contribution of the second eigenmode,

$$\langle \tilde{x}_2^2 \rangle = \frac{1}{\mu - \lambda_2} = \frac{1}{(\mu - \lambda_1) + (\lambda_1 - \lambda_2)} \,. \tag{35}$$

To estimate the difference $\lambda_1 - \lambda_2$ we use Wigner semicircle law, and integrate it from λ_2 to λ_1. We expect this integral to be asymptotically equal to $\frac{1}{N}$, as there are N eigenvalues in W. We obtain

$$\frac{1}{N} \sim \int_{\lambda_2}^{\lambda_1} d\epsilon \sqrt{\epsilon} \sim (\lambda_1 - \lambda_2)^{\frac{3}{2}} \,. \tag{36}$$

Therefore, $\lambda_1 - \lambda_2 \sim N^{-\frac{2}{3}} \gg \mu - \lambda_1 \sim N^{-1}$. We deduce that the contribution of Eq. (35) to $F(\mu)$ is negligible when $N \to \infty$.

What is the interpretation of the transition phenomenon taking place in $\sigma = 1$? Let us consider the spread of the configurations in the $\boldsymbol{x}$ space. For $\sigma < 1$, the variance of the first component

$$\langle \tilde{x}_1^2 \rangle = \frac{1}{\mu - \lambda_1} = \frac{1}{(\sigma - 1)^2} \qquad (\sigma < 1) \tag{37}$$

is finite, and so are the smaller variances of all other components $\tilde{x}_a$, with $a > 2$. For $\sigma > 1$, the calculation above shows that

$$\langle \tilde{x}_1^2 \rangle = \frac{1}{\mu - \lambda_1} = N\left(1 - \frac{1}{\sigma}\right) \qquad (\sigma > 1) \,. \tag{38}$$

Hence, the spread of the configurations is very large and diverges as $\sqrt{N}$ along the $\boldsymbol{e}_1$ direction. This phenomenon can be seen as a condensation of configurations along $\boldsymbol{e}_1$. The elongation along the second direction, $\boldsymbol{e}_2$, is of the order of $N^{\frac{1}{3}}$ according to Eq. (36), while the spreads along most directions a remain finite.

II. LECTURE 2. SOLUTION OF THE SPHERICAL SPIN GLASS MODEL WITH REPLICAS.

In this second lecture, we show how the condensation transition taking place in the spherical spin glass model can be found back with the replica method. Our aim is to illustrate the nature of the order parameters of replicas, which quantify the similarity between different solutions that one 'throws' in the same landscape we have characterized at the end of the first lecture. We also showcase the power of the replica approach in extracting additional detailed statistical information about the full probability distribution of the model by manipulating the number of replicas.

Setup of problem for the replica solution

As in the previous lecture, we consider the spherical spin glass model with N spins, a relaxed version of spin glasses. Quenched interactions form a symmetric matrix W sampled from the Gaussian Orthogonal Ensemble (GOE), with off-diagonal elements with variance σ^2/N. From the solution of this model in the previous section, we know that at large N, this model exhibits a phase transition as σ is varied. The order parameter is the projection of the spin configuration vector $\mathbf{x}$ along the top eigenvector $\mathbf{e}_1$ associated with the eigenvalue λ_1. Based on results from Lecture 1, we expect a projection of the order of $\sqrt{N}$ along the top eigenvector when $\sigma > 1$, and of the order of 1 when $\sigma < 1$.

Our goal is to derive this result using the replica approach and show the additional insights this approach provides. Our starting point is the partition function of the measure ρ in Eq. (23),

$$Z(W) = \int d\boldsymbol{x}\, e^{\frac{1}{2}\sum_{ij} W_{ij} x_i x_j}\, \delta\Big(\boldsymbol{x}^2 - N\Big). \tag{39}$$

The inverse of this partition function reads

$$\frac{1}{Z(W)} = \lim_{n\to 0} Z(W)^{n-1} = \lim_{n\to 0} \int d\boldsymbol{x}^2 \dots d\boldsymbol{x}^n e^{\frac{1}{2}\sum_{ij} W_{ij} \sum_{a=2}^{n} x_i^a x_j^a} \prod_{a=2}^{n} \delta\Big((\boldsymbol{x}^a)^2 - N\Big).$$

Writing the empirical density with the partition function expressed above gives us

$$\rho(\mathbf{x}^1) = \overline{\rho(\mathbf{x}^1|W)}^W = \lim_{n\to 0} \int \prod_{a=2}^{n} d\mathbf{x}^a \prod_{a=1}^{n} \delta\Big((\mathbf{x}^a)^2 - N\Big)\, \overline{e^{\frac{1}{2}\sum_{ij} W_{ij} \sum_{a=1}^{n} x_i^a x_j^a}}^W,$$

where averaging over the disorder over W is denoted by the bar symbol.

The crucial quantity in the replica approach is the effective energy, we express it using the index $a = 1, \dots, n$ for n replicas of the system, and the averaging over the disorder over W creates an effective coupling interaction between the replicas.

$$e^{-E_{eff}(\mathbf{x}^1,\dots,\mathbf{x}^n)} = \overline{e^{\frac{1}{2}\sum_{ij} W_{i,j} \sum_{a=1}^{n} x_i^a x_j^a}}. \tag{40}$$

Using independence of entries of W, exploiting its symmetry to eliminate $1/2$ factor by looping over the upper-triangular entries only and considering diagonal entries separately

$$E_{eff}(\mathbf{x}^1, \dots, \mathbf{x}^n) = -\log\left[\prod_i \overline{e^{\frac{1}{2} W_{ii} \sum_a x_i^a x_i^a}} \times \prod_{i<j} \overline{e^{\frac{1}{2} W_{ij} \sum_a x_i^a x_j^a}}\right] \tag{41}$$

$$= -\sum_i \frac{2\sigma^2}{N} \times \frac{1}{2}\left(\frac{1}{2}\sum_a x_i^a x_i^a\right)^2 - \sum_{i<j} \frac{\sigma^2}{N} \times \frac{1}{2}\left(\frac{1}{2}\sum_a x_i^a x_j^a\right)^2 \tag{42}$$

$$= -\frac{\sigma^2}{4N}\sum_{ij}\left(\sum_a x_i^a x_j^a\right)^2 = -\frac{\sigma^2}{4N}\sum_{a,b}\left(\sum_i x_i^a x_i^b\right)^2 = -\frac{\sigma^2}{4} N \sum_{a,b} q^{ab}(x)^2 \tag{43}$$

where we compute the averages using the Gaussian measure and the $n \times n$ overlap matrix is defined by its entries

$$q^{ab}(x) = \frac{1}{N}\sum_i x_i^a x_i^b = \frac{1}{N}\mathbf{x}^a \cdot \mathbf{x}^b\ . \tag{44}$$

These overlaps are smaller than 1 (in absolute value) due to the normalization of the configurations.

Using overlaps, we can understand the trade-off between entropy and energy in the problem above. Naturally, due to entropic reasons, randomly chosen configurations are orthogonal on the high-dimensional hypersphere, which is represented by small overlap, $q \to 0$, and high value of the energu E_{eff} in Eq. (41). Conversely, large overlaps, $q \to 1$, result from large scalar products between configurations; the energetic terms 'push' vectors towards the same direction which is a trace of them obeying the same matrix W. In other words, the effective energy is lower as the replicas are very similar, i.e., have large overlaps, which correspond to low entropies.

Calculation of the entropy

In this subsection, we introduce the quantity

$$\Xi(r) = \int \prod_{a=1}^{n} d\mathbf{x}^{\mathbf{a}} \exp\Big[-\sum_{a,b} r^{ab}\, \mathbf{x}^a \cdot \mathbf{x}^b \Big] , \tag{45}$$

which is the generating function of the distribution of the overlaps. It depends on a $n \times n$ positive definite matrix r with elements r^{ab}, which act as forces on the overlaps q^{ab}. By calculating $\Xi(r)$ in two ways we will get an expression for the entropy of the system.

Way 1. Write the expression in standard form to compute a Gaussian integral explicitly in the components of $\mathbf{x}$, under the assumption that r is positive definite

$$\begin{aligned}\Xi(r) &= \int \prod_{a=1}^{n} \prod_{i} dx_i^a e^{-\sum_{a,b} r^{ab} \sum_i x_i^a \cdot x_i^b} = \left(\int dx^1 \dots dx^n e^{-\frac{1}{2}\sum_{a,b} 2r^{ab}\, \mathbf{x}^a \cdot \mathbf{x}^b} \right)^N \\ &= \left(\frac{(2\pi)^{n/2}}{\sqrt{\det(2r)}} \right)^N . \end{aligned} \tag{46}$$

Way 2. In this calculation, we introduce entropy explicitly. Recall that entropy is the logarithm of the multiplicity of the overlap matrix. Noting that Ξ scales exponentially with N and using Eq. (44) to express the scalar product in terms of overlap, we write

$$\Xi(r) = \int \prod_{a<b} dq^{ab} \int \prod_{a} d\mathbf{x}^{\mathbf{a}} \prod_{a \le b} \delta\left(q^{ab} - \frac{1}{N} \mathbf{x}^a \cdot \mathbf{x}^b \right) e^{-N \sum_{a,b} r^{ab} q^{ab}} . \tag{47}$$

The following quantity is the weight associated to a given matrix q^{ab} for all elements a, b. We expect it to scale exponentially with N as we are working in an $N-$dimensional space. We can thus write a relation involving the entropy $S(\{q^{ab}\})$ interpreted as the log-multiplicity of the overlap matrix as

$$\int \prod_{a} d\mathbf{x}^{\mathbf{a}} \prod_{a \le b} \delta\left(q^{ab} - \frac{1}{N} \mathbf{x}^a \cdot \mathbf{x}^b \right) = e^{NS(\{q^{ab}\}) + o(N)} \tag{48}$$

As N becomes large, using the Laplace (saddle-point) method, we obtain

$$\frac{1}{N} \log \Xi(r) = \max_{Q=\{q^{ab}\}} \Big[S(Q) - \mathrm{Tr}\,(r \cdot Q) \Big] = -\frac{1}{2} \log\det(2r), \tag{49}$$

where we have used Eq. (46) and neglected r-independent additive constant.

At this stage, we give an interpretation of the matrix r. It is difficult to compute the entropy explicitly, but one can compute its Legendre transform with relative ease — it is a common trick in statistical mechanics to 'jump' from one ensemble to the other. In other words, working at fixed q^{ab} is difficult, but imposing a fixed pressure r^{ab} associated with each q^{ab} will 'force' the overlaps to have a definite value. The expression in Eq. (49) will be correct for all choices of r; to get the entropy we match it with the result of the first derivation Eq. (46) which can be done for all r. Then, we perform the inverse Legendre transform and recover the entropy $S(\{q^{ab}\})$. Applying the inverse Legendre transform on Eq. (49), that is, calculating the partial derivative w.r.t. r on both sides, and using the fact that we are working with a symmetric matrix we get

$$Q = (2r)^{-1} . \tag{50}$$

Plugging this back into Eq. (49), we have

$$S(\{q^{ab}\}) = -\frac{1}{2} \log\det(Q^{-1}) + \underbrace{\mathrm{Tr}\,(r \cdot (2r)^{-1})}_{\text{const.}} + c_1 = \frac{1}{2} \log\det(Q) . \tag{51}$$

Note that the result of this calculation is could have been guessed based on the Maximum Entropy principle: the distribution with maximal entropy at fixed second moments is Gaussian, and the entropy of a multivariate Gaussian is $\frac{1}{2}\log\det(\Sigma)$ where Σ is the covariance matrix.

Replica symmetric ansatz

Let us go back to the computation of the normalization of the density of $\mathbf{x}^1$. We write

$$\overline{Z(W)^n} = \int d\mathbf{x}' \rho(\mathbf{x}') = \int \prod_a d\mathbf{x}^a \prod_a \delta\Big((\mathbf{x}^a)^2 - N\Big) \exp\left(\frac{\sigma^2}{4} N \sum_{a,b} q^{ab}(x)^2\right). \tag{52}$$

We now replace the integral over the configurations with the integral over the overlap matrix Q, taking into account the multiplicities of the sets of n configurations associated to a given Q. The hard constraint in the delta function imposes $q^{aa}(x) = 1, \forall a$. Thus we write

$$\overline{Z(W)^n} = \int \prod_{a<b} dq^{ab} \exp\Big(N\big[S(Q) + \frac{\sigma^2}{4}\mathrm{Tr}\, Q^2\big]\Big) = \int \prod_{a<b} dq^{ab} e^{\frac{N}{2}V(Q)}, \tag{53}$$

where

$$V(Q) = \log\det(Q) + \frac{\sigma^2}{2}\mathrm{Tr}\,(Q^2) \tag{54}$$

We want to find the maximum of $V(Q)$:

$$\frac{\partial V}{\partial q^{ab}} = (q^{-1})^{ab} + \sigma^2 q^{ab} = 0 \quad \text{for} \quad a < b. \tag{55}$$

At this point, we need to find an Ansatz that will allow us to do the $n \to 0$ continuation. The obvious thing to say is that all the replica indices $a, b = 1, \ldots, n$ are generic indices, and the quantities that we compute should be invariant under all possible permutations of replicas. We assume that this invariance, in other words, symmetry of the function we are optimizing is actually reflected in the solution at the maximum. This is what we call the *replica symmetric Ansatz*, that is, all the off-diagonal entries of the overlap matrix are equal $q^{ab} = q$ for all $a \neq b$ and $q^{aa} = 1$ for all a. We express the inverse of the overlap matrix Q as

$$Q^{-1} = \begin{pmatrix} 1 & & q \\ & \ddots & \\ q & & 1 \end{pmatrix}^{-1} = \begin{pmatrix} A & & B \\ & \ddots & \\ B & & A \end{pmatrix}, \quad \text{where } \begin{cases} A = \frac{1+(n-2)q}{1+(n-2)q-(n-1)q^2} \\ B = \frac{-q}{1+(n-2)q-(n-1)q^2}. \end{cases}$$

Therefore, we rewrite Eq. (55) with the replica-symmetric Ansatz as

$$\frac{-q}{(1-q)(1+(n-1)q)} + \sigma^2 q = 0 \tag{56}$$

and solve for q to obtain the solutions $q = 0$ or $q = 1 - \frac{1}{\sigma}$ as $n \to 0$. Which solution should we choose? Recall that our goal is to maximise the potential function V — plugging the replica symmetric (RS) overlap matrix into Eq. (54) gives us

$$V(q, n) = \log(1 + (n-1)q) + (n-1)\log(1-q) + \frac{\sigma^2}{2} n(1 + (n-1)q^2)\,. \tag{57}$$

Notice that the potential $V(q, 0)$ vanishes, which is desirable because $\overline{Z(W)^n} = \int d\mathbf{x}' \rho(\mathbf{x}')$ expressed in Eq. (53) is equal 1 as we take $n \to 0$. For small n, we have $V(q, n) = n v(q) + o(n^2)$ where

$$v(q) = \log(1-q) + \frac{q}{1-q} + \frac{\sigma^2}{2}(1 - q^2). \tag{58}$$

We now look for maximum values of $v(q)$ by plugging in the solutions of $v'(q) = 0$ which are consistent with the saddle solutions:

$$\begin{cases} v(0) = \frac{\sigma^2}{2} \\ v\big(1 - \frac{1}{\sigma}\big) = 2\sigma - \frac{3}{2} + \log\big(\frac{1}{\sigma}\big) \end{cases} \tag{59}$$

Fig. 3 shows the plot of the two candidates as functions of σ; they intersect in $\sigma = 1$.

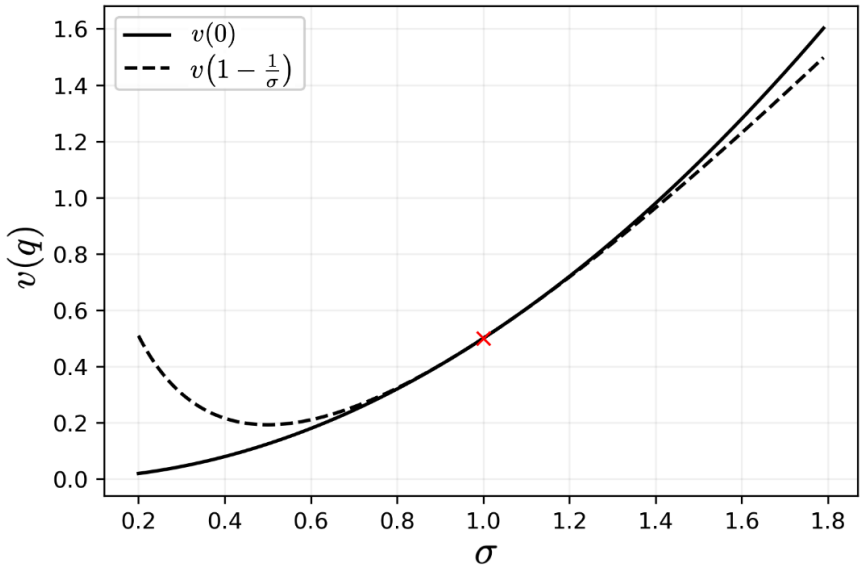

FIG. 3. Plot of the saddle point solutions for the overlap q, see Eq. (61), vs. σ. The red cross shows their point of intersection in $\sigma = 1$.

Naively, it would seem that if we want to maximize v, we should choose the solution $q = 1 - \frac{1}{\sigma}$ for $\sigma < 1$ and the solution $q = 0$ for $\sigma > 1$ — this is obviously wrong. In reality, the number of replicas n is determining whether we should be minimizing or maximizing. This can be best seen by expanding the potential in the vicinity of $n = 1$ replicas:

$$V(q,n) = \frac{\sigma^2}{2} + (n-1)\left(q + \log(1-q) + \frac{\sigma^2}{2}(1+q^2)\right) + O((n-1)^2) \tag{60}$$

For $n > 1$, looking at the first order term in $(n-1)$ above,we find that maximizing V gives a non-trivial overlap at high σ, as there is a clear maximum $q > 0$ for any $\sigma > 1$. For any $\sigma < 1$ the potential is maximal for $q = 0$. This phenomenology holds for all $n > 1$, see the plot of the potential $V(q, n = 3)$ in Figure 4.

Crucially, as soon as $n < 1$, if we want to track the $q \neq 0$ solution, we have to reverse the prescription of maximizing Eq. (55) into minimization. The reason for this is that there are $\frac{1}{2}n(n-1)$ degrees of freedom to minimize/maximize over, and this number changes signs when n lies between 0 and 1. This is one of the oddities of the replica approach. This phenomenon can also be seen in the Hessian matrix, which we will comment on at a later stage. As a result, minimizing $V(q,n)$ for $0 < n < 1$, we recover the result found in the first lecture

$$\begin{cases} q = 0 & \text{for} \quad \sigma < 1, \\ q = 1 - \frac{1}{\sigma} & \text{for} \quad \sigma > 1. \end{cases} \tag{61}$$

This result nicely agrees with the outcome of the previous lecture:

a. Case $\sigma < 1$. We know from Lecture 1 that vectors $\mathbf{x}$ sampled from the spherical spin model are such that $\mathbf{x} = \tilde{x}_1\mathbf{e}_1 + \tilde{x}_2\mathbf{e}_2 + \cdots + \tilde{x}_a\mathbf{e}_a + \cdots + \tilde{x}_N\mathbf{e}_N$ with $\tilde{x}_a \sim \mathcal{N}(0, 1/(\mu - \lambda_a))$. The overlap between two independently drawn vectors is

$$q = \frac{1}{N}\langle \mathbf{x}\cdot\mathbf{x}'\rangle = \frac{1}{N}\sum_a \langle \tilde{x}_a \tilde{x}'_a\rangle = 0\,. \tag{62}$$

b. Case $\sigma > 1$. A typical configuration will be $\mathbf{x} = \sqrt{N\left(1-\frac{1}{\sigma}\right)}\mathbf{e_1} + \cdots$ where we omit further orthogonal terms with Gaussian projections which are of order one and do not have a significant contribution. Sampling another vector $\mathbf{x}'$ independently of the first one, we obtain

$$q = \frac{1}{N}\langle \mathbf{x}\cdot\mathbf{x}'\rangle = \frac{1}{N}\sqrt{N\left(1-\frac{1}{\sigma}\right)}\sqrt{N\left(1-\frac{1}{\sigma}\right)} = 1 - \frac{1}{\sigma}\,. \tag{63}$$

In other words, a typical configuration lie on the cone of axis $\mathbf{e}_1$ and angle $\theta = \cos^{-1}(\sqrt{q})$, see Eq. (38). Two vectors lying on the cone have therefore typically an overlap $(\sqrt{q})^2$, in agreement with Eq. (61).

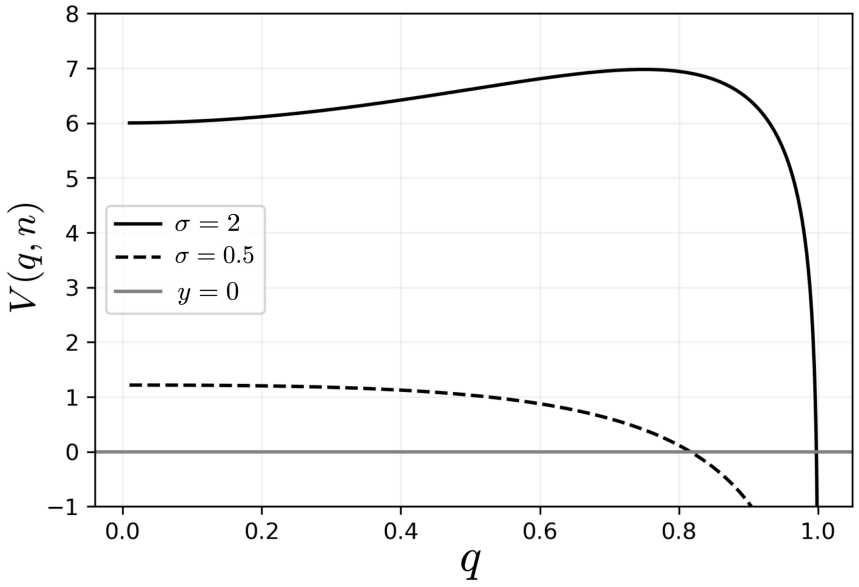

FIG. 4. Plot of the potential $V(q,n)$ as expressed in Eq. (60) for $n = 3$. The potential has a clear maximum for $\sigma = 2$, this is true for any $\sigma > 1$, while there is no clear maximum for $\sigma = 0.5$.

Two Remarks

First, looking for the saddle-point of the replica potential, as done above, is generally not sufficient. To better understand whether the solution obtained with the replica-symmetric Ansatz is correct, one should always check the stability of the solution. More specifically, we should check whether the eigenvalues of the Hessian matrix

$$H_{ab,cd} = \frac{\partial^2 V}{\partial q^{ab} \partial q^{cd}} \tag{64}$$

have the expected sign, depending on whether we are maximizing or minimizing.

Second, even when the solution is locally stable, there might be other valid solutions in the q^{ab} space not symmetric under permutations of replica indices. For example, in this problem, defining $s^a, s^b = \pm 1$, and redefining the entries of the overlap matrix $q^{ab} \to s^a s^b q^{ab}$ for any combination of s^a, s^b gives us a perfectly valid solution. This is because the determinant of both matrices and the trace of the square of both matrices are equal, thus giving the same potential $V(q)$. As a result, there actually are 2^n solutions. Another case is when replica symmetry is broken, i.e., the Ansatz of off-diagonal elements of the overlap matrix being equal is wrong.

What about non zero numbers of replicas?

In the last part of this lecture, we show the power of the replica method by manipulating the number n of replicas and extracting detailed statistical information about the full probability distribution of the model as a result. While n is usually sent to zero, other values are also interesting, depending on what we want to capture. To get additional information, we extend the interpretation of the calculation to n being different than zero.

Let us come back to the expression of the partition function $Z(W)$ in Eq. (39). From now on we consider the interaction matrix W is sampled from the GOE ensemble with off-diagonal entries of variance $\frac{1}{N}$, and explicitly factor out the parameter σ as an inverse temperature. As we take σ to be large, the partition function can be approximated as

$$Z(W) \overset{\sigma \text{ large}}{\simeq} \exp\left(\frac{\sigma}{2} N \lambda_1(W)\right), \tag{65}$$

where the N term come from the norm of $\mathbf{x}$ and $\lambda_1(W)$ is the top eigenvalue of W. There exists extensive work in Random Matrix Theory on the large deviations of the top eigenvalues of random matrices. It is in particular possible that the top

eigenvalue $\lambda_1(W)$ is much larger than its typical value 2. Pulling it away to the right from the bulk of the spectrum whose boundary is at 2 will have an exponential cost of the order of $e^{-N\phi_+(\lambda_1)}$ in terms of probability, where

$$\phi_+(\lambda_1) = \frac{\lambda_1}{2}\sqrt{\left(\frac{\lambda_1}{2}\right)^2 - 1} + \log\left(\frac{\lambda_1}{2} - \sqrt{\left(\frac{\lambda_1}{2}\right)^2 - 1}\right). \tag{66}$$

See [10], [28] for details on the derivation and background.

To illustrate the power of the replica approach, let us find back the result for the large deviation rate function in Eq. (66). We compute the result for the right-hand side, that is, increasing the top eigenvalue with replicas. Let us consider $\overline{Z(W)^n}$ where n is now not set equal to zero but kept positive:

$$\overline{Z(W)^n} \overset{\sigma \text{ large}}{=} \overline{\exp\left(\frac{N}{2}\sigma\lambda_1(W)n\right)} = \int_2^\infty d\lambda_1 e^{N[\frac{\sigma}{2}n\lambda_1 - \phi_+(\lambda_1)]} \simeq e^{N\max_{\lambda_1}[n\frac{\sigma}{2}\lambda_1 - \phi_+(\lambda_1)]}, \tag{67}$$

where the integral was over $\lambda_1 \geq 2$ as the positivity of n favors large values of the top eigenvalues. We will briefly evoke the $n < 0$ case later on. Recall that we know that

$$\overline{Z(W)^n} = e^{\frac{N}{2}\underset{q}{opt}V(q,n,\sigma)} \tag{68}$$

where the type of optimization (min/max) we perform over q depends on n. We see that the potential V is the Legendre transform of the large deviation function ϕ_+ of the top eigenvalue. We thus have to carry out the inverse Legendre to get back to $\phi(\lambda_1)$ from the knowledge of V. This is the last calculation of this lecture.

We choose the parameter for the Legendre transform to be $\mu = n\frac{\sigma}{2}$, thus the number of replicas we have is $n = \frac{2\mu}{\sigma}$. The number of replicas will be small but nonzero because typically μ is of order one and we are working in the large σ regime. The Legendre relation cannot be written as

$$\frac{1}{2}V\Big(q, n = \frac{2\mu}{\sigma}, \sigma\Big) = \max_{\lambda_1}[\mu\lambda_1 - \phi_+(\lambda_1)] \tag{69}$$

The saddle point equation we have to solve for q, now with $n = \frac{2\mu}{\sigma}$, is

$$\frac{1}{\sigma^2} = (1-q)(1-q+nq)\,, \tag{70}$$

according to Eq. (56). We look for a solution of the form $q = 1 - \frac{\Delta(\mu)}{\sigma}$ where Δ is a nontrivial function of μ. Note for $\mu = 0$ we have $\Delta(\mu) = 1$, in agreement with the solution for large σ we discussed previously in Eq. (61) as we sent n to zero. Plugging the ansatz $q = 1 - \frac{\Delta(\mu)}{\sigma}$ into the saddle–point equation Eq. (70) and considering the largest order in $1/\sigma$, we have

$$\frac{1}{\sigma^2} = \frac{\Delta(\mu)}{\sigma}\left(\frac{\Delta + 2\mu}{\sigma}\right) + \cdots \tag{71}$$

which gives us

$$\Delta(\mu) = \sqrt{1+\mu^2} - \mu\,.$$

We now know the overlap q as a function of the number of replicas, and puts it back into the potential:

$$\begin{aligned}
&\frac{1}{2}V\Big(q = 1 - \frac{\sqrt{1+\mu^2}-\mu}{\sigma}, n = \frac{2\mu}{\sigma}, \sigma\Big) \overset{\sigma \text{ large}}{=} \\
&= \frac{1}{2}\log\Big(\frac{\Delta+2\mu}{\sigma}\Big) + \frac{1}{2}\Big(\frac{\mu}{\sigma} - 1\Big)\log\Big(\frac{\Delta}{\sigma}\Big) + \frac{1}{2}\frac{\sigma^2}{2}\frac{2\mu}{\sigma}\Big(2\frac{\Delta}{\sigma} + \frac{2\mu}{\sigma}\Big) \\
&\overset{\sigma\to\infty}{=} \log(\sqrt{1+\mu^2} + \mu) + \mu\sqrt{1+\mu^2}\,.
\end{aligned} \tag{72}$$

This is a suitable expression for the LHS of Eq. (69) to perform an inverse Legendre transform which will give us an explicit expression for $\phi(\lambda_1)$. We have

$$\max_{\lambda_1}[\mu\lambda_1 - \phi_+(\lambda_1)] = \log(\sqrt{1+\mu^2} + \mu) + \mu\sqrt{1+\mu^2} \tag{73}$$

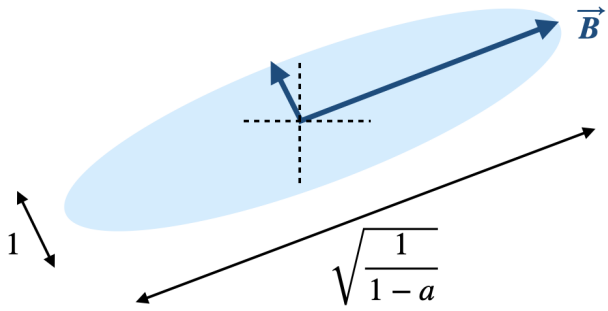

FIG. 5. Schematic representation of the influence of the biasing direction $\boldsymbol{B}$. The component of a vector along this direction is stretched by a factor $\sqrt{\frac{1}{1-a}}$ while orthogonal directions are left unchanged.

$$\overset{\frac{\partial}{\partial \mu}}{\Longrightarrow} \quad \lambda_1 = 2\sqrt{1+\mu^2} \quad \text{i.e.} \quad \mu = \sqrt{\left(\frac{\lambda_1}{2}\right)^2 - 1} \tag{74}$$

As expected, we see that a positive number of replicas corresponds to $\lambda_1 > 2$. Substituting μ with its expression in terms of λ_1 in Eq. (73) we find back the large deviation rate function ϕ_+ in Eq. (66).

A natural question at this point is: what happens if we consider a non-zero but negative number of replicas? In principle, one should now be able to explore the large devations of the distribution of eigenvalues corresponding to $\lambda_1 < 2$. However Eq. (74) shows that the calculation above makes sense for $\mu > 0$ only ... The reason of this apparent paradox is that there is a fundamental asymmetry between the left and right tails in the distribution of λ_1. To evaluate the probabilistic cost to make the top eigenvalue $\lambda_1(W)$ smaller than its average value 2, one must push all the eigenvalues in the bulk to the left. This is more unlikely and has a far greater cost than the previous case, when we considered having $\lambda_1(W) > 2$. For instance, this may be done by shrinking all W's, which is of exponential cost in N^2. Large deviations of the top eigenvalue λ_1 below its typical value have indeed very low probabilities $e^{-N^2\phi_-(\lambda_1)}$ decaying exponentially with N^2 and not N. If we compare this scaling with the situation on the right side in Eq. (67) it is clear that these large deviations can be captured in the replica framework if the number of replicas is made both negative and of the order of N, *i.e.* $n = N\,\frac{2\mu}{\sigma}$ with negative μ of the order of the unity. Carrying out replica calculations in this regime is very difficult as fluctuations around the saddle-point cannot be neglected any longer, and remains largely an open problem, see however [35].

III. LECTURE 3. LEARNING ONE OUT OF MANY DIRECTIONS.

Unsupervised learning

Let us suppose we have a N-dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_N)$ taken from distribution biased towards one direction denoted by $\boldsymbol{B} = (B_1, \ldots, B_N)$:

$$\rho\left(\boldsymbol{x}|\boldsymbol{B}\right) \propto \exp\left\{-\frac{1}{2}\boldsymbol{x}^2 + \phi(\boldsymbol{x}\cdot\boldsymbol{B})\right\}, \tag{75}$$

where ϕ is a generic function and $\boldsymbol{B}$ is a fixed vector of norm one: $\|\boldsymbol{B}\| = \sqrt{\sum_{i=1}^N B_i^2} = 1$.

If ϕ is a quadratic function, $\phi(u) = a/2\,u^2$, then ρ is the distribution of a Gaussian multivariate random variable with covariance $\langle x_i x_j\rangle = \left[\left(\mathbf{I} - a\boldsymbol{B}\boldsymbol{B}^\top\right)^{-1}\right]_{ij} = \delta_{ij} + \frac{a}{1-a}B_iB_j$. Note that this only makes sense if $a < 1$ since one needs the covariance matrix to be (semi)-definite positive.

If one has an enormous number of points $P \gg N$ drawn from this distribution ρ, one should be able to retrieve the direction $\boldsymbol{B}$ through principal component analysis (PCA). The regime where the number of points is in the same order of magnitude as their dimensions $P = \mathcal{O}(N)$ has received a lot of attention in statistics and in the case of a quadratic function ϕ in random matrix theory, since such model is known as *spiked Wishart matrix*.

Here we are interested in generic function ϕ, for example:

- $\phi(u) := a/2\,u^2 + b\,u$, where the linear term favors configurations of $\boldsymbol{x}$ with a *positive* dot product with the vector $\boldsymbol{B}$ (if $b > 0$).

- $\phi(u) := \cdots + c|u| + du^3, \ldots$

For this generic setting, one would like to infer the unknown vector $\boldsymbol{B}$ from P samples of $\boldsymbol{x}$:

$$\boldsymbol{B} \to \{\boldsymbol{x}^\mu\}_{\mu=1,\ldots,P} \to \boldsymbol{B}' \stackrel{?}{=} \boldsymbol{B} \tag{76}$$

Let us consider the *posterior distribution* of the direction:

$$P_{\text{posterior}}\left(\boldsymbol{B}'|\{\boldsymbol{x}^\mu\}\right) \propto \delta(\boldsymbol{B}'^2 - 1) \mathrm{e}^{\sum_{\mu=1}^P \Phi(\boldsymbol{B}'\cdot\boldsymbol{x}^\mu)} \tag{77}$$

In the case of quadratic $\phi(u) = a/2\, u^2$, one retrieves the spherical spin glass model studied in the previous lectures:

$$\sum_{\mu=1}^P \phi(\boldsymbol{B}'\cdot\boldsymbol{x}) = \frac{a}{2}\sum_{i,j} \tilde{B}'_i W_{ij} \tilde{B}'_j \tag{78}$$

where $W_{ij} = \sum_{\mu=1}^P x_i^\mu x_j^\mu$ is the entry of a Wishart matrix. There are two important parameters here: the constant a measuring the strength of the bias towards $\boldsymbol{B}$ and the ratio P/N. Depending on these two parameters, there will be a phase diagram with a region where it is possible to infer the vector $\boldsymbol{B}$ from the one where it is not.

One can repeat the replica calculation done in Lecture 2, see for example [36]. Let us briefly describe the main difference with the computation of the previous lecture. When we do the average over the quenched disorder (here, the data points) to get the effective energy coupling the different replicas of the $\boldsymbol{B}'$ vector, the order parameters will be the overlaps $q^{ab} = \boldsymbol{B}'^a \cdot \boldsymbol{B}'^b$ and there appear new parameters $r^a = \boldsymbol{B}'^a \cdot \boldsymbol{B}$ encoding the overlap with the ground-truth direction.

For a large number of data point $P \gg N$, one should expect the overlap r with the true direction to be very close to its maximal value. As we decrease the ratio $\alpha = P/N$, this overlap should also decrease and two scenarios can take place. In the first one, the overlap r *smoothly* decreases as one decreases $\alpha = P/N$ and vanishes for $\alpha = 0$. In the second scenario, as one decreases α, the overlap reaches $r = 0$ at a value $\alpha_c > 0$ and remains equal to zero for $\alpha \in [0, \alpha_c]$. What pops out of the replica computation is that which one of the two scenarios holds depends on whether the quantity

$$\bar{u} := \int_{-\infty}^{\infty} \mathrm{d}u\, u\, \mathrm{e}^{-u^2/2 + \phi(u)}, \tag{79}$$

is equal to zero or not:

- if $\bar{u} = 0$, then we are in the second regime, that is there is a phase transition at $\alpha_c > 0$.
- if $\bar{u} \neq 0$, then there is no phase transition at finite α.

FIG. 6. Schematic representation of the overlap between generic data and the true direction depending on the value of the integral in Eq. (79).

Let us first consider the latter case $\bar{u} \neq 0$ since this is the simpler one. This case corresponds essentially to non-even function ϕ, *e.g.* with a linear term. This asymmetry means that the data points are naturally aligned with the vector $\boldsymbol{B}$ *and* in the same direction. Consequently, averaging over the data will point towards the direction of the vector $\boldsymbol{B}$ even if one has little data (that is for a small value of α). As more and more data point are added, that is, as α increases, the accuracy gets better, which explains the smooth increasing behavior of r.

The former case $\bar{u} = 0$ is conceptually more interesting and is referred to as *retarded learning* [46] since one needs a minimal amount of data before one can estimate the direction $\boldsymbol{B}$.

In the case of a quadratic function $\phi(u) = a/2u^2$, the value $\alpha_c(u)$ is well known from Random Matrix Theory. For $\alpha < \alpha_c$, the spectrum of the empirical covariance matrix of the data is the Marčenko-Pastur distribution and in particular, the top eigenvalue is at the edge of this distribution and is just noise: it does not contain any information about of the true direction $\boldsymbol{B}$. In the regime $\alpha > \alpha_c$, there will be one *spike* away from the bulk and the corresponding eigenvector as a non-zero overlap with $\boldsymbol{B}$ and hence can be used to estimate it.

To build intuition for generic ϕ, let us consider the following non rigorous argument, based on an entropy vs. energy competition. In high dimensions ($N \gg 1$), most of the vectors are orthogonal to a fixed direction (here, the vector $\boldsymbol{B}$). Hence, entropy favors a null overlap, $r = 0$. Subsequently, one should expect that having a non-zero overlap r induces a cost of the form $\sim \mathrm{e}^{Ns(r)}$ in the posterior probability where s is precisely the entropy of a vector having overlap r with the direction $\boldsymbol{B}$. However, alignment along $\boldsymbol{B}$ is energetically favorable. Thus, one may expect this term to induce a gain of the form $\sim \mathrm{e}^{-Pe(r)}$ in the posterior probability , exponentially increasing with the number of data. To sum up, we have for the posterior

$$P_{\text{posterior}}(r) \propto \mathrm{e}^{N(s(r)-\alpha\, e(r))}\,, \tag{80}$$

Let us expand the entropic and energetic terms for small values of r.

- The entropy term is by nature insensitive to the sign of the overlap, that is, $s(r) = s(-r)$. As it reaches its maximum at $r = 0$, one has the following expansion for a small value of r:

$$s(r) = s_0 - s_2 r^2 + \dots\,, \tag{81}$$

  with $s_2 > 0$.

- Similarly for the energy term we have:

$$e(r) = e_0 - e_1 r - e_2 r^2 + \dots\,. \tag{82}$$

For large N, the posterior probability in Eq. (80) is dominated by the value r given by

$$r = \operatorname{argmax}\{s(r) - \alpha e(r)\} = \operatorname{argmax}\left\{s_0 - s_2 r^2 - \alpha\left(e_0 - e_1 r - e_2 r^2\right) + \dots\right\}\,. \tag{83}$$

For $\alpha = 0$, the maximum is attained at $r = 0$ as one should expect. For small $\alpha > 0$ two situations may arise depending on whether $e_1 = 0$ or not, see Figure 6:

- If $e_1 \neq 0$, we find that the overlap is equal to

$$r = \frac{e_1}{2s_2}\alpha, \tag{84}$$

  to the lowest order in α. Hence, r linearly increases with α as soon as $\alpha > 0$.

- If $e_1 = 0$, one needs to maximize $(\alpha e_2 - s_2)r^2 + \dots$. When α is small, the coefficient in front of r^2 is negative and the maximum is in $r = 0$. The critical value α_c corresponds to the case where this coefficient is exactly equal to zero,

$$\alpha_c = \frac{s_2}{e_2}\,. \tag{85}$$

Supervised learning

Let us consider a set of input-output data

$$\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1,\dots,P} \qquad \text{with} \qquad y^\mu = f(\boldsymbol{x}^\mu)\,, \tag{86}$$

where the function f is unknown. The input vector $\boldsymbol{x}_\mu$ is N-dimensional and for simplicity we assume the output vector to be binary: $y^\mu \in \{-1, 1\}$. The goal is to learn or 'guess' the unknown function f and generalize it for other input data.

Let us start with an extremely simple model with:

- *unstructured data*: The y^μ are taken independently of the $\boldsymbol{x}^\mu$ with $y^\mu = \pm 1$ with probability one half . We further assume the component x_i^μ to be independent and such that $x_i^\mu = \pm 1$ with probability one half.
- *one-layer net*: The function f to learn is a perceptron model $f(\boldsymbol{x}) := \text{sign}(\boldsymbol{J} \cdot \boldsymbol{x})$, where $\boldsymbol{J} = (J_1, \ldots, J_N)$ and hence the goal is to learn the components $\boldsymbol{J}$.

This null model with unstructured data is by construction very far from a realistic learning model. Yet, it turns out to be an interesting model to understand the fundamental behavior of more realistic models as we will see later on.

For this linear classifier, we have for $\mu = 1, \ldots, P$

$$y^\mu = \text{sign}(\boldsymbol{J} \cdot \boldsymbol{x}^\mu) \,, \tag{87}$$

Because the output is in $\{-1, 1\}$, we always have $(y^\mu)^2 = 1$ and so by multiplying Eq. (87) by y^μ it follows that

$$1 = \text{sign}(\boldsymbol{J} \cdot \underbrace{(y^\mu \boldsymbol{x}^\mu)}_{=:\boldsymbol{\xi}^\mu}) \qquad \text{for } \mu = 1, \ldots, P \,. \tag{88}$$

One therefore gets the following set of inequalities for the vector $\boldsymbol{J}$:

$$\boldsymbol{J} \cdot \boldsymbol{\xi}^\mu > 0 \qquad \text{for } \mu = 1, \ldots, P \,. \tag{89}$$

Now since this inequality is very sensitive to a small variation of the input data, one usually prefers a more robust condition of the form

$$\boldsymbol{J} \cdot \boldsymbol{\xi}^\mu > \kappa_0 \qquad \text{for } \mu = 1, \ldots, P \,, \tag{90}$$

where κ_0 is a positive constant. For this inequality to be meaningful, one needs also to impose the norm of $\boldsymbol{J}$. For this reason, we restrict $\boldsymbol{J}$ to lie on the sphere of radius $\|\boldsymbol{J}\| = \sqrt{N}$. In order to have a non-trivial large–N behavior we also scale $\kappa_0 = \sqrt{N}\kappa$. We thus look for a vector $\boldsymbol{J}$ such that

$$\|\boldsymbol{J}\|^2 = N \qquad \text{and} \qquad \boldsymbol{J} \cdot \boldsymbol{\xi}^\mu > \kappa\sqrt{N} \qquad \text{for } \mu = 1, \ldots, P \,. \tag{91}$$

The space of solutions to this problem is convex (since we impose $\kappa > 0$), see Figure 7. As we increase P, we add more and more constraints and this shrinks the domain of possible solutions.

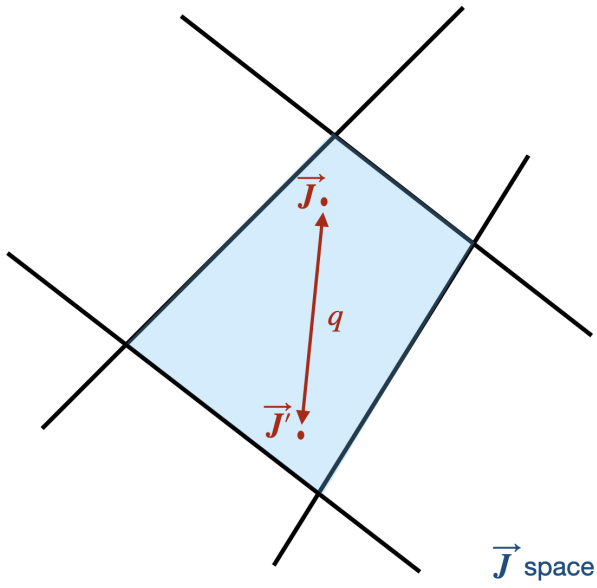

FIG. 7. Solution space of vectors $\boldsymbol{J}$. It is a convex set (for $\kappa > 0$) since it is the intersection of half spheres.

Let us consider two solutions $\boldsymbol{J}, \boldsymbol{J}'$ of this problem and estimate the similarity or overlap between them: $q = \boldsymbol{J} \cdot \boldsymbol{J}'/\sqrt{N}$. If $q \to 1$ the domain of solutions is shrinking, while q going to zero indicates an extremely large domain. Thus, the overlap

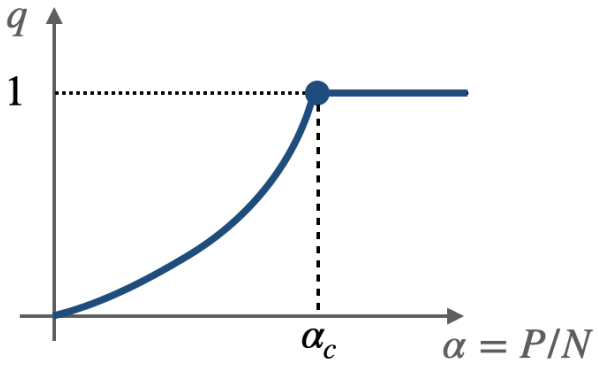

FIG. 8. Behavior of the overlap q between solutions in Figure 7 vs. ratio $\alpha = P/N$. The critical capacity α_c is the value of α for which $q = 1$.

is an indicator of the volume of the domain of solutions. Because adding more constraints reduces the set of possible solutions, we naturally expect q to be an increasing function of the ratio $\alpha = P/N$. The value α_c at which $q = 1$ is the *critical capacity* of the model.

Let us now compute the volume $V(\{\boldsymbol{\xi}^\mu\})$ of solutions of this problem:

$$V(\{\boldsymbol{\xi}^\mu\}) = \int \mathrm{d}\boldsymbol{J}\, \delta(\boldsymbol{J}^2 - N) \prod_{\mu=1}^{P} \Theta\left(\frac{\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}} - \kappa\right), \tag{92}$$

where $\Theta(\cdot)$ is the Heaviside function ($\Theta(x) = 1$ if $x \geq 0$ and zero otherwise). We now compute this volume thanks to the replica method, where the thermalized variables are the $\boldsymbol{J}$'s and the quenched disorder are the $\boldsymbol{\xi}^\mu$'s. If we denote as usual $\overline{\cdot}$ the average over the quenched disorder, we have:

$$\overline{V(\{\boldsymbol{\xi}^\mu\})^n} = \int \prod_{a=1}^{n} \mathrm{d}\boldsymbol{J}^a \prod_{a=1}^{n} \delta((\boldsymbol{J}^a)^2 - N) \prod_{\mu=1}^{P} \overline{\prod_{a=1}^{n} \Theta\left(\frac{\boldsymbol{J}^a\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}} - \kappa\right)}. \tag{93}$$

As the $\boldsymbol{\xi}^\mu$ are i.i.d., and $P = \alpha N$, we have:

$$\overline{V(\{\boldsymbol{\xi}^\mu\})^n} = \int \prod_{a=1}^{n} \mathrm{d}\boldsymbol{J}^a \prod_{a=1}^{n} \delta((\boldsymbol{J}^a)^2 - N) \left(\overline{\prod_{a=1}^{n} \Theta\left(\frac{\boldsymbol{J}^a\cdot\boldsymbol{\xi}^1}{\sqrt{N}} - \kappa\right)}\right)^{\alpha N}, \tag{94}$$

Let us introduce the variables

$$\Delta^a = \frac{\boldsymbol{J}^a\cdot\boldsymbol{\xi}^1}{\sqrt{N}} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i^a \xi_i^1 . \tag{95}$$

For large N, according to the central limit theorem, these quantities are Gaussian variables, and their distribution is completely determined by the mean and covariance:

$$\overline{\Delta^a} = 0, \tag{96}$$

$$\overline{\Delta^a \Delta^b} = \frac{1}{N} \sum_{i,j} J_i^a J_j^b \underbrace{\overline{\xi_i^1 \xi_j^1}}_{=\delta_{ij}} = \frac{1}{N} \sum_{i=1}^{N} J_i^a J_i^b =: q^{ab} . \tag{97}$$

As a consequence we may write the effective energy of the n replicas $\boldsymbol{J}^a$ as a function of the overlap matrix:

$$\exp\{-E_{\text{eff}}(Q)\} = \left(\overline{\prod_{a=1}^{n} \Theta(\Delta^a - \kappa)}\right) = \int_\kappa^\infty \prod_{a=1}^{n} \frac{\mathrm{d}\Delta^a}{\sqrt{2\pi}} \cdot \frac{\exp\left\{-\frac{1}{2}\sum_{a,b} \Delta^a (Q^{-1})^{ab} \Delta^b\right\}}{\sqrt{\det Q}} . \tag{98}$$

Thus, if we do the change of variable from the $\boldsymbol{J}$ to the replicas $\mathbf{q}$, we need to take into account the Jacobian of this change of variable which produces an entropic term we have already encountered in the previous lecture, see Eq. (51). All in all, we have

$$\overline{V(\{\boldsymbol{\xi}^\mu\})^n} = \int \prod_{a<b} \mathrm{d}q^{ab} \exp\left\{ -\alpha N\, E_{\text{eff}}(Q) + N \underbrace{S(Q)}_{=\frac{1}{2}\log\det \mathbf{q}} \right\}. \tag{99}$$

As before, let us assume a replica symmetry form for the matrix of overlaps:

$$Q = \begin{pmatrix} 1 & & q \\ & \ddots & \\ q & & 1 \end{pmatrix}. \tag{100}$$

To compute α_c, it is sufficient to focus on the regime $q = 1-\epsilon$ with ϵ small. In that case, the entropy is given, to the first order in ϵ, by

$$S = \frac{n}{2\epsilon} + n\,\mathcal{O}(\log\epsilon). \tag{101}$$

Similarly, for the effective coupling term, we have that the denominator in Eq. (98) is given by $\sqrt{\det Q} = 1+\frac{n}{2\epsilon}$ and for the numerator since $(Q^{-1})^{ab} = \frac{\delta^{ab}}{\epsilon} - \frac{1}{\epsilon^2} + \dots$ that is

$$\exp\left\{ -\frac{1}{2}\sum_{a,b} \Delta^a (Q^{-1})^{ab} \Delta^b \right\} = \exp\left\{ -\frac{1}{2\epsilon}\sum_{a=1}^n (\Delta^a)^2 + \frac{1}{2\epsilon^2}\left(\sum_{a=1}^n \Delta^a\right)^2 \right\}. \tag{102}$$

Next, to deal with the second term in the argument of the exponential which contains mixed product $\Delta^a\Delta^b$, let's introduce a dummy variable z to linearize this quadratic term. By Gaussian integration we have:

$$\exp\left\{ \frac{1}{2\epsilon^2}\left(\sum_{a=1}^n \Delta^a\right)^2 \right\} = \int_{-\infty}^{\infty} \frac{\mathrm{d}z}{\sqrt{2\pi}} \cdot \exp\left\{ -\frac{z^2}{2} + \frac{z}{\epsilon}\sum_{a=1}^n \Delta^a \right\}. \tag{103}$$

The effective energy now reads

$$-E_{\text{eff}}(Q) = \log \int_{-\infty}^{\infty} \frac{\mathrm{d}z}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}} \frac{\left(\int_\kappa^\infty \frac{\mathrm{d}\Delta}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\Delta^2}{2\epsilon}+\frac{z}{\epsilon}\Delta}\right)^n}{1+\frac{n\epsilon}{2}}, \tag{104}$$

$$= \log\left\{ 1 - \frac{n}{2\epsilon} + n\int_{-\infty}^{\infty} \frac{\mathrm{d}z}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}} \log\left[\int_\kappa^\infty \frac{\mathrm{d}\Delta}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\Delta^2}{2\epsilon}+\frac{z}{\epsilon}\Delta} \right] \right\}, \tag{105}$$

$$= \log\left\{ 1 + n\int_{-\infty}^{\infty} \frac{\mathrm{d}z}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}} \log\left[\int_\kappa^\infty \frac{\mathrm{d}\Delta}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(\Delta-z)^2}{2\epsilon}} \right] \right\}, \tag{106}$$

$$= -\frac{n}{2\epsilon}\int_{-\infty}^{\kappa} \frac{\mathrm{d}z}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}} \cdot (\kappa - z)^2 + \mathcal{O}(n\log\epsilon), \tag{107}$$

for small n. An interpretation of the selection of Δ in the passage from Eq. (106) to Eq. (107) is proposed in Figure 9.

Injecting the expression of the entropy given for small ϵ by Eq. (101) and the one of the effective energy given by Eq. (107) in Eq. (99), we get for the volume:

$$\overline{V(\{\boldsymbol{\xi}^\mu\})^n} = \exp\left\{ \frac{nN}{2\epsilon}\left(1 - \alpha\int_{-\infty}^{\kappa} \frac{\mathrm{d}z}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}} (\kappa - z)^2 \right) + \mathcal{O}(\log\epsilon) \right\}, \tag{108}$$

If the term in parenthesis is positive then for small ϵ, the average volume is exponentially large and exceeds the volume of the hypersphere. This means that the assumption that ϵ is very small is wrong. On the contrary, when this term is negative, this means that $\log\overline{V^n} \to -\infty$, and hence the volume of the solutions goes to zero. As a consequence, we deduce that the critical capacity of the model is given by, see Figure 10,

$$\alpha_c(\kappa) = \frac{1}{\int_{-\infty}^{\kappa} \frac{\mathrm{d}z}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}} (\kappa - z)^2}. \tag{109}$$

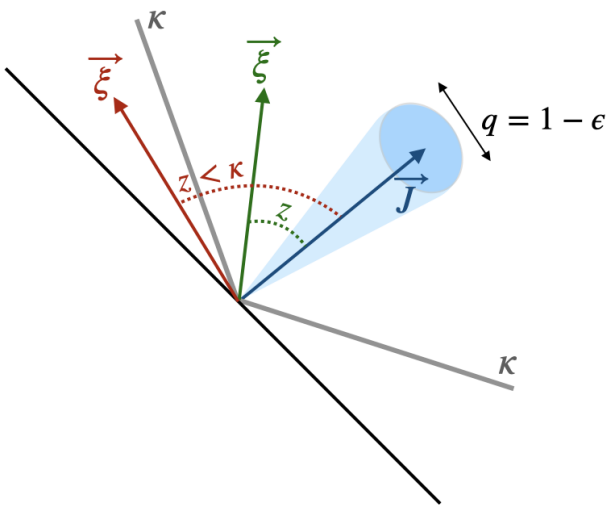

FIG. 9. Schematic interpretation of the integration over Δ in Eq. (106). In green, a 'good' situation where the overlap z between $\boldsymbol{\xi}$ and $\boldsymbol{J}$ is in the convex cone, and we have $\Delta = z$. In red, a 'bad' situation where $z < \kappa$. In this case, one needs to *push* the weight vector to force $z = \kappa$ and this introduces an important energetic cost.

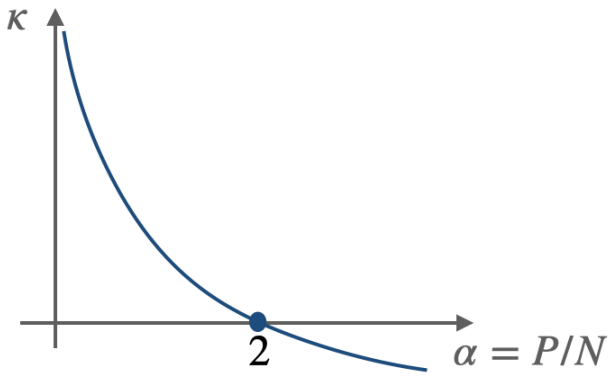

FIG. 10. Optimal stability κ as a function of $\alpha = P/N$ obtained from the RS calculation. Note that the $\kappa < 0$ part of the curve is not correct as RS is broken in this regime.

This result, first obtained by Gardner and by Gardner and Derrida [20] is a generalization of Cover's result [9] limited to the case $\kappa = 0$. After some work, one can check that the replica symmetry ansatz is locally stable as long as $\alpha < \alpha_c$ with $\kappa > 0$ by computing the eigenvalues of the associated Hessian matrix. For a negative value of $\kappa < 0$, the problem is not convex anymore and the replica Ansatz is wrong. We will see a manifestation of this result in the next lecture.

The computation above can be extended in many ways, see [16] for a comprehensive review:

- One can change the constraint on the J_i by looking for example at $J_i \in \{-1, 1\}$ or $J_i > 0$,...
- One can change the structure of the input data by considering for example components x_i, x_j to be correlated, or even two different samples $\boldsymbol{x}^\mu, \boldsymbol{x}^\nu$ to be correlated.
- One can consider recurrent neural networks rather than classifiers.
- One can extend this result to more realistic models, where the outputs are correlated with the inputs.
- One can consider more complex architectures than a single-layer perceptron, see next lecture.

IV. LECTURE 4. MULTI-LAYER NETWORKS: EXAMPLES AND DOMAINS OF SOLUTIONS

In this chapter, we will focus on a simple architecture of a multi-layer network and its domains of solutions using replica techniques. As shown in Figure 11, the network we will study has a tree-like structure with one hidden layer of binary units $\{\tau_1,\dots,\tau_k\}$ and a given set of input vectors $\boldsymbol{x}=\{x_1,\dots,x_N\}$, where each unit is determined by N/k components of $\boldsymbol{x}$. For simplicity we choose the following to fix the weights $\{W_1=1,\dots,W_k=1\}$, therefore the only free parameters, in this case, are the weights $\{\boldsymbol{J}_j\}=\{J_{ij}\}_{i\in[\![1,\dots,N]\!],j\in[\![1,\dots,k]\!]}$ for the hidden layer. This simplification is also known as the decoder representation. Finally, the output y and all hidden unit τ's of the tree are determined by

$$y=F(\tau_1,\dots,\tau_k)\quad\text{and}\quad\tau_j=\text{sign}(\boldsymbol{J}_j.\boldsymbol{x}_{|j}) \tag{110}$$

where $\boldsymbol{x}_{|j}$ represents the N/k components of $\boldsymbol{x}$ that will determine τ_j. Two examples of functions for F are the parity

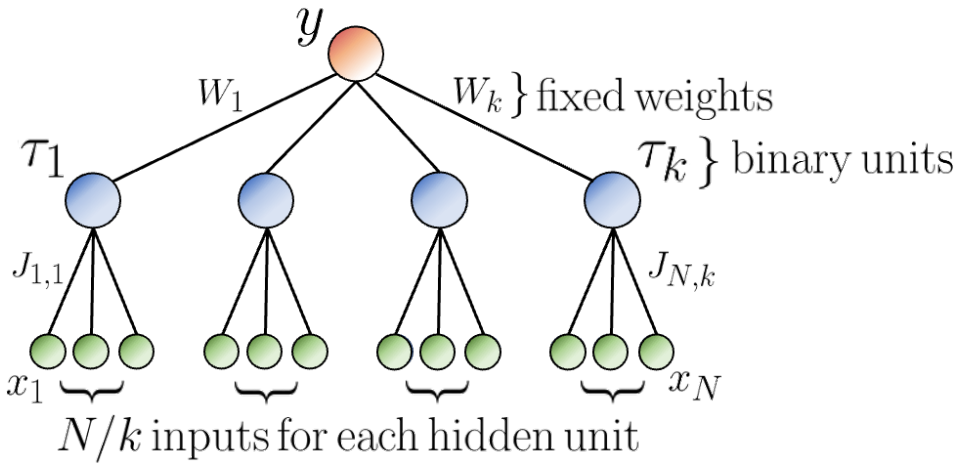

FIG. 11. Schematic representation of a tree-like network with one hidden layer. Here we have a set of k fixed binary units $\{\tau_1,\dots,\tau_k\}$ and N/k inputs per hidden unit. For simplicity the weight $\{W_1,\dots,W_k\}$ are fixed.

($F=\prod_j\tau_j$) and the committee machines ($F=\text{sign}[\sum_j\tau_j]$) [16].

If we now consider a set of input vectors $\{\boldsymbol{x}_\mu\}_{\mu\in[\![1,\dots,\alpha N]\!]}$ and output $\{y_\mu\}_{\mu\in[\![1,\dots,\alpha N]\!]}$ the number of solutions for the weights $\boldsymbol{J}$'s, for which the tree-like network gives the correct output for each vector $\boldsymbol{x}_\mu$, is

$$V(\{\boldsymbol{x}_\mu,y_\mu\})=\int\prod_{j=1}^{k}d\boldsymbol{J}_j\delta(\boldsymbol{J}_j{}^2-N/k)\sum_{\tau_{j,\mu}=\pm1}\prod_{j,\mu}\Theta(\tau_{j,\mu}\boldsymbol{J}_j\boldsymbol{x}_{\mu|j})\prod_{\mu}\Theta[y_\mu F(\tau_{1,\mu},\dots,\tau_{k,\mu})]. \tag{111}$$

For further simplification, in the previous definition, we choose to restrict the $\boldsymbol{J}$'s on a sphere of radius $\sqrt{N/k}$. Considering both random input vectors and outputs the replica-symmetric computation yields an expression of the form (taking $N\to+\infty$ and then $k\to+\infty$)

$$\overline{V(\{\boldsymbol{x}_\mu,y_\mu\})^n}=e^{Nng_{RS}(q,\alpha,K)}\quad\text{with}\quad q=\boldsymbol{J}_j^a.\cdot\boldsymbol{J}_j^b/(N/k)\,. \tag{112}$$

The replica-symmetric computation predicts that there are exponentially many solutions for the weights $\boldsymbol{J}$'s as long as $g_{RS}(q,\alpha,K)>0$. Straightforwardly, this will cease to be true solutions when $q=1$. As schematically represented in Figure 12, the storage capacity $\alpha_c(F)$, i.e. the maximum number of inputs $\boldsymbol{x}_\mu$'s for which we can parameterize a tree-like network outputting the correct y_μ's, corresponds to the point where the replica-symmetric computation predicts an overlap $q=1$ between the replicated weights $\vec{J}_j^a$. For $k\gg1$, the parity model gives $\alpha_c(F)\sim k^2$ while the committee gives $\alpha_c(F)\sim\sqrt{k}$.

However, the true storage capacities are much smaller than what the replica-symmetric computation predicts. Indeed, when computing $\overline{V(\{\boldsymbol{x}_\mu,y_\mu\})^n}$ with the replica-symmetric saddle-point it can be shown that the solution is in fact unstable towards further replica symmetry breaking. To solve this problem we can notice that by fixing the units τ's we obtain back the perceptron problem. This means in particular that the problem becomes replica symmetric in this case, see Figure 13

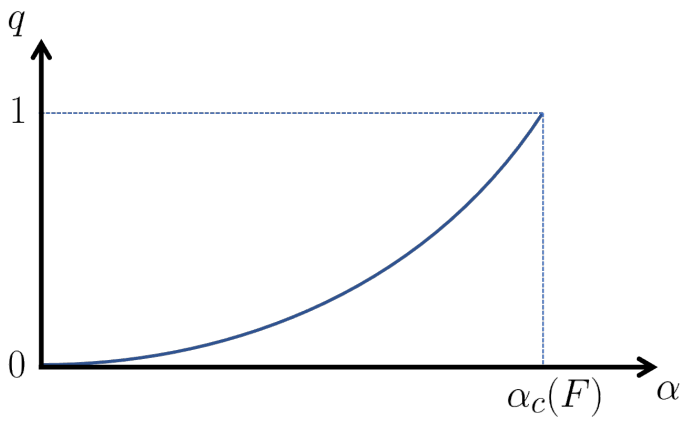

FIG. 12. Schematic representation of the overlap $q = (\boldsymbol{J}_j^a . \boldsymbol{J}_j^b)/(N/k)$ as a function of α (obtained with the replica-symmetric computation). The storage capacity $\alpha_c(F)$ of the tree-like network corresponds to the point where the overlap q between replicas is equal to one.

for a schematic representation of the landscape. Thus we will focus in the following on the modified volume

$$V_{\boldsymbol{\tau}}(\{\boldsymbol{x}_{\boldsymbol{\mu}}, y_\mu\}) = \int \prod_{j=1}^{k} d\boldsymbol{J}_j \delta(\boldsymbol{J}_j{}^2 - N/k) \prod_{j,\mu} \Theta(\tau_{j,\mu} \boldsymbol{J}_j \boldsymbol{x}_{\mu|j}). \tag{113}$$

This quantity can be seen as the volume of configurations $\boldsymbol{J}$'s giving the same representation $\boldsymbol{\tau} = \{\tau_{j\mu}\}_{j\in[\![1,\dots,k]\!],\mu\in[\![1,\dots,\alpha N]\!]}$. The number of representations yielding the same outputs $\{y_\mu\}_{\mu\in[\![1,\dots,\alpha N]\!]}$ scales exponentially as $c(k)^{\alpha N}$ with for example

$$c(k) = 2^{k-1}\,(\text{Parity}), \quad c(k) = \sum_{n=\text{int}(\frac{k+1}{2})^k} \binom{k}{n} \underset{k\gg 1}{\sim} 2^k\,(\text{Committee}). \tag{114}$$

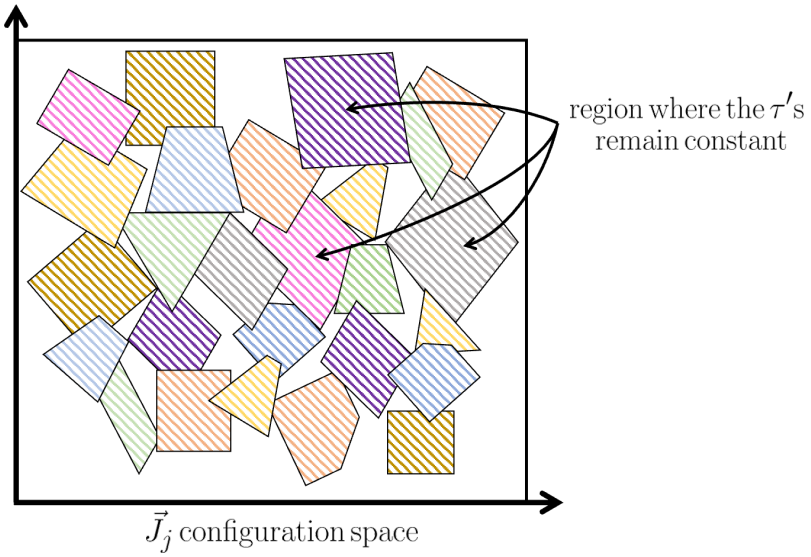

FIG. 13. Schematic representation of the landscape of the tree-like network. In a configurational region of $\boldsymbol{J}$'s where the hidden layer remains constant the landscape for $V(\{\boldsymbol{x}_{\boldsymbol{\mu}}, y_\mu\})$ appears to be replica symmetric.

Several questions can be asked about these regions where the hidden layer τ's remain constant:

- How many of these regions are there?

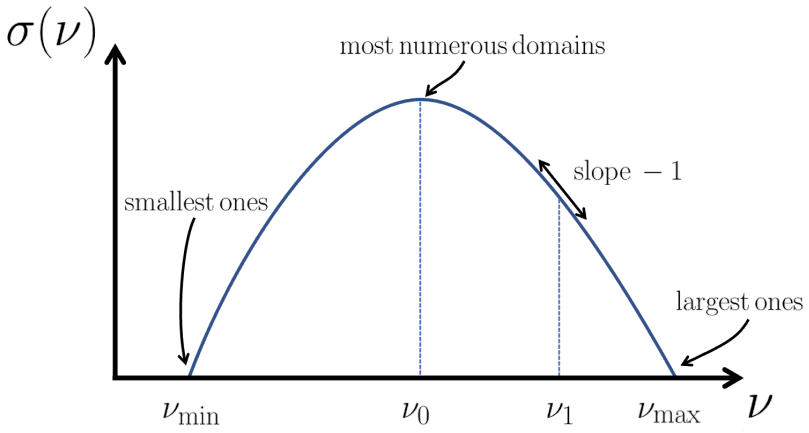

FIG. 14. Curve representing the behavior of $\sigma(\nu)$ as a function of ν. Four particular points can be highlighted. First, $\nu_{\min}$ and $\nu_{\max}$ represent respectively the smallest and largest domains that can be obtained with the tree-like network. Then, ν_0 corresponds to the size of domains that are the most numerous in the system. Last of all, ν_1 is the domain size dominating the total volume of configurations for the hidden layer, see Eq. (116).

- What is their distribution in terms of size? of relative distance?
- How does this population of domains evolve with α?

To answer these questions we first need to define the quantity

$$\left[\text{number of } \tau\text{'s s.t. } V_\tau(\{\boldsymbol{x_\mu}, \boldsymbol{y_\mu}\}) = e^{N\nu}\right] = e^{N\sigma(\nu)}. \tag{115}$$

This corresponds to the number of configurations of hidden units τ's which have the same volume $V_\tau(\{x_\mu, y_\mu\})$. From this follows that the total volume of configurations for the hidden layer is

$$\sum_\tau V_\tau(\{x_\mu, y_\mu\}) = \int d\nu\, e^{N\nu + N\sigma(\nu)} \underset{N\gg 1}{=} e^{N \max_\nu(\nu+\sigma(\nu))}. \tag{116}$$

Using this rewriting enables separating the total volume into two contributions: the first one being the domain size for a given configuration τ's ($e^{N\nu}$) and the second one being the number of configurations with the same domain size $e^{N\nu}$ ($e^{N\sigma(\nu)}$). Moreover, it appears now clear that the total volume for the hidden layer is in fact dominated by configurations τ's with the same volume $\nu_1 = \underset{\nu}{\text{argmax}}(\sigma(\nu)+\nu)$. This is due to the exponential scaling with N of the two contributions involved in Eq. (116), which allows us to evaluate the integral via a saddle-point approximation. For more clarity on the domain sizes, we represent in Figure 14 the behavior of $\sigma(\nu)$ as a function of ν.

To characterize the domains with more precision we compute the generating function

$$G(\beta) = \sum_\tau V_\tau(\{x_\mu, y_\mu\})^\beta \underset{N\to+\infty}{=} e^{N \max_\nu(\beta\nu+\sigma(\nu))}. \tag{117}$$

This technique, also known as "real replicas", enables us to "decouple" the domains entropy $\sigma(\nu)$ from their size ν. It allows us to obtain $\sigma(\nu)$ by inverse Legendre transformation over β[33]. We will see in the following that this computation is tightly linked to a 1-step replica symmetry breaking in the tree-like network.

In practice, and as written in Eq. (117), $G(\beta)$ depends on the data $\{\boldsymbol{x}_\mu, y_\mu\}$. Therefore to compute this quantity we will once again introduce replica and perform an average over $\{\boldsymbol{x}_\mu, y_\mu\}$, i.e. we will evaluate (for $\beta, n \in \mathbb{N}$)

$$\overline{G(\beta)^n} = \overline{\left(\sum_\tau V_\tau(\{x_\mu, y_\mu\})^\beta\right)^n} = \sum_{\tau^1,\dots,\tau^n} \overline{V_{\tau^1}(\{x_\mu, y_\mu\})^\beta \dots V_{\tau^n}(\{x_\mu, y_\mu\})^\beta} \tag{118}$$

with

$$V_{\tau^a}(\{x_\mu, y_\mu\}) = \int \prod_{j=1}^{k} \prod_{\tilde{a}=1}^{\beta} d\boldsymbol{J}_j^{a,\tilde{a}}\, \delta\left((\boldsymbol{J}_j^{a,\tilde{a}})^2 - N/k\right) \prod_{j,\mu} \Theta\left(\tau_{j,\mu}^a \boldsymbol{J}_j^{a,\tilde{a}} \boldsymbol{x}_{\mu|j}\right). \tag{119}$$

The new order parameter for this computation is then

$$q_j^{a\tilde{a},b\tilde{b}} = \frac{1}{N/k} \boldsymbol{J}_j^{a,\tilde{a}} \cdot \cdot \boldsymbol{J}_j^{b,\tilde{b}} . \tag{120}$$

The *Ansatz* we will take for this overlap is of the form

$$\forall j \in [\![1,\dots,k]\!], \quad q_j^{a\tilde{a},b\tilde{b}} = \begin{cases} 1 & \text{if } a=\tilde{a} \quad \text{and} \quad b=\tilde{b} \\ q^* & \text{if } a=b \quad \text{and} \quad \tilde{a}\neq\tilde{b} \\ q & \text{if } a\neq b \end{cases} \tag{121}$$

and consequently

$$\overline{G(\beta)^n} = e^{nN \underset{q,q^*}{\text{extr}}[g(q,q^*,\alpha,\beta)]} . \tag{122}$$

As mentioned earlier this *Ansatz* is similar to a 1-step replica symmetry breaking. This choice for the form of the order parameter can be explained as follows. First, for one given configuration τ^a of the hidden layer, the replica configurations $\{\boldsymbol{J}_j^{a,\tilde{a}}\}_{\tilde{a}\in[\![1,\dots,\beta]\!]}$ lie in a domain associated with a replica symmetric landscape. Thus, the self-overlap between two distinct replicas will take one value q^*. Then, if we look at two replica configurations $\{\boldsymbol{J}^{a,\tilde{a}}\}_{\tilde{a}\in[\![1,\dots,\beta]\!]}$ and $\{\boldsymbol{J}^{b,\tilde{b}}\}_{\tilde{b}\in[\![1,\dots,\beta]\!]}$ lying in two different domains ($\tau^a \neq \tau^b$, $a \neq b$) we will consider with a high probability that they have a finite overlap $q < q^*$. To visualize this better we show in Figure 15 the structure of the domains probed by $G(\beta)$ and the values taken by $q_j^{a\tilde{a},b\tilde{b}}$ under a matrix form.

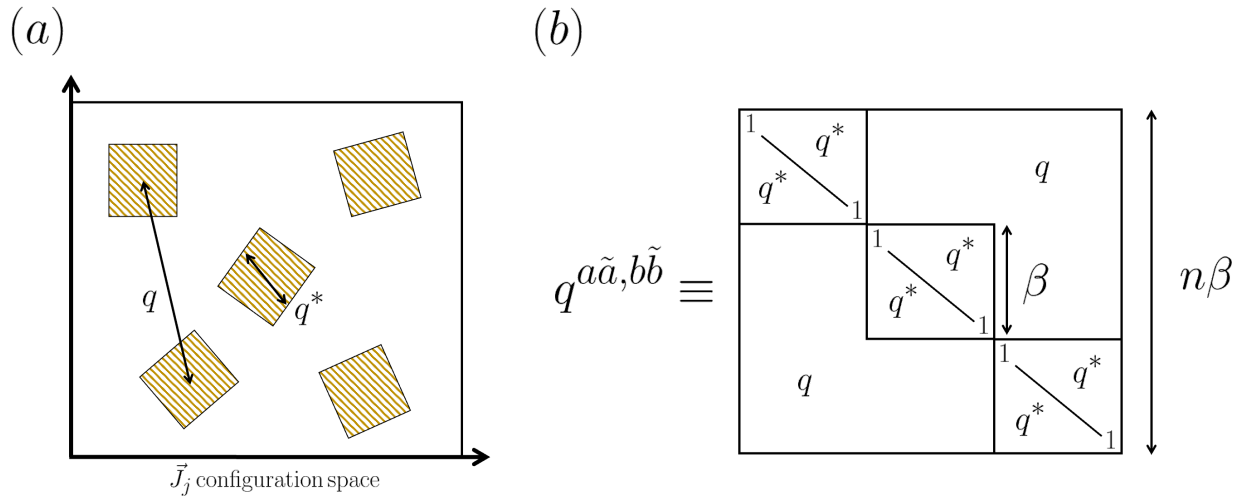

FIG. 15. (a) Representation of the structure of the domains probed by $G(\beta)$. Inside the same domain, τ^a replicas have a fixed overlap q^*. Two replicas lying in different domains will have overlap $q < q^*$. (b) Schematic representation under a matrix form of the order parameter $q^{a\tilde{a},b\tilde{b}}$

It is important to note that we reobtain the replica symmetric computation directly with this generating function $G(\beta)$ when setting $\beta = 1$. And moreover, we can recall that the $\beta = 1$ case corresponds to the dominant volumes as predicted in Eq. (116). Therefore, in order to seize properly the 1-step replica symmetry breaking structure of the dominant volumes it is important to determine the first correction in $\beta = 1+\varepsilon$ (with $\varepsilon \ll 1$) of the generating function. By doing so we obtain for the parity network

$$q^* = 1 - \frac{c}{(\alpha k)^2}, \quad q = 0 \quad \text{and} \quad \nu_1 = \log k - \alpha \log 2 \quad (1 \ll \alpha \ll k) \tag{123}$$

and for the committee network

$$q^* = 1 - \frac{c}{(\alpha k)^2}, \quad q = 1 - \frac{c'}{\alpha^2} \quad \text{and} \quad \nu_1 = \log k - \left(\frac{\pi}{16}\alpha\right)^2 \tag{124}$$

with c and c' two known constants. It is not surprising that the intra-domain overlap is controlled by the effective load αk. A fixed representation $\boldsymbol{\tau}$ defines a perceptron problem, with load $P/(N/k) = \alpha k$, see Figure 11.

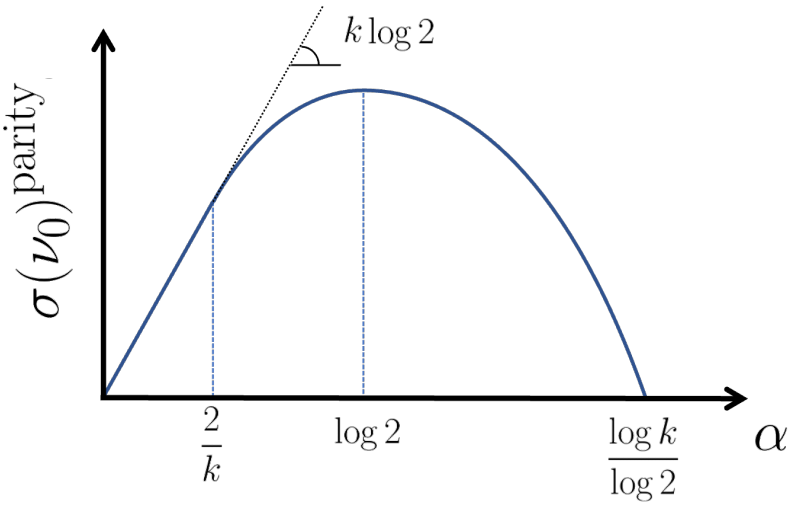

FIG. 16. Schematic representation of $\sigma(\nu_0)$ as a function of α for the parity machine. Below $\alpha = 2/k$ all representations are realizable, a behavior independent of the machine's type (parity, committee, ...). In this case $\sigma(\nu_0) = \alpha k \log 2$. Above $\alpha = 2/k$ the behavior depends on the machine's type. In the case of the parity machine, no cluster can be found above $\alpha = \log k / \log 2$.

In the case of the parity network, an obvious bound for the replica symmetric cluster is $\alpha_c = \frac{\log k}{\log 2}$. For larger α the volume becomes sub-exponential in N, see Eq. (123). Moreover, we emphasize that the behavior of the perceptron with α is quite different from the one of the multi-layer network. While the landscape in the perceptron corresponds to a replica-symmetric domain which shrinks when we increase α, increasing α in a multi-layer network results in the killing of the domains with the largest and smallest volume v. Notice that the RS Ansatz can still be valid despite the presence of an exponentially large number of connected components, since the only overlap that can be probed by picking up two solutions at random is the inter-overlap q.

Finally, we can also probe the most numerous domains with the generating function (corresponding to ν_0 in Fig. 14). To do so we simply have to set β to zero. When $\alpha < 2/k$ all representations are realizable and the result is straightforward. Indeed this means that the number of the most numerous domains is of order $(2^k)^p$, and thus $\sigma(\nu_0) = \alpha k \log 2$. However in this case the size of these domains is sub-exponential, i.e. $\nu_0 < 0$ and $V_\tau(\{\boldsymbol{x}_\mu, y_\mu\}) \ll \mathcal{O}(1)$. Above $\alpha = 2/k$ we obtain for example with the parity machine

$$\sigma(\nu_0)^{\text{parity}} = \alpha k \log k - (\alpha k - 1)\log(\alpha k - 1) - \alpha \log 2 \quad \text{and} \quad \alpha < \frac{\log k}{\log 2}. \tag{125}$$

We summed up the behavior of $\sigma(\nu_0)$ for the parity machine in Figure 16.

To close this section, let us briefly discuss generalization in this domain-based framework. In other words, if we have a teacher, where does it lie in the domain picture? In the case of a random teacher (random set for $\{\boldsymbol{x}_\mu, y_\mu\}$), it will lie in one of the most probable domains with $\nu = \nu_1$. In general, for small α, student and teacher will lie in different clusters and the generalization error will be large. This error will drop when, after disappearance of many domains, the student will lie in the same domain as the teacher.

V. LECTURE 5. RESTRICTED BOLTZMANN MACHINES: OVERVIEW AND APPLICATIONS

Restricted Boltzmann Machines [1, 23] are prototypical models for features extraction from a structured dataset in an unsupervised way. Their training is based on the maximization of the likelihood function. Since the likelihood function refers to the best-fitting probability distribution of reality (the dataset), they are also useful as generative models, *i.e.* to produce different but similarly distributed examples w.r.t. the training samples. This Lecture is devoted to a brief overview of Restricted Boltzmann Machines, and their relation with statistical-mechanical models.

A brief introduction to Restricted Boltzmann Machines

Let us go back to the unsupervised problem of inferring a preferred direction, say $\boldsymbol{W}$, in a high-dimensional space from a set of data $\mathcal{D} = \{\boldsymbol{x}_d\}_{d=1}^D$, $\boldsymbol{x}_d \in \mathbb{R}^N$ for all d. Assuming a Gaussian prior on the vectors $\boldsymbol{x}_d$, the probability distribution

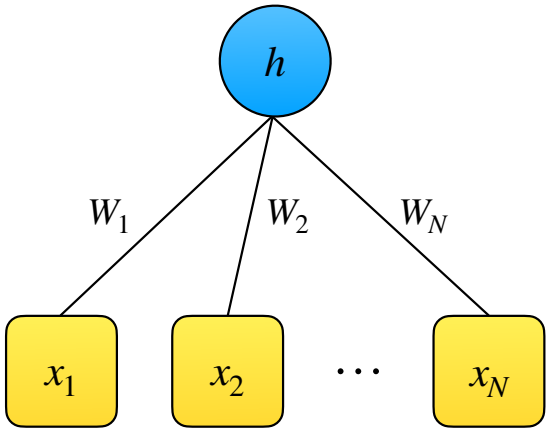

FIG. 17. Schematic representation of RBM with one hidden neuron. The figure reports a representation of the probabilistic model in Eq. (126) through the formal expression for the function ϕ. The direction $\boldsymbol{W}$ biasing the data in the high-dimensional space $\mathbb{R}^N$ plays, in this equivalence, the role of interaction weights between the visible neurons and the hidden unit.

describing the statistics of the data is given by

$$P(\boldsymbol{x}|\boldsymbol{W}) = \frac{1}{\mathcal{Z}} \exp\Big(-\frac{1}{2}\sum_{i=1}^{N} x_i^2 + \phi(\boldsymbol{x}\cdot\boldsymbol{W})\Big), \tag{126}$$

where $\mathcal{Z}$ is the partition function (ensuring normalization) and ϕ is a generic function quantifying the bias of the data points in the direction of $\boldsymbol{W}$, and whose form can be arbitrarily complex, see Lecture 3.

As we saw, we can consider the problem of inferring the direction $\boldsymbol{W}$ by considering the probability distribution $P(\boldsymbol{W}|\mathcal{D})$, which – for simple enough $\phi(u)$ – can be approached with statistical mechanics, see for example [36]. In this case, we would rather follow another route, and assume that the function ϕ can be represented in a formal way as

$$\exp(\phi(u)) = \int_{-\infty}^{\infty} dh \exp(hu - U(h)) \,. \tag{127}$$

For instance, if $\phi(u) = au^2/2$ (apart from unessential multiplicative constants in front) the previous equality reduces to a Hubbard-Stratonovich transformation, leading to $U(h) = h^2/(2a)$. Thus, the probability distribution (126) can be written as

$$P(\boldsymbol{x}|\boldsymbol{W}) = \int_{-\infty}^{\infty} dh P(\boldsymbol{x}, h),$$

with

$$P(\vec{x}, h) = \frac{1}{\mathcal{Z}} \exp\Big(-\frac{1}{2}\sum_{i=1}^{N} x_i^2 - U(h) + h\sum_{i=1}^{N} W_i x_i\Big). \tag{128}$$

In other words, we introduced an auxiliary variable h so that the 'complex' dependence on $\boldsymbol{x}$ through the function ϕ is linearized. The probabilistic model in Eq. (128) can be interpreted as an interacting bipartite graph where the variables x_i are associated with visible nodes, while the h variable is called a hidden node. The resulting scheme is a particular example of Restricted Boltzmann Machine (RBM), as represented in Figure 17, where "restricted" stands for the fact that there is no direct interaction within the visible layer. Furthermore, the direction vector $\boldsymbol{W}$ plays the role of interaction weights between visible units and the hidden neuron. Due to the restricted nature of the model, an important role is covered by the conditional probabilities $P(\boldsymbol{x}|h)$ and $P(h|\boldsymbol{x})$, as we will see in the next Section. In particular, the former factorizes as $P(\boldsymbol{x}|h) = \prod_{i=1}^{N} P(x_i|h)$.

More generally, it is possible to extend the model in order to capture information from more than one direction in the $\mathbb{R}^N$ space, say $\boldsymbol{W}_\mu$ with $\mu = 1, \ldots, M$. The most general model of Restricted Boltzmann Machine is described by the probability distribution

$$P(\boldsymbol{x}, \boldsymbol{h}) = \frac{1}{\mathcal{Z}} \exp\Big(-\sum_{i=1}^{N} V_i(x_i) - \sum_{\mu=1}^{P} U_\mu(h_\mu) + \sum_{i\mu} W_{i\mu} x_i h_\mu\Big), \tag{129}$$

where $V_i(x_i)$ and $U_\mu(h_\mu)$ encode resp. the priors on the visible and hidden units. A representation for the probabilistic model defined in Eq. (129) is given in Figure 18, with a structure analogous to models for cognitive science [38]. The resulting RBM is a bipartite graph composed of a visible layer (consisting of N variables x_i) and a hidden (or *representation*) one, made up of M variables h_μ. The activity of neurons in each layer is described by the conditional probabilities $P(\boldsymbol{x}|\boldsymbol{h}) = \prod_{i=1}^N P(x_i|\boldsymbol{h})$

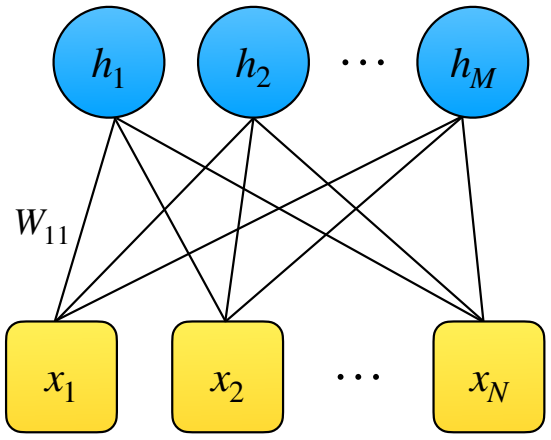

FIG. 18. Schematic representation of RBM with M hidden units. The figure shows the natural extension of the 1-hidden unit case. In this scenario, the model is equivalent to a situation in which the dataset is biased by M different direction $\boldsymbol{W}_\mu$ in the high-dimensional space $\mathbb{R}^N$.

and $P(\boldsymbol{h}|\boldsymbol{x}) = \prod_{\mu=1}^M P(h_\mu|\boldsymbol{x})$ – again, factorization follows from the restricted nature of the model, with

$$P(x_i|\boldsymbol{h}) \;\propto\; \exp(-V_i(x_i) + x_i \sum_\mu W_{i\mu} h_\mu), \tag{130}$$

$$P(h_\mu|\boldsymbol{x}) \;\propto\; \exp(-U_\mu(h_\mu) + h_\mu \sum_i W_{i\mu} x_i). \tag{131}$$

Let us now focus on the activity of the hidden layer. From Eq. (131), we see that the most probable response of one of the h_μ variables, say $\mu = 1$, is given by

$$h^\star \quad : \quad -U_1'(h^\star) + \sum_{i=1}^N W_{i1} x_i = 0.$$

Equivalently, in case the first derivative of U_1 is invertible, we can recast the previous equation as

$$h^\star = (U_1')^{-1}(u) \equiv \varphi_1(u),$$

where $u = \sum_{i=1}^N W_{i1} x_i$ is the net input signal to the hidden node h_1, and $(U_1')^{-1}$ plays the role of activation function φ_1. A schematic representation of the situation is reported in Fig. 19 (left).

For example, for quadratic $U_1(h) = ah^2/2$, we have $h^\star = u/a$, corresponding to a linear activation function with slope $1/a$. Instead, taking

$$U_1(h) = \begin{cases} \frac{h^2}{2} + h\theta & h \geq 0, \\ +\infty & h < 0, \end{cases} \tag{132}$$

the associated activation function potential $\varphi_1(u) = [u - \theta]_+$ is a ReLU with threshold θ (which can be properly chosen), see Figure 19 (right) for a representation.

The same analysis can be performed on units in the visible layer. In the case of quadratic potential $V_i(x_i)$, the conditional probability $P(x_i|\boldsymbol{h})$ is a Gaussian distribution centered around the value $\tilde{u}_i = \sum_{\mu=1}^M W_{i\mu} h_\mu$, which is the net input signal coming from the hidden layer.

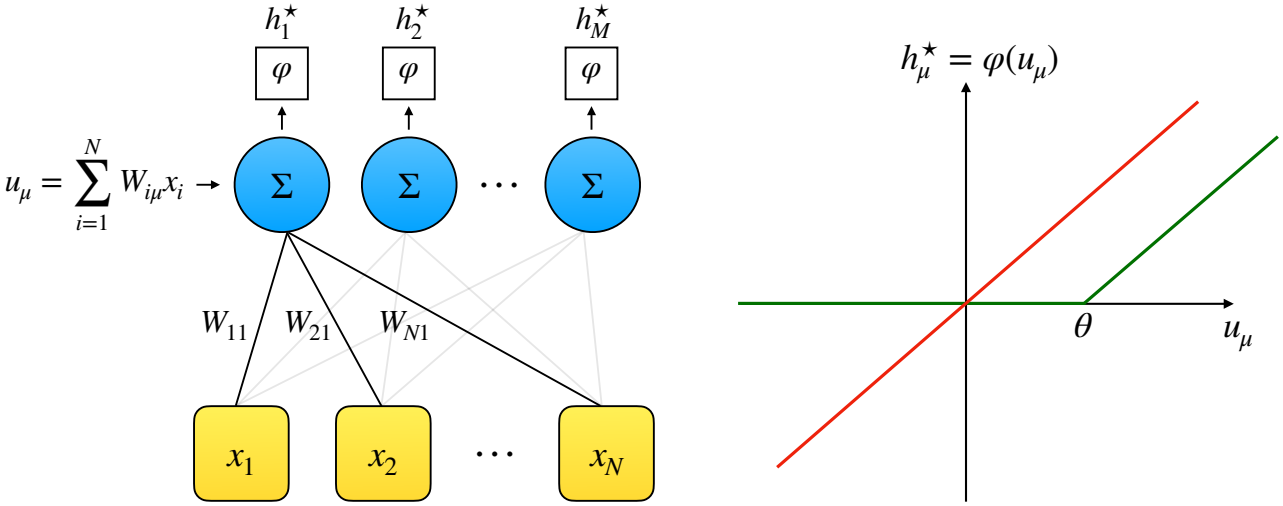

FIG. 19. Hidden layer activity of the RBM. On the left side, we have a schematic representation of hidden layer activity. The visible layer is fixed to a specific data point $\boldsymbol{x}$ in the dataset $\mathcal{D}$. Thus, the signal for each of the hidden units coming from the visible ones is the weighted sum of the inputs: $u_\mu = \sum_i W_{i\mu} x_i$. The most probable configuration of the hidden layer is then obtained by subjecting the input signal to the activation function: $h_\mu^\star = \varphi(u_\mu)$. On the right side, two specific examples of activation functions: the red curve refers to a linear response for the hidden layer (corresponding to a Gaussian prior for the hs variables), while the green one to a ReLU activation function with threshold θ, resulting from the potential defined in Eq. (132).

Training of Restricted Boltzmann Machines

As we stated in the introduction, Restricted Boltzmann Machines are the prototype models for inferring in an unsupervised way specific directions (or features) $\boldsymbol{W}_\mu$ characterizing a set of data $\{\boldsymbol{x}_d\}_{d=1}^D$ in a N-dimensional space. The aim of the training procedure of RBMs is to model the target distribution describing the statistics of the dataset in terms of the model distribution $P(\boldsymbol{x})$. As a consequence, these models can also be used as a sampling algorithm (meaning that they are *generative models*). The training procedure of RBM is usually performed in terms of the (log-)likelihood function (an empirical version of the Kullback-Leibler divergence [26]), which is defined as

$$\mathcal{L}(\{\boldsymbol{W}_\mu\}_{\mu=1}^M|\mathcal{D}) = \frac{1}{D}\sum_{\boldsymbol{x}_d\in\mathcal{D}} \log P(\vec{x}_d), \tag{133}$$

where $P(\boldsymbol{x}_d)$ is the probability of the data point $\boldsymbol{x}_d$ according to the model distribution:

$$P(\boldsymbol{x}_d) = \frac{\int \prod_\mu dh_\mu \exp\left(-\sum_{\mu=1}^M U_\mu(h_\mu) - \sum_{i=1}^N V_i(x_i^d) + \sum_{i\mu} W_{i\mu} h_\mu x_i^d\right)}{\int \prod_i dx_i \prod_\mu dh_\mu \exp\left(-\sum_{\mu=1}^M U_\mu(h_\mu) - \sum_{i=1}^N V_i(x_i) + \sum_{i\mu} W_{i\mu} h_\mu x_i\right)}.$$

By setting up a maximum likelihood problem via gradient ascent algorithm, it is possible to properly the network parameters (the interactions weights $W_{i\mu}$) as

$$\frac{dW_{i\mu}}{dt} = \frac{\partial}{\partial W_{i\mu}} \mathcal{L}(\{\boldsymbol{W}_\mu\}_{\mu=1}^M|\mathcal{D}).$$

By explicit computations, it is easy to prove that the update rule for the network weights is simply (in the discrete version)

$$\Delta W_{i\mu} = \epsilon\left(\langle x_i^d \langle h_\mu\rangle_c(\boldsymbol{x}^d)\rangle_{\mathcal{D}} - \langle x_i h_\mu\rangle_f\right), \tag{134}$$

where ϵ is the learning rate, and

$$\langle x_i^d \langle h_\mu\rangle_c(\boldsymbol{x}^d)\rangle_{\mathcal{D}} = \frac{1}{|\mathcal{D}|}\sum_{\boldsymbol{x}_d\in\mathcal{D}} x_i^d \int \prod_\mu dh_\mu P(\boldsymbol{h}|\boldsymbol{x}_d) h_\mu, \tag{135}$$

$$\langle x_i h_\mu\rangle_f = \int \prod_i dx_i \int \prod_\mu dh_\mu P(\boldsymbol{x},\boldsymbol{h})\, x_i h_\mu. \tag{136}$$

In the first quantity, we compute the empirical mean of $x_i h_\mu$ on the dataset after taking the average over the hidden layer activity – according to the conditional probability distribution $P(\boldsymbol{h}|\boldsymbol{x}_d)$ – with the visible neurons are fixed to the data (c stands for conditioned-RBM with fixed visible layer activity). Conversely, in the second line, we compute the correlation function on the full model, *i.e.* w.r.t. the joint probability distribution $P(\boldsymbol{x},\boldsymbol{h})$ of the full RBM. Thus, the learning rule in Eq. (134) is eventually a moment-matching criterion, and the fixed point is achieved when the correlation function of the model distribution recovers the data-conditioned behavior of the model, *i.e.* when the RBM account for the statistics of the data.

Correlation functions as in Eq. (135) is hard to evaluate in practical scenarios for two main points. First of all, depending on the values of the network parameters (and the number M of relevant directions in the dataset), it could be hard to sample from the target distribution $P(\boldsymbol{x},\boldsymbol{h})$. We will deepen this point later on, when considering the relation between binary-binary RBMs and the Hopfield model. Further, from a practical point of view, Eq. (134) is hard to handle as it is since it would require the computation of the full partition function of the model. As a consequence, training RBMs can be addressed by setting up a numerical estimation of the model correlation function. The simplest way to do it is via Monte Carlo Markov Chains (MCMC) whose relaxation towards equilibrium is described by the joint distribution $P(\boldsymbol{x},\boldsymbol{h})$, see [21, 31] and references therein. However, these procedures are typically very slow (the MCMC needs a large number of iterations for the sampling according to the true distribution), and the estimated gradient suffers of large variance. As an alternative procedure, one can start the MCMC at the data distribution, *i.e.* the visible layer is clumped to data points $\boldsymbol{x}_d$, then the system is updated for a small number K of steps according to the conditional probabilities in Eqs. (130) and (131). Together with the update rule in Eq. (134), this procedure is the so-called Contrastive Divergence (CD_K) [24]. An example of the training procedure for a RBM with the MNIST dataset is reported in Figure 20. Other alternative procedure (similar in spirit to the CD prescription) can be found in the literature, see for example the persistent Contrastive Divergence (PCD) [18, 42].

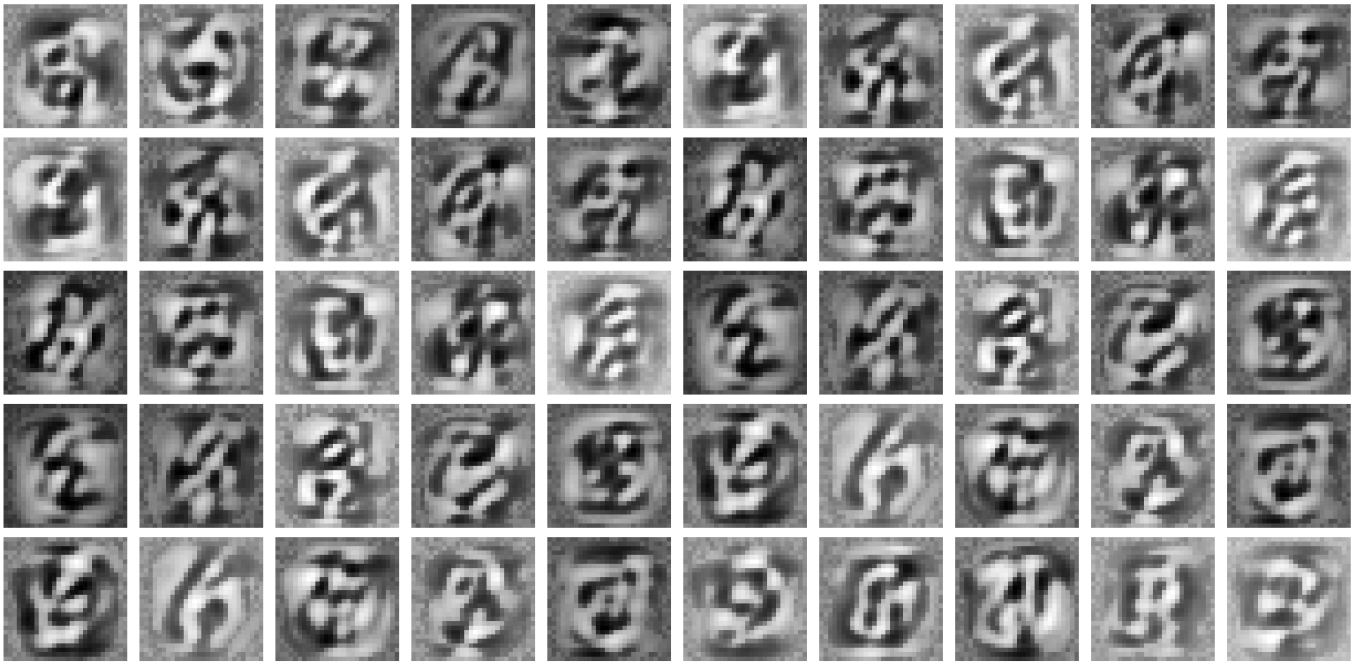

FIG. 20. Graphical representation of extracted features with a RBM. The figure shows the graphical representation of $M = 50$ extracted features from the MNIST dataset with a RBM trained with CD_{10}. The learning strength is fixed to $\epsilon = 5 \cdot 10^{-5}$, the number of epochs is 2000, and the gradient is averaged over minibatches of 200 examples each.

In general, understanding the learning dynamics of unsupervised training of RBMs is a highly non-trivial problem. By expanding the log-likelihood at small $W_{i\mu}$, we have

$$\frac{\partial \mathcal{L}}{\partial W_{i\mu}} \simeq \sum_j C_{ij} W_{i\mu} - W_{i\mu},$$

where $C_{ij} = D^{-1} \sum_{d=1}^{D} x_i^d x_j^d$ is the empirical covariance matrix of the dataset. It is possible to show that, in this regime (*i.e.* during the very early stages of learning), the RBM training is driven by the principal modes of the dataset, as the modes of

the SVD of the feature matrix are modified accordingly [13].[34] Several approaches have been carried out in the literature to understand learning, see for instance [19, 25, 40] for a TAP equations approach, and [11, 12] for a statistical mechanics description. A detailed description of the emergence of feature extraction in the case of invariant data can be found in [22].

A. Binary-Gaussian random RBM and the Hopfield model

In the context of understanding information processing principles in RBM learning procedures, Statistical Mechanics approach is fruitful in investigating the working regimes of the model, especially in determining the behavior of free-energy as a function of the external parameters (and consequently the phase transitions of the model), and to figure out the conditions for which it is possible to sample points according to $P(\boldsymbol{x})$. Indeed, as Statistical Mechanics teaches us, it could be possible for the system to be located – within the parameter space – in a spin-glass phase, meaning that an exponential time is needed to escape from spurious minima for the free-energy. Even in the ideal region (which is called the *retrieval* phase), where the free-energy landscape is dominated by "good" minima (from the point of view of dataset statistics), it could happen that the probability to randomly jump in these wells is low (because of the presence of large barriers in the landscape). Thus, in principle, it could be difficult to reach a good sampling, and consequently a good numerical estimation of relevant correlation functions.

To understand this, let us consider a binary-Gaussian RBM (resp. visible-hidden layers), with i.i.d. random extracted features $W_{i\mu} = \xi_i^\mu/\sqrt{N}$, with $P(\xi_i^\mu = \pm 1) = 1/2$. Thus, the joint probability distribution of the model reads

$$P(\boldsymbol{x}, \boldsymbol{h}) = \frac{1}{\mathcal{Z}}\left(\frac{1}{2\pi\beta}\right)^{M/2} \exp\Big(-\frac{1}{2\beta}\sum_\mu h_\mu^2 + \sum_{i\mu} W_{i\mu} x_i h_\mu\Big). \tag{137}$$

Notice that here $h_\mu \sim_{i.i.d.} \mathcal{N}(0, \beta^{-1})$. Since the ultimate goal of training the RBM consists in capturing the statistics of the dataset $\mathcal{D}$ and sample according to the model, it would be useful to consider the marginal distribution $P(\boldsymbol{x})$, describing the relaxation towards equilibrium of the visible layer, which turns out to be

$$P(\boldsymbol{x}) = \mathcal{Z}^{-1} \exp\Big(\frac{\beta}{2}\sum_{ij} J_{ij} x_i x_j\Big), \tag{138}$$

with $J_{ij} = N^{-1}\sum_{\mu=1}^M \xi_i^\mu \xi_j^\mu$. This distribution is the same as the one of the Hopfield model, meaning that the latter exhibits the same equilibrium dynamics of the RBM once the hidden layer is marginalized out [5]. Remarkably, the statistical mechanics of the Hopfield model has been extensively studied, and in particular the phase diagram in the (α, T) parameter space is well-known (where $\alpha = \lim_{N\to\infty} M/N > 0$ and $T = \beta^{-1}$) [4], and it is reported in Figure 21. The phase diagram consists of three main regions (below, $\overline{\cdot}$ is the average w.r.t. the probability distribution of the patterns $\vec{\xi}^\mu$):

- In the region I (retrieval or *ferromagnetic*) phase), a single pattern $\boldsymbol{\xi}^\mu$ is retrieved. The value of μ depends on the initial condition over the visible configuration; we hereafter assume with no loss of generality that $\mu = 1$. We have $N^{-1}\overline{\sum_i \langle x_i\rangle \xi_i^\mu} = m\delta_{\mu,1}$ with $m > 0$. The equilibrium configuration of the model has a non-zero overlap with the pattern $\boldsymbol{\xi}^1$ and vanishing one with all of the others vector $\boldsymbol{\xi}^\mu$ with $\mu > 1$;

- In the region II (spin-glass phase), we have $N^{-1}\overline{\sum_i \langle x_i\rangle \xi_i^\mu} = 0$ for all μ but $\overline{\sum_i \langle x_i\rangle^2} = q \neq 0$, meaning that each spin has vanishing overlap with all of the patterns but there exists a locally self-organizing behavior.

- In the region III (*paramagnetic*) phase), the system is fully random, as the thermal noise dominates. The region is delimited from below by the equation $T_c(\alpha) = 1/\beta_c = 1 + \sqrt{\alpha}$ [3].

This argument implies that, in region III, it is impossible to generate data points with desired properties, since the thermalization of the system towards equilibrium completely destroys the information about the stored patterns $\boldsymbol{\xi}^\mu$. In region II, in principle, the measure P over the activity configurations depends on the patterns, but is extremely complex to characterize, and generation is in practice uncontrollable.

Let us now try to translate this analysis in terms of the RBM, which is summarized in Figure 22. Recalling that the model under consideration has potential $U(h) = \frac{1}{2\beta}h^2$, we see that the most probable value for hidden layer activity is

$$h_\mu^\star \equiv u_\mu = \frac{1}{\sqrt{N}}\sum_{i=1}^N \xi_i^\mu x_i = \beta m\sqrt{N}\delta_{\mu,1} + \mathcal{O}(N^0),$$

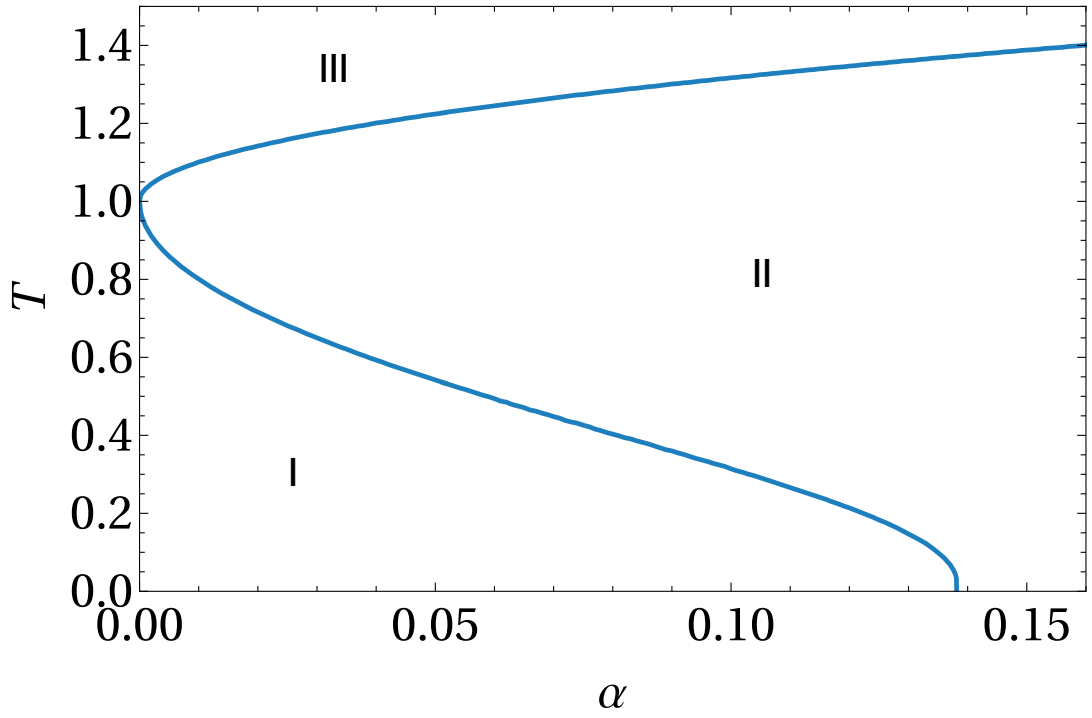

FIG. 21. Phase diagram of the Hopfield model. Region I is the retrieval phase, in which the system has a non-zero magnetization m at the equilibrium. Region II is the spin glass phase, consisting in vanishing magnetization, while the two-replica overlap is non-zero, meaning that the system exhibits local self-organizing behavior. The region III is the paramagnetic phase, where all of the order parameters vanish.

where the last equality holds in the ferromagnetic phase under the hypothesis that only the pattern $\boldsymbol{\xi}^1$ is retrieved to memory. This means that the hidden neuron h_1 (the one associated to the recalled pattern) exhibits a large activity, while all the others exhibit values close to zero. Conversely, the input for the visible layer is

$$\tilde{u}_i = \frac{1}{\sqrt{N}} \sum_{\mu} W_{i\mu} h_\mu = \beta m \xi_i^1 + \mathcal{O}\big(\sqrt{M/N}\big).$$

As is clear, the input to the visible layer in this case is much weaker w.r.t. its hidden counterpart. Further, the noise generated by non-retrieved patterns is $O(\sqrt{M/N})$, meaning that it is more and more important as the storage capacity α increases (in the thermodynamic limit we deal with finite α, the so-called *high storage regime*). This clearly affects the retrieval capabilities of the RBM, as the noise in this case would destroy the information coming from the hidden layer. Also, in the spin-glass phase, where there is no correlation between samples from $P(\boldsymbol{x}, \boldsymbol{h})$ and statistics underlying the dataset, both hidden and visible layer are subjected to noisy contributions, and thus – even preparing the latter in a configuration close to one of the patterns – and thus information is immediately lost as the network relaxes toward equilibrium.

B. Binary-ReLU random RBM

In the previous Section, we saw that the relaxation towards equilibrium of the Binary-Gaussian RBM with Rademacher weights is equivalent to the equilibrium dynamics of the Hopfield model. Further, under the hypothesis of single-pattern retrieval, the corresponding hidden unit receive a large signal (of order of $\sqrt{N}$) while all the others are subjected to noise of the order of 1; however, the visible units receive a signal of the order of 1, which (in the high storage limit) is comparable to the noise generated by non-retrieved patterns, thus harming the information retrieval performances of the RBM. Further, it is worth to stress that - in this scenario - the recall of a single pattern is associated to the strong activation of a single hidden unit, which – in representative sense – is non-optimal. In this Section, we will consider a modification of the RBM architecture, whose main distinctive features are the following:

- The hidden units have a ReLU activation function with threshold θ, and the latter is tuned in order to filter out noisy contributions coming from the visible layer;

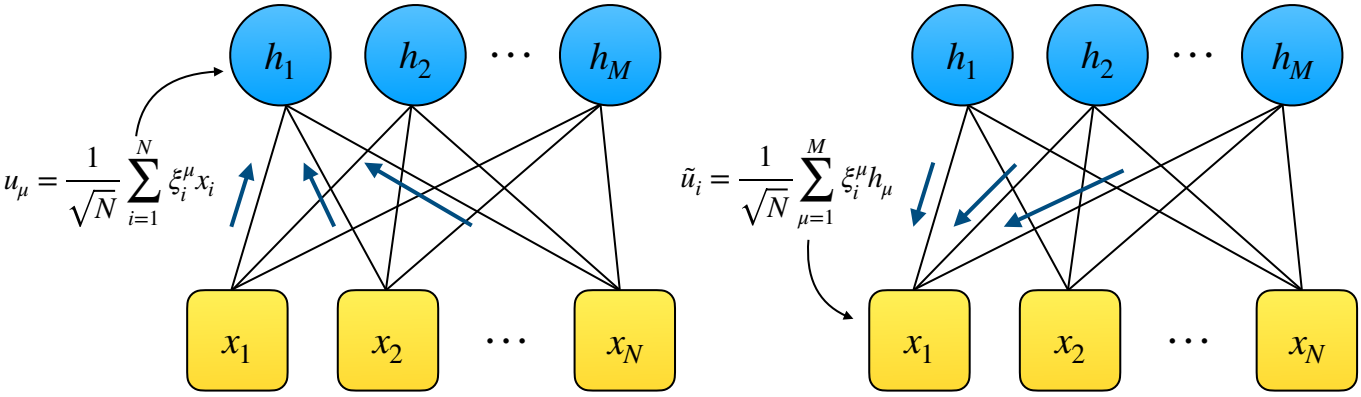

FIG. 22. Input signals for the hidden and visible layer respectively in the binary-Gaussian RBM. On the left side, the input signal for the hidden layer. If the visible layer is aligned with the pattern $\boldsymbol{\xi}^1$, the hidden unit h_1 receives a large signal of order N, while all the others are subjected to an input $\mathcal{O}(N^0)$. On the right side, the input signal for the visible layer. In the high storage regime $M = \alpha N$, the contributions for hidden units with large or weak activity are of the same order.

- In order to increase the representative power of the model, we want that more than one hidden units be active with large signal. A possible way to increase the representation power of the RBM is to introduce sparsity (or *dilution*) in the weights matrix [2, 44, 45] according to the following scheme:

$$W_{i\mu} = \begin{cases} 0 & 1-p \\ \frac{W}{\sqrt{N}} & \frac{p}{2} \\ -\frac{W}{\sqrt{N}} & \frac{p}{2} \end{cases},$$

  where the parameter p controls the sparsity in the feature representation.

In addition, we assume that the visible units $x_i \in \{0, 1\}$ visible units, and are subjected to an external field g (whose effect is to bias the direction of the x_i depending on its sign). The joint probability distribution of the model thus reads

$$P(\boldsymbol{x}, \boldsymbol{h}) = \frac{1}{\mathcal{Z}} \exp\Big(-\sum_\mu U(h_\mu) + \sum_\mu \theta h_\mu + \sum_{i\mu} W_{i\mu} x_i h_\mu + g \sum_i x_i \Big),$$

where now the h_μ are constrained to be positive because of the ReLU potential: $U(h) = \frac{1}{2}h^2$ if $h \geq 0$, and $U(h) = +\infty$ for $h < 0$.

When N is large and $p < 1$, the number of zero components in each feature vector $\boldsymbol{W}_\mu$ is $\sim (1-p)N$. Thus, for p small enough, configurations aligning with more than one pattern (in particular, with high positive overlap $\boldsymbol{W}_\mu \cdot \boldsymbol{x}$) are more probable than the single-retrieval case. In the case of $g > 0$, such configurations are also favored, since its contributions would bias the visible units to take values 1. Thus, it is reasonable to expect that, with proper choices of the network parameters, there exists a "compositional" phase, where more than one hidden units receive a large signal, while all the others receive small perturbation around zero (and can be filtered out by means of the ReLU threshold θ). Let us assume that a number $L > 1$ of hidden units receive are strongly active. The most probable configuration of hidden units is

$$h_\mu^\star = \sum_{i=1}^{N} W_{i\mu} x_i - \theta,$$

if $h_\mu^\star \geq 0$, and zero otherwise. Let us now assume that x_i are strongly correlated with positive entries of the feature $\boldsymbol{W}_\mu$ (for simplicity, let us forget for the moment about the threshold). Thus, we have

$$h_\mu^\star \simeq \frac{W}{\sqrt{N}} \frac{p}{2} N \sim p\sqrt{N},$$

corresponding to units experiencing a large input signal. Conversely, if the visible layer configuration is not correlated with the μ-th feature, we have

$$|h_\mu^\star| \simeq \frac{W}{\sqrt{N}} \sqrt{pN} \sim \sqrt{p}.$$

Thus, setting a threshold $\theta \gtrsim \sqrt{p}$leads to a large shutdown of hidden neurons with weak input signal.

Let us now consider the effects of introducing the threshold in the activation function. The input signal without the thresholding mechanism (and neglecting the homogeneous field g) would be

$$\tilde{u}_i^{w/o} \sim \frac{W}{\sqrt{N}} p(Lp\sqrt{N}) = LWp^2 \ ,$$

where the weakly activated hidden units are silenced through the appropriate choice of the threshold.

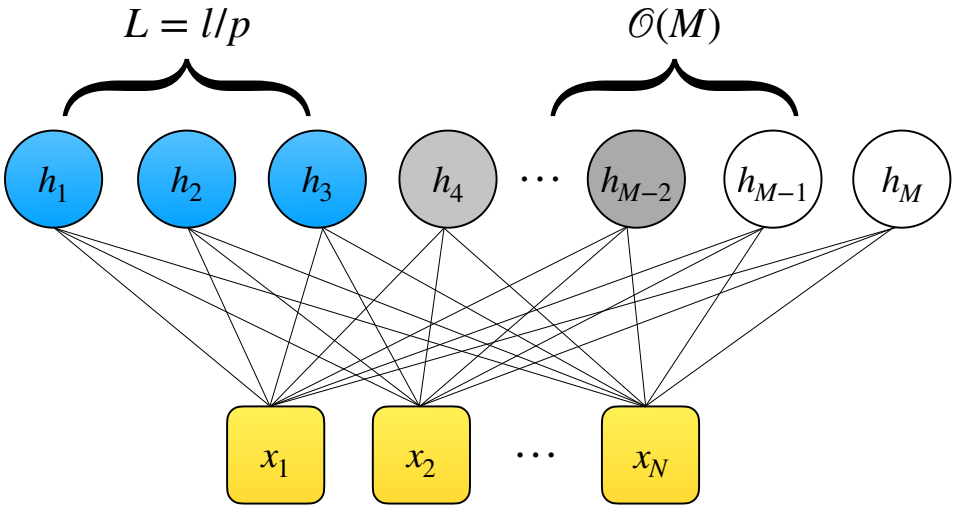

FIG. 23. Schematic representation of compositional phase for random RBM with thresholded hidden units. When the feature vectors are sparse ($p \ll 1$), a compositional phase in the parameter space exists, and L hidden units can exhibits a large activity, while all the others (a number $\mathcal{O}(M)$ of hidden neurons) can be weakly active or even shutted down by the thresholding mechanism.

The statistical mechanics of the Binary-ReLU random RBM can be studied via replica theory. Within the replica-symmetry Ansatz, an explicit expression for the quenched free-energy can be derived [43] and studied in terms of the control parameters p, α and g. In particular, it is found that, for sparse enough weight matrix (*i.e.* $p \ll 1$), a compositional phase is observed with a number of strong-activity hidden neuron $L = l/p$ with l finite, and a number $O(N)$ of inactive or weakly-activated ones. In this sense, the model exhibits a better representation power w.r.t. the usual RBM, since more hidden units can be used to code the information coming from the data, and the machine is able to simultaneously process different features. This result extends the results obtained for diluted Hopfield model with a finite number of patterns [2] to non quadratic potentials, *i.e.* to models with high-order interactions between the visible units. As a final comment, this machinery can be used also to extract information from real data, represent them and sample according to the resulting model statistics, while controlling the weight matrix sparsity with regularization terms in the training cost-function. The whole machinery can be used for extracting compositional representations of real structured datasets, see for example [43] for applications to MNIST and [44, 45] to alignments of protein sequences.

Further topics on RBM

The applications of statistical mechanics to RBM are numerous, for a recent review see [14]. We briefly expose two of them below; this choice is purely subjective.

RBM for data representation disentangling

Consider a two-population distribution of data points in a high dimensional space. The goal is to represent them – by means of RBM – in a latent space in which the two classes are disentangled, see Figure 24 for a schematic description of the problem.

Disentangling representations is a general problem in unsupervised learning. Two common strategies to achieve this goal is to enforce orthogonality between the latent factors. Principal component analysis, arguably the simplest high-dimensional

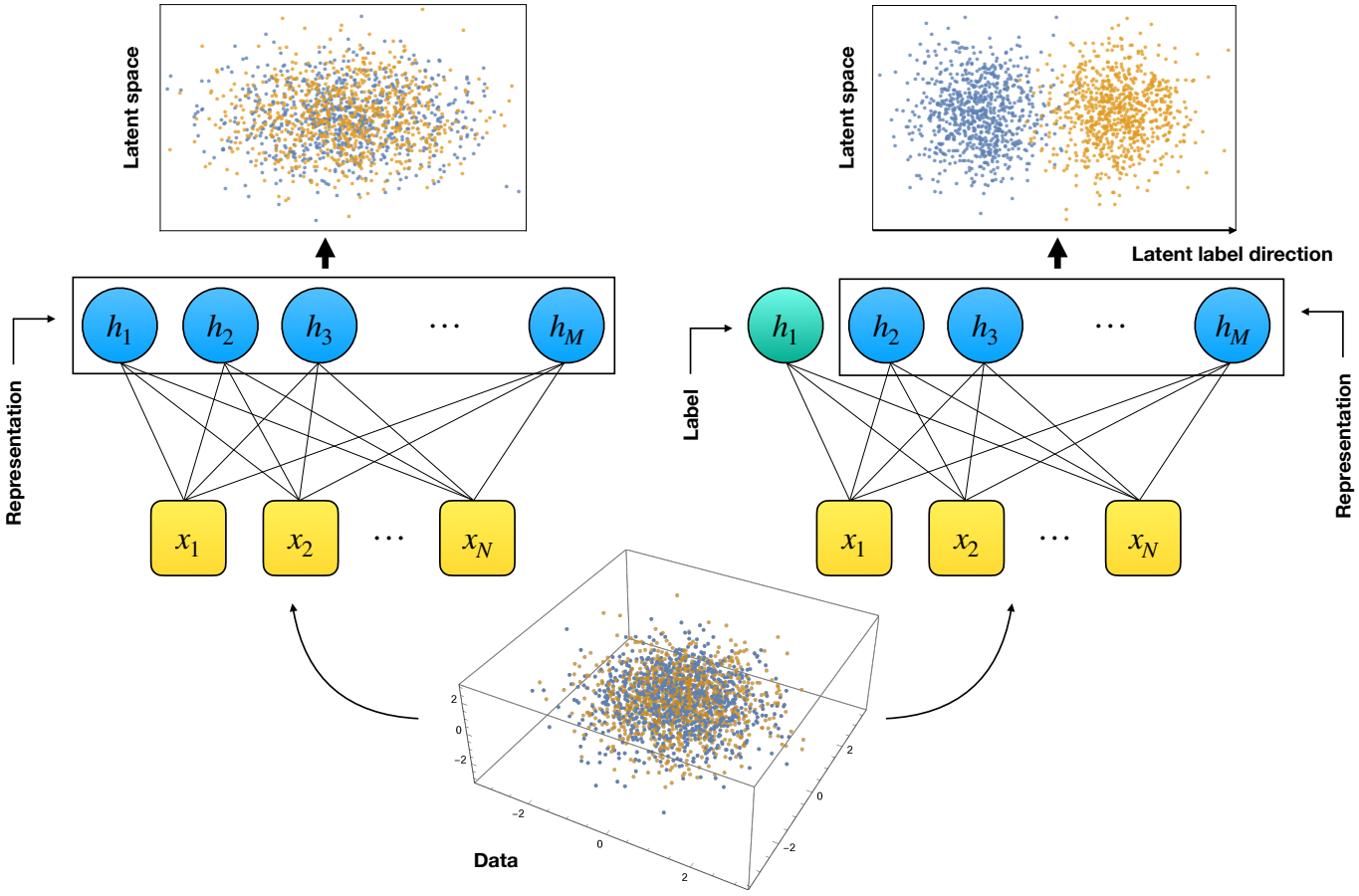

FIG. 24. Schematic representation of the disentangled representation problem. In the last row, we have a population of data points (splitted in two superposed classes) in a high-dimensional space. Training a RBM in a standard way (left side) would result in a lower-dimensional representation in which the data population are merged. Conversely, forcing one or few hidden neurons to receive information on the labels while the others are constrained only to form a representation of each class, the data clouds can be disentangled, and the label-dedicated hidden neurons can be manipulated for generating new examples of a single class.

unsupervised learning method, is doing just that. More sophisticated approaches based on variational auto-encoders favor the factorization of the distribution of latent variables through the so-called $\beta-$VAE model [7]. Another strategy is based on the use of adversarial training, see [27] for an example. Briefly speaking, a discriminator aims at removing the information about a label (binary valued for the case of the two classes in Figure 24) from the latent representation. This approach is conceptually very appealing, but suffers from the well-known difficulties in adversarial training.

Recently, the authors of [17] proposed a simple framework, equivalent to adversarial learning for limited classes of discriminators. This approach consists in learning RBM as usual, but with additional constraints on the weight vectors $\boldsymbol{W}^{\mu}$ to impede hidden units to capture one or more directions in the data space crucial to determine the label values. This frameworks allows to localize the information about the label on one (or few) hidden units of the RBM, and therefore to generate new data with desired label values. It can also be used to generate new data with ambiguous labels. Applications to human faces, MNIST, the Ising model and protein taxonomy can be found in [17]. Remarkably, it is also possible to estimate analytically (within some degree of approximation) the cost in terms of log-likelihood of the generated data induced by the disentanglement of the representations.

Deep tempering with RBM

Sampling complex energy landscape is key to statistical and computational physics. Following the introduction of Monte Carlo (MC) methods by Metropolis, several approaches have been considered to speed up sampling. Among them are replica exchange MC [39], also called parallel tempering. Parallel tempering consists in simulating more than one copy of the system, at higher temperatures than the target temperature of interest. These replicated systems are likely to be

easier to sample, especially at very high temperatures for which the effective barriers in the energy landscape are low and easy to cross. The idea is then to allow for exchange of configurations between copies of the system thermalized at different temperatures. Hence, low-temperature systems will benefit for the capability of high-temperature systems to quickly explore the configuration space, avoiding therefore to be indefinitely stuck in the landscape valleys. The procedure requires that the exchange, which must satisfies detailed balance, has a reasonable probability of occurring, which implies that the two temperatures should not be too far away from one another.

From a conceptual point of view, the idea of having a chain of different systems with slowly changing Hamiltonians, which is key to parallel tempering, is more general than the standard implementation in which all these Hamiltonians are identical up to global rescalings encoding the temperatures of the systems. In this context Bengio and collaborators proposed to use RBM to exploit a more general version of parallel tempering, called deep tempering, in which the Hamiltonians are all different, and built from RBMs of different sizes [15].

This approach was recently extended in [37]. Informally speaking, a stack of nested RBMs, using the representations of a RBM as 'data' for the next RBM along the stack, are learned. These RBMs define more and more simplified versions of the true distributions, which become increasingly easier to sample with standard MC dynamics. The RBMs are then coupled through each other, allowing them to exchange their configurations. These exchanges are made possible by the nested structure of the stack, *i.e.* the compatibility between the sizes of the layers of contiguous RBMs. This algorithm can be shown to offer a considerable speed up with respect to standard Gibbs sampling of a single RBM.

Acknowledgments. These are notes from the lecture by R. Monasson given at the summer school "Statistical Physics & Machine Learning", which was organized by F. Krzakala and L. Zdeborova and took place in Les Houches School of Physics in France from 4th to 29th July 2022.

[1] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

[2] Elena Agliari, Adriano Barra, Andrea De Antoni, and Andrea Galluzzi. Parallel retrieval of correlated patterns: From hopfield networks to boltzmann machines. *Neural Networks*, 38:52–63, 2013.

[3] Daniel J Amit and Daniel J Amit. *Modeling brain function: The world of attractor neural networks*. Cambridge university press, 1989.

[4] Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55:1530–1533, Sep 1985.

[5] Adriano Barra, Alberto Bernacchia, Enrica Santucci, and Pierluigi Contucci. On the equivalence of hopfield networks and boltzmann machines. *Neural Networks*, 34:1–9, 2012.

[6] J-P. Bouchaud, M. Mézard, and J. Dalibard. *Complex Systems: Lecture Notes of the Les Houches Summer School 2006*. ISSN. Elsevier Science, 2011. R. Monasson, Chapter 1: Introduction to Phase Transitions in Random Optimization Problems.

[7] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae, 2018.

[8] Tommaso Castellani and Andrea Cavagna. Spin-glass theory for pedestrians. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(05):P05012, may 2005.

[9] Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965.

[10] David S. Dean and Satya N. Majumdar. Large deviations of extreme eigenvalues of random matrices. *Phys. Rev. Lett.*, 97:160201, Oct 2006.

[11] Aurélien Decelle, Giancarlo Fissore, and Cyril Furtlehner. Spectral dynamics of learning in restricted boltzmann machines. *EPL (Europhysics Letters)*, 119(6):60001, 2017.

[12] Aurélien Decelle, Giancarlo Fissore, and Cyril Furtlehner. Thermodynamics of restricted boltzmann machines and related learning dynamics. *Journal of Statistical Physics*, 172(6):1576–1608, 2018.

[13] Aurélien Decelle and Cyril Furtlehner. Restricted boltzmann machine: Recent advances and mean-field theory. *Chinese Physics B*, 30(4):040202, 2021.

[14] Aurélien Decelle and Cyril Furtlehner. Restricted boltzmann machine, recent advances and mean-field theory, 2020.

[15] Guillaume Desjardins, Heng Luo, Aaron Courville, and Yoshua Bengio. Deep Tempering. *arXiv:1410.0123 [cs, stat]*, October 2014. arXiv: 1410.0123.

[16] A. Engel, C. Van den Broeck, and C. Broeck. *Statistical Mechanics of Learning*. Statistical Mechanics of Learning. Cambridge University Press, 2001.

[17] Jorge Fernandez-de Cossio-Diaz, Simona Cocco, and Remi Monasson. Disentangling representations in restricted boltzmann machines without adversaries. *Physical Review X*, 13:021003, 2023.

[18] Asja Fischer and Christian Igel. Training restricted boltzmann machines: An introduction. *Pattern Recognition*, 47(1):25–39, 2014.

[19] Marylou Gabrié, Eric W. Tramel, and Florent Krzakala. Training restricted boltzmann machines via the thouless-anderson-palmer free energy. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 640–648, Cambridge, MA, USA, 2015. MIT Press.

[20] E Gardner and B Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271, jan 1988.

[21] Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

[22] Moshir Harsh, Jérôme Tubiana, Simona Cocco, and Rémi Monasson. 'place-cell' emergence and learning of invariant data with restricted boltzmann machines: breaking and dynamical restoration of continuous symmetries in the weight space. *Journal of Physics A: Mathematical and Theoretical*, 53(17):174002, apr 2020.

[23] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[24] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[25] Haiping Huang and Taro Toyoizumi. Advanced mean-field theory of the restricted boltzmann machine. *Phys. Rev. E*, 91:050101, May 2015.

[26] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.

[27] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes, 2017.

[28] Satya N. Majumdar and Massimo Vergassola. Large deviations of the maximum eigenvalue for wishart and gaussian random matrices. *Phys. Rev. Lett.*, 102:060601, Feb 2009.

[29] Olivier C. Martin, Rémi Monasson, and Riccardo Zecchina. Statistical mechanics methods and phase transitions in optimization problems. *Theoretical Computer Science*, 265(1):3–67, 2001.

[30] M. Mézard, G. Parisi, and M. Virasoro. *Spin Glass Theory and Beyond*. World Scientific Lecture Notes in Physics: Volume 9, 1986.

[31] Radford M Neal. *Probabilistic inference using Markov chain Monte Carlo methods*. Department of Computer Science, University of Toronto Toronto, ON, Canada, 1993.

[32] Notice that the fluctuations of λ_1 around the right edge of the semi-circle, 2σ, are much larger than the distance to μ represented by the gap. These fluctuations of λ_1 indeed obey the Tracy-Widom distribution, and are of the order of $N^{-1/6}$.

[33] Note that $G(1)$ simply coincides with the full volume of solution in Eq. (111).

[34] Notice also that the Gaussian-Gaussian RBM case, the learning dynamics can be written exactly, and it emerges that is achieved SVD of the dataset by keeping modes above a given threshold [13].

[35] Giorgio Parisi and Tommaso Rizzo. Universality and deviations in disordered systems. *Physical Review B*, 81, 01 2009.

[36] P Reimann, C Van den Broeck, and G J Bex. A gaussian scenario for unsupervised learning. *Journal of Physics A: Mathematical and General*, 29(13):3521, jul 1996.

[37] Clément Roussel, Jorge Fernandez-De-Cossio-Diaz, Simona Cocco, and Rémi Monasson. Deep Tempering with Nested Restricted Boltzmann Machines, January 2023. HAL preprint hal-03919483.

[38] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science, 1986.

[39] Robert H. Swendsen and Jian-Sheng Wang. Replica monte carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57:2607–2609, Nov 1986.

[40] Chako Takahashi and Muneki Yasuda. Mean-field inference in gaussian restricted boltzmann machine. *Journal of the Physical Society of Japan*, 85(3):034001, 2016.

[41] Michel Talagrand. The parisi formula. *Annals of Mathematics*, 163(1):221–263, 2006.

[42] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1064–1071, New York, NY, USA, 2008. Association for Computing Machinery.

[43] J. Tubiana and R. Monasson. Emergence of compositional representations in restricted boltzmann machines. *Phys. Rev. Lett.*, 118:138301, Mar 2017.

[44] Jérôme Tubiana, Simona Cocco, and Rémi Monasson. Learning compositional representations of interacting systems with restricted boltzmann machines: Comparative study of lattice proteins. *Neural Comput.*, 31(8):1671–1717, aug 2019.

[45] Jérôme Tubiana, Simona Cocco, and Rémi Monasson. Learning protein constitutive motifs from sequence data. *eLife*, 8:e39397, mar 2019.

[46] T L H Watkin and J P Nadal. Optimal unsupervised learning. *Journal of Physics A: Mathematical and General*, 27(6):1899, mar 1994.