

# Spin glass theory and its new challenge: structured disorder

M Mézard\* 

Bocconi University, Milan, Italy

Received: 04 September 2023 / Accepted: 06 November 2023

**Abstract:** This paper first describes, from a high-level viewpoint, the main challenges that had to be solved in order to develop a theory of spin glasses in the last fifty years. It then explains how important inference problems, notably those occurring in machine learning, can be formulated as problems in statistical physics of disordered systems. However, the main questions that we face in the analysis of deep networks require to develop a new chapter of spin glass theory, which will address the challenge of structured data.

## Spin glasses

Statistical physics is more or less one and a half century old. Its creation was based on renouncing to follow the trajectories of single particles and moving rather to a coarser, statistical description of systems with many interacting particles. This radical move allowed to handle the specific effects that emerge when the number of particles becomes large, as summarized in Phil Anderson's famous paper "More is different" [1]. One of its great achievements is the understanding and analysis of phase transitions, and the discovery of universality classes at second-order phase transitions, where the divergence of the correlation length wipes out many of the microscopic details of the particles.

About fifty years ago, statistical physics developed a new research direction, the one of strongly disordered systems. An important building piece of its construction is the theory of spin glasses, magnetic systems with disordered interactions. In this section, we shall mention some of the formidable challenges that had to be solved in order to develop a theory of spin glasses, keeping to the case of classical systems (parallel developments in the field of quantum statistical physics deserve a separate presentation).

As is well known, magnetism has played an important role in the development of statistical physics. The solution by Onsager of a "simple" model of ferromagnet, the Ising model in two dimensions, was crucial in establishing the

concept of spontaneous symmetry breaking. And the understanding of Ising models in  $d$ -dimensions, including non-integer values of  $d$ , also played a major role in the development of the renormalization group.

A few well-known landmarks from the ferromagnetic Ising model

It is useful to set the stage and prepare the discussion of spin glasses, starting with a very short sketch of the well-known case of ferromagnetism, which can be found in any standard book of statistical physics. In the Ising model, two-state spins  $s_i = \pm 1$  located on the  $N = L^d$  vertices of a hypercubic  $d$ -dimensional lattice interact through pair interactions, with an interaction energy which is a sum over pairs of adjacent spins

$$E(s) = -J \sum_{(ij)} s_i s_j \quad (1)$$

where  $J > 0$  is the ferromagnetic coupling constant.

The order parameter is the magnetization density

$$M = \frac{1}{N} \sum_i \langle s_i \rangle, \quad (2)$$

where  $\langle s_i \rangle$  is the expectation of the spin  $s_i$  with respect to the Boltzmann measure  $P(s) = (1/Z)e^{-\beta E(s)}$ . This measure is even under the simultaneous flipping of all the spins  $s_i \rightarrow -s_i$ ; therefore, for a fixed  $N$  one has  $M = 0$  at any inverse temperature  $\beta$ . On the other hand, if one adds a small symmetry breaking term to the energy,  $E(s) \rightarrow E(s) - B \sum_i s_i$ , then

\*Corresponding author, E-mail: marc.mezard@unibocconi.it

$$\lim_{B \rightarrow 0^+} \lim_{N \rightarrow \infty} M = \pm M^* , \quad (3)$$

with  $M^*(\beta) > 0$  in the low-temperature phase  $\beta > \beta_c$ , where the inverse critical temperature  $\beta_c$  is finite for  $d \geq 2$ . This is the ferromagnetic phase transition, associated with the phenomenon of spontaneous symmetry breaking.

In order to get a first qualitative understanding of this phase transition, one can use the mean-field approximation. Starting from the exact relation

$$\langle s_i \rangle = \left\langle \tanh \beta \left( B + J \sum_{j \in D_i} s_j \right) \right\rangle \quad (4)$$

where  $D_i$  is the set of neighbors of spin  $i$ , one neglects fluctuations, substituting the expectation value of the tanh by the tanh of the expectation value (this is the mean field). Seeking a homogeneous solution  $\langle s_i \rangle = M$  (which is correct far from the boundaries, or using periodic boundary conditions), one finds

$$M = \tanh \beta (B + zJM) \quad (5)$$

where  $z = |D_i|$  is the number of neighbors of each spin. This equation predicts a ferromagnetic phase transition when  $B \rightarrow 0$ , with an inverse critical temperature  $\beta_c = 1/(zJ)$ .

The mean-field approximation is better in larger dimensions; it actually becomes exact when  $d \rightarrow \infty$ , while it wrongly predicts the existence of a phase transition when  $d = 1$ . A popular model where the mean-field approximation becomes exact is the Curie–Weiss model, where all the pairs of spins interact with a rescaled coupling  $J = \tilde{J}/N$ . In this model, the mean-field equation  $M = \tanh \beta (B + \tilde{J}M)$  is exact and the phase transition takes place at  $\beta_c = 1/\tilde{J}$ .

### Spin glasses

A simple model for spin glasses is the Edwards–Anderson model [12]. This has the same ingredients as the Ising model, except that the coupling constant between two spins  $i, j$  depends on the pair. The energy becomes

$$E(s) = - \sum_{(ij)} J_{ij} s_i s_j . \quad (6)$$

Depending on the pair  $(ij)$ , the coupling constant can be ferromagnetic ( $J_{ij} > 0$ , favoring the alignment of spins at low temperatures), or antiferromagnetic ( $J_{ij} < 0$ , favoring spins pointing in opposite directions at low temperatures).

With respect to the ferromagnetic case, this modification is crucial and poses a number of remarkable challenges that had to be solved in order to elaborate a theory of spin glasses. This elaboration is an outstanding achievement which culminated in the solution by Parisi of the mean-

field Sherrington–Kirkpatrick model [48] (Parisi’s Nobel lecture [44] gives a nice summary, and the recent book [7] gives an idea of the applications that it has had in several branches of science).

In this paper, I shall not enter any detail of spin glass theory, but adopt a high-level point of view, trying to point out the four most important challenges.

### First challenge: ensembles of samples

The first challenge that can be identified is the characterization of a spin glass sample. In order to define the energy, and therefore the Boltzmann probability, one needs to know all the coupling constants  $\mathcal{J} = \{J_{ij}\}_{1 \leq i < j \leq N}$ . If the interactions are short range, this is a number of parameters which grow proportionally to the size of the system  $N$ . This raises two problems. On the one hand, for macroscopic systems it is impossible to even write the energy function: the description of a given sample requires to know a number of parameters of the order of the Avogadro number. On the other hand, for each new sample characterized by these couplings  $\mathcal{J}$ , there is a new Boltzmann probability

$$P_{\mathcal{J}}(s) = \frac{1}{Z_{\mathcal{J}}} e^{\beta \sum_{(ij)} J_{ij} s_i s_j} , \quad (7)$$

where  $Z_{\mathcal{J}}$ , called the partition function, is a sample-dependent normalization constant which ensures that the total probability is normalized to one.

In a step that mimics the one which was taken when statistical physics was first introduced, this double problem was solved by introducing a second level of probability, namely a probability distribution in the space of samples. The couplings  $\mathcal{J}$  are supposed to be generated from a probability distribution  $\mathcal{P}(\mathcal{J})$ . A given realization of  $\mathcal{J}$  is a sample. For instance, in the Edwards–Anderson model [12] one assumes that for each pair  $(ij)$  of neighboring spins we draw  $J_{ij}$  independently at random, from a distribution with probability density  $\rho$ . In the SK model each of the  $N(N-1)/2$  couplings  $J_{ij}$  is drawn at random from a normal distribution with mean 0 and variance  $1/N$ .

We have now two levels of probability. The first one draws a sample  $\mathcal{J}$  generated from the probability  $\mathcal{P}(\mathcal{J})$ . Then, one studies the Boltzmann law  $P_{\mathcal{J}}(s)$  for this sample. The averages of spin configurations with respect to  $P_{\mathcal{J}}(s)$  are called thermal averages, while the averages over samples, with respect to  $\mathcal{P}(\mathcal{J})$ , are called quenched averages. I’ll call  $\mathcal{P}(\mathcal{J})$  the quenched probability, to distinguish it from Boltzmann’s probability.

Then one is led to make a distinction between two types of properties.

On the one hand, there are properties which depend on the sample. For instance, the ground state configuration of

spins, the one which minimizes the energy, obviously depends on  $\mathcal{J}$ . Actually, all the details of the energy landscape depend on  $\mathcal{J}$ .

On the other hand, some properties turn out to be 'self-averaging', meaning that they are the same, for almost all samples (with a quenched probability that goes to one in the large  $N$  limit). For instance, in the EA or SK model the internal energy density

$$U_{\mathcal{J}} = \frac{1}{N} \sum_s P_{\mathcal{J}}(s) E_{\mathcal{J}}(s) \quad (8)$$

is self-averaging (this is easily proven in EA because one can cut a sample into many pieces and neglect the interactions between pieces which are of relative order surface to volume; the proof is less easy for models in the SK family [21]). This means that the distribution of  $U_{\mathcal{J}}$  (when one picks a sample at random from the quenched probability) has a probability density that concentrates, when  $N \rightarrow \infty$ , around a given value  $u$  that depends only on the inverse temperature  $\beta$  and on the statistical properties of the distribution of  $\mathcal{J}$ . The typical sample-to-sample fluctuations of  $U_{\mathcal{J}}$  around this value are of order  $1/\sqrt{N}$ . In the limit  $\beta \rightarrow \infty$ , this also implies that the ground state energy density is self-averaging. The same is true for all the extensive thermodynamic properties. For instance, the magnetization density in the presence of a magnetic field, or its linear dependence at small fields, the magnetic susceptibility, are self-averaging. This property of self-averageness is crucial: it is the reason why the measurements of magnetic susceptibilities or specific heat of two distinct spin glass samples with the same statistical properties (take, for instance, two samples of CuMn with 1% of Mn) give the same result: these are reproducible measurements because the measured property is self-averaging.

Notice that, for the properties which are not self-averaging, one can study their quenched distribution. A typical example is the order parameter function that we shall discuss below [37].

### Second challenge: inhomogeneity

The second challenge that spin glass theory had to face is inhomogeneity. The lesson we learn from detailed studies of the SK model is the following. For a typical sample  $\mathcal{J}$  there exists a low-temperature 'spin glass' phase in which the spins develop nonzero local magnetizations:

$$\langle s_i \rangle = m_i \quad (9)$$

Because of the disorder in the coupling constants  $J_{ij}$ , contrarily to the ferromagnetic case these magnetizations are not uniform. Analyzing a spin glass order in detail thus requires to use as order parameter the set of all the

magnetizations. This is a  $N$ -component order parameter. Thouless Anderson and Palmer were able to write a closed system of  $N$  equations that relate all these components [54]. The TAP equations, which generalize (5) to the spin glass case, are:

$$m_i = \tanh \left[ \beta \left( \sum_j J_{ij} m_j - \beta(1-q)m_i \right) \right] \quad (10)$$

where  $q = (1/N) \sum_j m_j^2$ . With respect to the naive mean-field equations  $m_i = \tanh \left[ \beta \sum_j J_{ij} m_j \right]$ , they are characterized by the appearance of the "Onsager reaction term". This basically says that, when one computes the mean of the local magnetic field on site  $i$ , one should subtract from the naive estimate  $\sum_j J_{ij} m_j$  the part of  $m_j$  which is polarized by  $i$  itself. This means using a "cavity" magnetization  $m_j^c = m_j - \chi_j J_{ji} m_i$  where  $\chi_j = \beta(1 - m_j^2)$  is the local magnetic susceptibility of an Ising spin.

When  $N$  is not too large, say a few tens of thousands, TAP-like equations can be used as an algorithm, and they can be solved by iteration using a specific iteration schedule that was found by Bolthausen [6]. This gives information on the behavior of a given sample  $\mathcal{J}$ .

On the other hand, when  $N$  is very large, for instance of the order of the Avogadro number, one cannot write explicitly or solve the TAP equations. One must use a statistical study of the properties of these solutions. It turns out that this cannot be done directly on the TAP equations themselves, because the Onsager reaction term creates subtle correlations. The cavity method [38, 39] allows to circumvent this problem, by first analyzing the statistics of the cavity field, the field acting on a spin in absence of this spin. This allows to build a full solution to the problem.

### Third challenge: the many-valleys landscape

Keeping to the SK model, it was found that there actually exist many different 'states' where the system can freeze, and therefore many solutions of the TAP equations. Each state  $\alpha$  is characterized by  $N$  magnetizations  $m_i^\alpha$ , so the order parameter is actually a  $N$ -component vector. This generalizes the situation of the ferromagnet. Instead of two states, identified by their average magnetization, we have many states. In each of them the average magnetization in the absence of external field,  $(1/N) \sum_i m_i^\alpha$ , vanishes in the thermodynamic limit.

Defining these states correctly is actually difficult. If one parallels the construction of the two pure states that we introduced in (3) for the ferromagnet, the natural generalization is to introduce for each state  $\alpha$  a site-dependent small magnetic field  $B_i^\alpha$  and take the limit where all these

local fields go to zero after the thermodynamic limit. This leads to

$$m_i^\alpha = \lim_{B^z \rightarrow 0} \lim_{N \rightarrow \infty} \langle s_i \rangle_{B^z} \quad (11)$$

The weakness of this definition is that we do not know how to choose the local orientations of  $B_i^z$ : on which site should they be positive and on which site should they be negative? Solving this problem requires knowing the signs of  $m_i^z$ . So, while this definition of the order parameter is interesting, in practice it is useless.

The replica method which was used to solve the SK model [38, 42] actually has an interesting interpretation from this point of view. The idea is that, if we do not know the preferred orientations where the spins will polarize, the systems knows them. So, for theoretical understanding, one can introduce, for a given sample  $\mathcal{J}$ , two replicas of spins,  $s$  and  $\sigma$ , with the same energy function  $E_{\mathcal{J}}$ . In this system, the probability of the two configurations is

$$P_{\mathcal{J}}(s, \sigma) = \frac{1}{Z_{\mathcal{J}}^2} e^{-\beta(E_{\mathcal{J}}(s) + E_{\mathcal{J}}(\sigma))} \quad (12)$$

One can introduce the overlap  $q = (1/N) \sum_i s_i \sigma_i$ , and ask what is the distribution  $P_{\mathcal{J}}(q)$  of this overlap, in the thermodynamic limit  $N \rightarrow \infty$ . In the high-temperature paramagnetic phase, one finds  $P_{\mathcal{J}}(q) = \delta(q - q_0)$  where  $q_0 = 0$  in the absence of an external field, but it becomes  $q_0 > 0$  in the presence of a uniform field. In the spin glass phase,  $P_{\mathcal{J}}(q)$  becomes non-trivial, it has a support  $[q_0, q_1]$  with  $q_1 > q_0$ , and it fluctuates from sample to sample. So, this is a non-self-averaging quantity [37]. Its quenched average,  $P(q) = \int d\mathcal{J} \mathcal{P}(\mathcal{J}) P_{\mathcal{J}}(q)$ , is the order parameter for the spin glass phase. It is this order parameter which appears naturally and is computed in the replica method with replica symmetry breaking, as shown in Parisi's seminal work [43].

A simple way to define the existence of a spin glass phase using two replicas is to introduce a small coupling between them. The energy of a pair of configurations  $s, \sigma$  now becomes:

$$E_{\mathcal{J}}^\epsilon(s, \sigma) = E_{\mathcal{J}}(s) + E_{\mathcal{J}}(\sigma) - \epsilon \sum_i s_i \sigma_i \quad (13)$$

Sampling the pairs of configurations with the corresponding Boltzmann weight, one can compute the expectation value of the overlap  $\langle q \rangle^\epsilon = \int dq P_{\mathcal{J}}^\epsilon(q) q$ . Taking the limit  $\epsilon \rightarrow 0^\pm$  after the thermodynamic limit, one finds

$$q_1 = \lim_{\epsilon \rightarrow 0^+} \lim_{N \rightarrow \infty} \langle q \rangle^\epsilon ; \quad q_0 = \lim_{\epsilon \rightarrow 0^-} \lim_{N \rightarrow \infty} \langle q \rangle^\epsilon \quad (14)$$

These are the two limits of the support of  $P(q)$ . The existence of a spin glass phase is signaled by  $q_1 > q_0$ . This definition gives a very intuitive interpretation to the use of

replicas: one takes two replicas coupled by a small attractive interaction ( $\epsilon > 0$ ). When this interaction vanishes, if the spins in each of the two replicas remain correlated, this signals the spin glass phase. This criterion can also be used in glassy systems without disorder, like structural glasses [15, 36].

The whole ‘‘landscape structure’’ of the spin glass phase can be analyzed as follows: in a given sample there exist many pure states  $\alpha$ . Each of them is characterized by the  $N$ -dimensional vector of magnetizations  $m^\alpha = \{m_1^\alpha, \dots, m_N^\alpha\}$ , and its free energy  $F^\alpha$ . All the states that contribute to the thermodynamics have the same free energy density  $\lim_{N \rightarrow \infty} F^\alpha/N$ , but they have finite free energy differences  $F^\alpha - F^\gamma = O(1)$ , and therefore, each state contributes to the Boltzmann measure with a weight  $P^\alpha$ . Therefore,

$$P(q) = \int d\mathcal{J} \mathcal{P}(\mathcal{J}) \sum_{\alpha, \gamma} P^\alpha P^\gamma \delta(q - q^{\alpha\gamma}) \quad (15)$$

Various types of glassy phases are characterized by different types of  $P(q)$  functions, two extremes being the simple ‘‘one-step RSB’’ characteristic of the structural glass transition ones having only two  $\delta$  peaks at  $q_0$  and  $q_1$  [9, 11, 20], and the ‘‘full-RSB’’ which occurs in the SK model and where the support of  $P(q)$  is the full interval  $[q_0, q_1]$ , with an infinity of states organized in a hierarchical structure called ultrametric [37], and a  $\delta$  peak of  $P(q)$  at the Edwards–Anderson order parameter  $q = q_1$ . This order parameter characterizes the size of the states in the sense that two randomly chosen configurations within the same state will have overlap  $q_1$ .

Fourth challenge: out of equilibrium dynamics

The last big challenge that spin glass theory had to face was the one of equilibrium. The whole description that I gave so far is based on the idea that a given sample of a spin glass can be characterized by the Boltzmann measure  $P_{\mathcal{J}}(s)$ . However, this is true only in the case where the system reaches equilibrium. However, experiments precisely teach us that equilibrium is not reached in the spin glass phase. For instance, measuring the magnetic susceptibility by first cooling the system to a temperature  $T < T_{\text{sg}}$  and then adding a small uniform magnetic field  $B$  gives a ‘‘zero-field-cooled susceptibility’’  $\chi_{\text{ZFC}}$  which is different from the one found by placing the sample in the magnetic field  $B$  at high temperature (above the spin glass transition  $T_{\text{sg}}$ ), and then cooling it to  $T$ . This last procedure gives a ‘‘field-cooled’’ susceptibility  $\chi_{\text{FC}}$  which is in general larger than  $\chi_{\text{ZFC}}$ . In both cases the measurement of the susceptibility is done at the same point  $T, B$  of the phase diagram, but the results differ, proving that the spin glass is

out of equilibrium. Then a very legitimate question is: how can the equilibrium theory be of any use ?

One of the first successes of the Parisi theory has been to give a qualitative explanation of this difference by assuming that the FC susceptibility corresponds to the reaction of the system when perturbing an equilibrium which is a superposition of pure states, while the ZFC susceptibility corresponds to a perturbation within one pure state (see, e.g., [44]). In fact, if one introduces a constrained perturbation to a SK spin glass, in which the system reacts to a small magnetic field, but it is constrained to remain at an overlap larger than  $q$  from its initial state, then the corresponding susceptibility is

$$\chi(q) = \beta \int_q^1 dq' P(q')(1 - q') \quad (16)$$

which gives in the two limiting cases:

$$\chi_{ZFC} = \beta \left[ 1 - \int_0^1 dq' P(q')q' \right] ; \quad \chi_{FC} = \beta[1 - q_1] \quad (17)$$

One can also go beyond and try to study directly the dynamics of mean-field models like the SK model. In the spin glass phase, the time to reach equilibrium diverges in the thermodynamic limit. One can then study what happens on various diverging timescales, as in the first works of Sompolinsky and Zippelius [50].

An alternative approach which gives very interesting insight is to solve the out of equilibrium dynamics, as was proposed initially by Cugliandolo and Kurchan [10]. Focusing again on the mean-field models, one can derive a closed set of equations for the two-time correlation  $C(t_w + t, t_w) = (1/N) \sum_i \langle s_i(t_w + t) s_i(t_w) \rangle$  and the two-time response function  $R(t_w + t, t_w)$ , which is the linear response measured at time  $t_w + t$  of a system which has started its dynamics at time 0, and to which a small magnetic field has been added at the time  $t_w$ . In systems which reach their equilibrium, after a long waiting time  $t_w$ , the functions  $C$  and  $R$  become time-translation invariant, i.e., they depend only on the measurement time  $t$ . This invariance is broken in the spin glass phase: the  $t$  dependence of these two functions depend on the age  $t_w$  of the system, and they keep evolving when  $t_w$  increases, a phenomenon called aging which is often observed in glassy systems. The simplest scenario of aging would be one in which  $C$  and  $R$  become function of  $t/t_w^a$ . For instance, approximate  $t/t_w$  scaling with  $a = 1$  is often observed. In link with its hierarchical static structure, the SK model shows a more complicated behavior, with various timescales characterized by distinct exponents  $a$  playing a role. In link with the aging phenomenon, one also finds a modification of the standard fluctuation dissipation theorem (FDT).

In an equilibrium system, at large enough  $t_w$ , the standard FDT relation between fluctuation and response is  $R(t) = \beta(C(0) - C(t))$ . (Notice that we use here an integrated response function, as defined above.)

In the spin glass phase, this is modified and becomes a relation between  $C(t_w, t)$  and  $R(t_w, t)$  that holds when both the waiting time  $t_w$  and the measurement time  $t$  are large:

$$\frac{\partial R(t_w, t)}{\partial t} = -\beta X(C(t_w, t)) \frac{\partial C(t_w, t)}{\partial t} \quad (18)$$

The function  $X(C)$  is the ‘‘fluctuation dissipation ratio’’. When computed from spin glass theory, one finds that it is equal to the total probability of an overlap larger than  $C$ :

$$X(C) = \int_C^1 P(q) dq \quad (19)$$

It can thus be measured by plotting parametrically  $R$  versus  $C$ . We have thus a way to measure the equilibrium order parameter  $P(q)$  from an out of equilibrium measurement of correlation and response. This was done by [24], and the reader can find a discussion in [44].

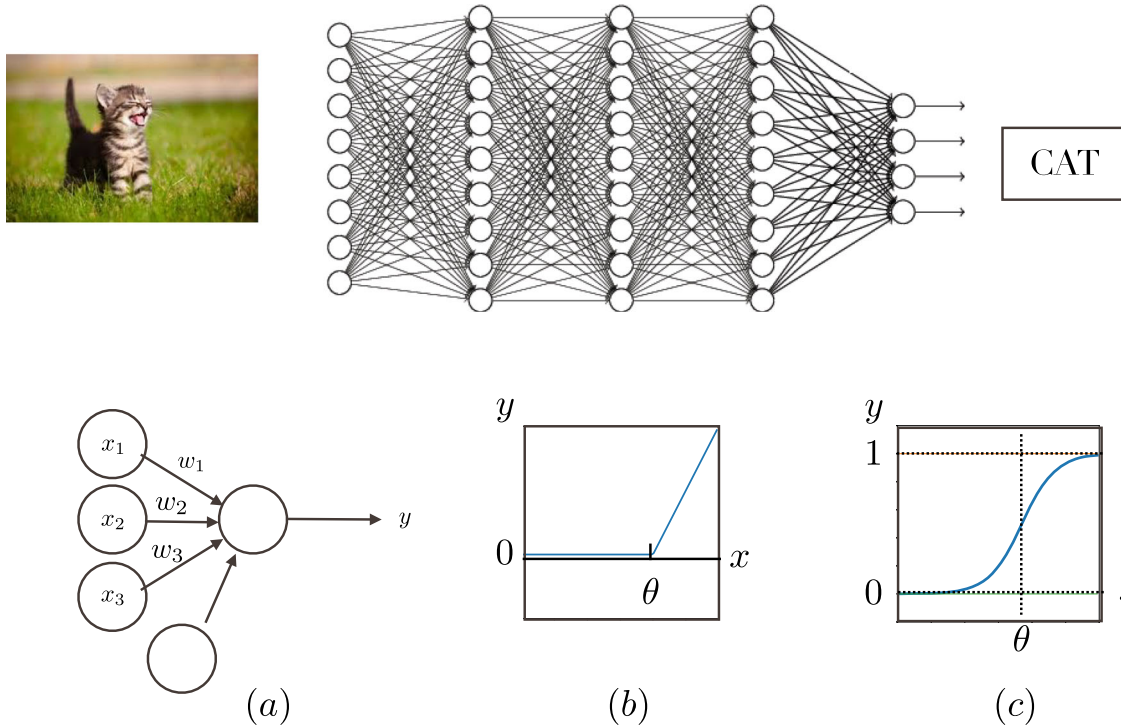
## Statistical physics of inference

### Machine learning as a statistical physics problem

Spectacular recent developments of artificial intelligence are based on machine learning. I’ll sketch here the formal framework of supervised learning, in order to relate it to statistical physics of disordered systems. Recent introductions can be found in [28, 58].

Machine learning aims at learning a function from a  $d$ -dimensional input  $\xi \in \mathbb{R}^d$  to a  $k$ -dimensional output  $y$ . Usually, one is interested in large-dimensional input like an image, so  $d$  is large, in practice it can be  $10^6$  or more, and a small-dimensional output. Taking the famous example of handwritten digits, the image could be an image of a digit, and the output would be the digit. In a one-hot encoding, one would use  $k = 10$  and the digit  $r$  would be associated to  $y_r = 1$  and  $y_{r'} = 0$  for  $r' \neq r$ . One thus wants to learn a target function  $y = f_i(\xi)$ . Actually, in practical applications we do not have a full definition of the function, but we have examples, in the form of a database of pairs input–output  $\mathcal{D} = \{\xi^\mu, y^\mu\}$ , with  $\mu \in \{1, \dots, P\}$ .

Modern deep networks are based on artificial neurons organized in layers. Each neuron in layer  $L$  is a simple unit that receives a signal from the neurons in the previous layer, applies a nonlinear function and sends this processed signal to the neurons of the next layer (see Fig. 1). The activity of neuron  $i$  in layer  $r$  is given by



**Fig. 1** Top: typical structure of a feedforward neural network. The input (image of a cat) is presented on the left layer. Data is processed, layer after layer, until the output is given in the right. Bottom: (a) Each layer is built from artificial neurons. They receive inputs from the neurons of the previous layer on their left hand side, these

inputs are weighted, and the linear combination of the reweighted inputs is then transformed by a nonlinear transfer function. (b, c) Two examples of such functions are shown here: the ReLu function (b) and a sigmoid (c)

$$x_i^r = \psi_i^r \left( \sum_j W_{ij}^r x_j^{r-1} \right) \quad (20)$$

where  $W^r$  is a matrix of the “synaptic efficacies” between neurons in layer  $r - 1$  and  $r$ . The nonlinear function  $\psi_i^r$  can be, for instance, a sigmoid or a rectified linear unit. Usually, it depends on the layer  $r$  but not on the precise neuron in the layer.

The layer 0 is the input,  $x^0 = \zeta$ , and the layer  $L$  is the output,  $x^L = y$ ; other layers are called hidden layers. A given realization of the neural network is given by its architecture (the depth  $L$  and the width of each layer), the choice of nonlinear functions, and the values of the weights  $W = \{W^r\}$ ,  $r \in \{1, \dots, L\}$ . We shall denote by  $N$  the total number of weights. If the width of each hidden layers is constant equal to  $h$ , then  $N = dh + (L - 2)h^2 + hk$ . A given network with parameters  $W$  implements a function from input to output  $y = f(W, \zeta)$ .

In most applications the architecture is chosen by the engineer, based on previous experience, but the weights are learnt. Indeed, machine learning designates the process by which the parameters (in this case the weights) are not given to the machine, but the machine learns them from data. This learning is called the training phase. In order to

train a network, one defines a “loss function”  $L(W)$  which measures the errors made by the machine with parameters  $W$  on the database. For instance, one could use a quadratic loss

$$L_D(W) = \sum_{\mu=1}^P (y^\mu - f(W, \zeta^\mu))^2 \quad (21)$$

but many other choices are possible. Then the training phase consists in finding the  $W$  that minimizes the loss. In practice, people use a form of gradient descent called stochastic gradient descent, in which one moves in the  $W$  landscape using iteratively noisy versions of the gradient computed from partial sums involving some batches of  $\mu$  indices.

Once the learning has been done, one can use the couplings  $W^*$  which have been found during training, and see how well the network generalize when it is presented some new data that it has never seen. The test—or generalization—loss has the same expression as (21), but with new, previously unseen, input–output pairs.

## Data as disorder

One can also introduce a probability distribution in the space of weights  $W$ , of the form

$$P_{\mathcal{D}}(W) = \frac{1}{Z_{\mathcal{D}}} P_0(W) e^{-\beta L_{\mathcal{D}}(W)} \quad (22)$$

where  $Z_{\mathcal{D}}$  is a normalization constant, and  $P_0(W)$  is a prior on the weights; one can choose it as a factorized prior  $P_0(W) = \prod_{r,i,j} \rho(W_{ij}^r)$  (for instance, one can use for  $\rho$  a Gaussian if one wants to avoid too large weights). One could also normalize  $\sum_j (W_{ij}^r)^2 = 1$ , but I will keep here for simplicity to the factorized case. The parameter  $\beta$  is an auxiliary inverse temperature parameter. When  $\beta$  is large, this probability distribution is concentrated on the sets of weights  $W$  which minimize the loss. This formalism amounts to an approach of statistical physics in the space of weights. It was pioneered by Elizabeth Gardner [16, 17] in the study of the simplest network, the perceptron which has no hidden unit (and is therefore limited to linearly separable tasks).

In recent applications of deep networks like the large language model Chat-GPT, the total number of weights can be of order  $10^{11}$ , and the total number of operations used in order to train a network on extremely large databases of basically all available text is easy to remember; it is of the order of  $10^{24}$ , a “mole of operations.” So, the distribution (22) is a measure in a large  $N$ -dimensional space. The elementary variables, the weights, are real-valued variables with a measure  $\rho$ . They are coupled through an energy which is the loss  $L_{\mathcal{D}}(W)$ . This energy is a complicated function of the variables  $W$ , and it depends on a large set of parameters, namely all the input–output pairs  $\mathcal{D}$ . Therefore, the measure (22) has all the ingredients of a statistical physics systems with a quenched disorder which is the data.

Having in mind our discussion of disordered systems, we can immediately identify several questions. One can work on a given data base (a given sample) and ask about the landscape for learning, and for generalization. But clearly, for theoretical studies, one would like to have an ensemble of samples, which is a probability measure in the space of inputs, and for each input the corresponding output. With such a setup of a *data ensemble*, one can draw a database by choosing inputs independently at random from the ensemble, one can study the property of self-averageness (which properties of the optimal network  $W^*$  are dependent on the precise realization of the database, and which ones are not—in the large  $N$  limit ?). The test loss becomes easy to define: it is the expectation value of the loss, over pairs of input–output generated from the data ensemble.

Working with a single dataset and with a data ensemble are two rather different approaches. In many practical applications the engineer’s approach is to use a single dataset, and the definition of an ensemble is not obvious. For instance, if one wants to identify if there is a cat or a dog on an image, so far what is done is use huge databases of images of cats and dogs, randomly choosing part of them for training, and another part for the test phase. However, in such a single database setup it is not easy to develop a theory: on the one hand, there is a risk of developing a theory which is too much tailored to this precise database, and from which one cannot draw general conclusions; also, one cannot use probabilities to compute the generalization error. So, the use of data ensembles is clearly welcome from a theoretical point of view, but then one faces the difficulty of defining the ensemble in such a way that it will include some essential features of real databases, but it should be smooth enough so that one can interpolate through it reasonably (the ensemble which would use a probability law that is a sum of  $\delta$  peaks on each point of a database is useless), and simple enough that it can be studied. So, the question of finding good ensembles is a fundamental question of how to model the “world”, i.e., the set of all possible inputs that can be presented. Here by modeling one intends it in a physics’ approach, namely being able to identify the key features that should be incorporated into the ensemble, neglecting less important “details”.

Interestingly, this quest for modeling of data meets with an important recent direction of development of machine learning, which are generative models. In parallel to, and in symbiosis with, the supervised machine learning that I have briefly exposed, very significant progress has been made on generative models. These aim at generating data ‘similar to’ a given database. Among the processes that have been explored, one can mention generative adversarial networks (GAN), or the very physical generative diffusion models which take a database, degrade it using a Langevin process until it has been transformed into pure noise, and then reverse the Langevin process to reconstruct artificial data from noise (see [49, 51, 52]); for a recent review see [57], and for a statistical physics perspective: [5]).

## Surprises

*Perceptrons*

Training a neural network in supervised learning amounts to finding the ground state of a strongly disordered system. One can thus ask what properties of spin glasses one can find in neural networks. Early studies on perceptrons have provided important benchmarks. Two main categories of tasks have been studied: learning arbitrary labels, or

learning from a “teacher rule”. In both cases, the database consists of  $P$  independent input datapoints each consisting of  $d$  i.i.d. numbers from a distribution  $\rho_0(\xi)$ . In the case of learning from arbitrary labels, for each input one generates a desired output which is drawn randomly, independently from the input. In this case one studies only the training phase, generalization has no meaning. In the case of learning a teacher rule, one generates the output from a “teacher” set of weights,  $W_t$ , through  $y = f(W_t, \xi)$ . The quality of training can be monitored by computing some distance between  $W^*$  and  $W_t$ , like, for instance,  $|W^* - W_t|^2 / |W_t^2|$ .

The behavior of the training depends a lot on the a priori measure on the weights,  $\rho$ . If  $\rho$  is Gaussian or imposes a spherical constraint, the training problem is convex and the landscape is simple. The training and generalization error decrease continuously with  $\alpha = P/N$ . If  $\rho$  is discrete, corresponding to Ising spins, then the training on random labels shows a replica symmetry breaking phase at  $\alpha_c = .83$  [27] which has a strange nature. On the one hand, the typical configurations at  $\alpha$  close to  $\alpha_c$  are isolated points, building a golf-course potential [26]. On the other hand, atypical, exponentially rare regions of phase space concentrate a large number of neighboring solutions [2], and are easy to find. As far as learning a teacher rule is concerned, with binary synapses, the generalization error shows a first order phase transition to perfect generalization [22] when  $\alpha$  is larger than a threshold  $\alpha_g \simeq 1.25$ . The phase diagram can be studied rigorously, and when  $\alpha > \simeq 1.5$  one can also use iterative message-passing algorithms based on the cavity-TAP method [35] in order to find the optimal weights defined by the teacher [4].

### Deep networks

The recent experimental successes of deep networks have triggered a lot of analyses, but the situation is less clear than in perceptrons. So far, statistical physics approaches can be used efficiently in multilayer networks either when the transfer functions are linear [32], or when there is a single layer of learnable parameters with a size that diverges in the thermodynamic limit, like, for instance, in committee machines or parity machines.

Also, the empirical observations of the learning process show a picture which is rather different from the usual spin glass landscape. The first observation is that very complex functions involving billions of weights are learnable from examples, using the simple stochastic gradient descent algorithm. This means that the loss function which is optimized is not as rough as one would have in a spin glass. Typically, stochastic gradient descent, when initiated from generic initial conditions (with small weights), finds a set

of weights  $W^*$  not far from the initial condition, which has a small loss. Surprisingly, in this large-dimensional space, the set of weights with small loss is not sparse, as one could have expected from our experience of optimizing in large dimensions.

Once the network has been trained, typically using a number  $N$  of weights that is of the same order as the number of points in the database, one must study its generalization properties. From a statistics’ perspective, what has been done in the training phase is fitting a complicated  $N$ -dimensional function using  $P$  datapoints. This is possible because  $N$  is large, but one should expect to be in a regime of overfitting, and therefore a poor generalization. This is not the case. Actually, increasing the depth of the network, and therefore the number  $N$  of fitting parameters, one observes that the generalization errors keeps decreasing, while it should shoot-up in the overfitting regime. Among all these minima of the loss, some generalize better than others, and this seems to be correlated with the flatness of the landscape around the minimum [3].

These facts indicate that deep learning landscapes are rather different from the ones that have been explored in spin glass theory or in perceptrons. What are the ingredients responsible for the relatively easy training and the lack of overfitting in deep networks? Three directions are being explored: (1) the architecture of the networks, and in particular the importance of using deep enough networks, with many layers; in practice the design of the architecture, including the choice of nonlinearities, is an engineer’s decision based on previous experience; (2) the learning algorithm; stochastic gradient descent started from weights with small values seems efficient at finding out first the main pair correlation in the data, then gradually improving [46]; and (3) the structure of data: practical problems deal with highly structured data, whether they are text, image, amino-acid sequences. In the next section I shall focus on this last point, argue about the relevance of structured data and describe the challenge it poses to statistical physics.

### The new challenge of spin glass theory: structured disorder

Data are highly structured, and a major objective is to develop mathematical models for the datasets on which neural networks are trained. Most theoretical results on neural networks do not model the structure of the training data. Statistical learning theory [40, 55] usually provides bounds that hold in the worst case, but are far from describing typical properties seen in experiments. On the other hand, traditional statistical physics approaches use a setup where inputs are either drawn component-wise i.i.d. from some probability distribution, or are Gaussian-



distributed [13, 47]. Labels are either random or given by some random, but fixed function of the input. Despite providing valuable insights, these approaches ignore key structural properties of real-world datasets.

In recent years several aspects of data structure have been explored, and the first ensembles of structured data have started to be developed. The challenge is of course to create ensembles which contain some of the essential structure, but are at the same time simple enough to be analyzed. I will mention here three categories of data properties which are being studied: effective dimensionality, correlations, and combinatorial/hierarchical structure.

### Effective dimension

Let us consider perhaps the simplest canonical problem of supervised machine learning: classifying the handwritten digits in the MNIST database using a neural network [29]. The input patterns are images with  $28 \times 28$  pixels, so *a priori* we work in the high-dimensional space  $\mathbb{R}^{784}$ . However, the inputs that may be interpreted as handwritten digits, and hence constitute the “world” of our problem, span but a lower-dimensional manifold within  $\mathbb{R}^{784}$ . Although this manifold is not easily defined, its dimension can be estimated based on the distance between neighboring points in the dataset [8, 19, 31, 53]. In fact, if we consider  $P$  independent datapoints in a  $D$ -dimensional space, we expect that the distance between nearest neighbors scales like  $P^{-1/D}$ . Analyzing the MNIST data base, one finds the effective dimension to be around  $D \approx 15$ , much smaller than  $N = 784$ . The “perceptual submanifold” associated with each digit also has an effective dimension, ranging from  $\approx 7$  for the digit 1 to  $\approx 13$  for the digit 8 [23]. Therefore, the task of identifying a handwritten digit consists in finding these ten perceptual submanifolds, embedded in the 15-dimensional “world” manifold of handwritten digits. Of course, the problem is that these manifolds are nonlinear, folded, and it is hard to find them (see [14] for algorithmic approaches). The same phenomenon of reduction in effective dimension is found in other datasets. For instance, images in CIFAR10 are defined in dimension  $N = 1024$ , but have an effective dimension  $D \approx 35$ . In most machine learning problems, the effective “world” on which we train our networks has an effective dimension  $D \ll N$  (in fact, a good practice would be to train the networks so that they can identify when they see an input which is far from the world in which they were trained, and refuse to give an answer in such cases).

A simple attempt at including this effective dimensionality in ensemble of data is the “hidden manifold model” [18]. In this model, the seed  $s$  of a datapoint is generated i.i.d. in a  $D$ -dimensional “latent” space, for

instance, from a Gaussian distribution. Then the datapoint components  $\xi_i$  are generated as

$$\xi_i = g \left( \sum_{r=1}^D F_{ir} s_r \right) \quad (23)$$

where  $F_{ir}$  are given and define the model, as well as  $g$  which is a nonlinear function. It turns out that, when the components of  $F$  are well balanced (and in particular if they are generated i.i.d. from a well-behaved distribution), one can generalize the statistical physics studies of the perceptrons or shallow networks to data which has this hidden manifold structure. The reason is that the hidden units actually receive an input which becomes Gaussian-distributed. This “Gaussian equivalence theorem” allows to use the whole traditional spin glass machinery. It also tells that this kind of model has its limitations, as it is equivalent to some type of Gaussian distributed inputs.

Note that the hidden manifold structure of data defined in (23) can receive a different interpretation, where one would like to learn from the latent signal  $s$  in  $D$  dimensions, but one first projects it to a  $N$ -dimensional space of random features which are fixed, and not learnt, a problem which has been studied in detail when the matrix  $F$  is generated from a random matrix ensemble [34] (but Gaussian equivalence holds beyond this, as long as matrix elements are well balanced, like, for instance, in Hadamard transformation).

Actually, the construction of hidden manifolds can be elaborated by using, instead of (23), an iterative construction based on several layers of projections, as done in the GAN approach. In that case, Gaussian equivalence is conjectured to hold, although it has not been proven yet [18].

### Correlations

From the database, one can construct the empirical pair correlation  $C_{ij}$  between two components of the input

$$C_{ij} = \frac{1}{P} \sum_{\mu=1}^P \xi_i^\mu \xi_j^\mu \quad (24)$$

as well as higher-order correlations. (Here we assume that we use centered data, in which the empirical mean of  $\xi_i$  has been subtracted.) A distinguishing property of practical datasets is that correlations are highly structured, and actually some of this structure is already seen at the level of the pair correlation.

For instance, if one diagonalizes the matrix of pair correlations, which is of Wishart type, one finds a spectrum of eigenvalues which differs notably from the Marcenko Pastur one that would be obtained if the components  $\xi_i^\mu$  were distributed independently and identically. Instead,

one typically gets a power-law distribution of the large eigenvalues, and it has been argued that this power-law scaling is actually related to the power-law decay of the loss with respect to either  $N$  or  $P$ , found in large language models [33]. A simple attempt at including this effective dimensionality in an ensemble of data is to use random Wishart matrices with a power-law distributed spectrum [30, 33].

Note that this power-law scaling (with small exponents) of eigenvalues of the correlation matrix points to the existence of some type of long-range correlations. In fact, very structured and long-range correlations in data are very important, and the recently developed "attention mechanism" is precisely built in order to handle such correlations [56]. These are of a type which is rather different from what one is used to in statistical physics. The easiest way to illustrate them is through language models. In these models, one decomposes the sentences into tokens (typically words or—for composite words—portions of words) and the language models are trained from a large corpus, at the task which is to take a text, interrupt it somewhere, and give the best guess of the next token. Clearly the simplest approach would be to sample the conditional probability distribution: take the previous  $k$  tokens before being interrupted, and look in the database at sentences which have exactly this sequence of  $k$  tokens, and compute from this database the most probable next token. This approach was started very early on, by Shannon himself. But clearly it is limited to small values of  $k$ : beyond  $k$  of order a dozen, one does not have the statistics to infer the conditional probability. But it turns out that key tokens, which are crucial for guessing the next one, can be found much earlier in the text. Take, for instance, this sentence written above: "*Instead, one gets typically a power law distribution of the large eigenvalues, and it has been argued that this power-law scaling is actually related to the power-law decay of the*". In order to guess the next word, 'loss', it would be useful to focus on portions of this paper which appear much earlier, where the loss is defined. It is this type of long-range correlation that is handled by the attention mechanism.

### Combinatorial and hierarchical structure

A third distinctive structure of datasets used in practice is its combinatorial nature. Imagine, for instance, a photo of a lecture hall: it is composed by a group of students, each sitting at his desk. Then each student is "composed" of head, chest, arms, and each head is "composed" of eyes, nose, mouth, hair, and the eyes are "composed" by pigmented epithelial cells, etc. This is actually typical, and most of the images that we want to analyze have this type of combinatorial structure with a hierarchy of features and

subfeatures related to the scale at which one looks. This structure is also related to the decoding that happens when learning from images with a deep network: one typically finds that the first layers of the network decode small scales elements like edges, and going further into the network one gradually identifies larger scale properties, until in the final layers one is able to decide the content of an image. Interestingly, the same type of analysis, from small scale to larger scales, takes place in the sequence of visual areas used in the brains of primates. One also finds the same combinatorial/hierarchical structure in text, for instance, and also in protein sequences with their primary, secondary and ternary structures.

The first attempts at building ensembles with combinatorial/hierarchical properties are still rather rudimentary. An easy case, although not very realistic, is the one of linear structures. Interestingly, one can show that an associative memory network [25] trying to store such hierarchical patterns can be mapped onto a layered network where the first layers analyze the small scale features, and the information is then built gradually to larger scales, by combining smaller scale features of previous layers. Very recently, simple nonlinear versions of combinatorial/hierarchical data ensembles have started to be explored [41, 45].

### Conclusions

Constructing a theory of deep learning is an important challenge, both from the theoretical point of view, but also for applications: only a solid theory will be able to turn a deep network prediction from a black-box best guess into a statement which can be explained and justified, and whose worst-case behavior can be controlled. The main high-level challenge that is faced in deep network is the one of emergence: how is the information gradually elaborated when it is processed from layer to layer in the network? How is it encoded collectively? Contemporary networks are working in a high-dimensional regime, and what we need is a good control of the representations obtained from data of probability distributions in large dimensions. This is typically a problem of statistical physics. One big question is whether we will be able to elaborate a statistical physics of deep network which is based on a not-too-large number of order parameters that can be controlled statistically, as was done in spin glasses.

In order to be relevant, this approach to deep networks must be able to take into account important ingredients of the real 'world', and in particular its structure. So far, spin glass theory has been developed mostly for ensembles in which the coupling constants are identically and independently distributed. It is known that more structured

ensembles can be very hard to study. This is the case, for instance, of the EA model: in this model, the fact that the spins are coupled only among nearest neighbors on a cubic lattice is a type of Euclidean structure, and this problem has not been solved exactly so far. A fascinating new challenge of spin glass theory is to develop new ensembles of correlated disorder, including some of the most relevant ingredients that are found in real databases, like long-range correlations, hierarchy, combinatorial structures and effective dimensions, while being able to keep some analytic control of the problem.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- [1] P W Anderson *Science* **177** 393 (1972)
- [2] C Baldassi, A Inghrosso, C Lucibello, L Saglietti and R Zecchina *Phys. Rev. Lett.* **115** 128101 (2015)
- [3] C Baldassi, F Pittorino and R Zecchina *Proc. Natl. Acad. Sci.* **117** 1 161 (2020)
- [4] J Barbier, F Krzakala, N Macris, L Miolane, L Zdeborová Phase transitions, optimal errors and optimality of message-passing in generalized linear models (2017)
- [5] G Biroli and M Mézard, *J. Stat. Mech.* 093402 (2023)
- [6] E Bolthausen *Commun. Math. Phys.* **325** 1 333 (2014)
- [7] P Charbonneau, M Mézard, E Marinari, G Parisi, F Ricci-Tersenghi, G Sicuro and F Zamponi (eds.) *Spin-Glass Theory and Far Beyond* (Singapore: World Scientific) (2023)
- [8] JA Costa and AO Hero Learning Intrinsic Dimension and Intrinsic Entropy of High-Dimensional Datasets. In 2004 12th European Signal Processing Conference, p 369 (2004)
- [9] A Crisanti and H J Sommers *Zeitschrift für Phys. B Condens. Matter.* **87** 341 (1992)
- [10] L Cugliandolo and J Kurchan *Phys. Rev. Lett.* **71** 173 (1993)
- [11] B Derrida *Phys. Rev. B* **24** 5 2613 (1981)
- [12] S F Edwards and P W Anderson *J. Phys. F* **5** 965 (1975)
- [13] A Engel and C Van den Broeck *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press) (2001)
- [14] C Fefferman, S Mitter and H Narayanan *J. Am. Math. Soc.* **29** 983 (2016)
- [15] S Franz and G Parisi *J. Phys. I* **5** 11 1401 (1995)
- [16] E Gardner *J. Phys. A: Math. Gen.* **21** 257 (1988)
- [17] E Gardner and B Derrida *J. Phys. A: Math. Gen.* **22** 12 1983 (1989)
- [18] S Goldt, M Mézard, F Krzakala and L Zdeborová *Phys. Rev. X* **10** 4 041044 (2020)
- [19] P Grassberger and I Procaccia *Phys. Rev. Lett.* **50** 5 346 (1983)
- [20] D J Gross and M Mézard *Nucl. Phys. B* **240** 431 (1984)
- [21] F Guerra, FL Toninelli *Commun. Math. Phys.* **230** 71 (2002)
- [22] G Györgyi *Phys. Rev. A* **41** 12 7097 (1990)
- [23] M Hein, JY Audibert Intrinsic Dimensionality Estimation of Submanifolds in rd. In Proceedings of the 22nd International Conference on Machine Learning, p 289 (2005)
- [24] D Hérisson and M Ocio *Phys. Rev. Lett.* **88** 25 257202 (2002)
- [25] J J Hopfield *PNAS* **79** 8 2554 (1982)
- [26] H Huang, K M Wong and Y Kabashima *J. Phys. A Math. Theor.* **46** 37 375002 (2013)
- [27] W Krauth and M Mézard *J. de Phys.* **50** 20 3057 (1989)
- [28] C Lauditi, E Troiani, and M Mézard Sparse representations, inference and learning. arXiv preprint, [arXiv:2306.16097](https://arxiv.org/abs/2306.16097) (2023)
- [29] Y LeCun and C Cortes The MNIST database of handwritten digits (1998)
- [30] N Levi and Y Oz The underlying scaling laws and universal statistical structure of complex datasets. (2023). [arXiv:2306.14975](https://arxiv.org/abs/2306.14975)
- [31] E Levina and PJ Bickel Maximum likelihood estimation of intrinsic dimension. In Advances in Neural Information Processing Systems 17 (2004)
- [32] Q Li and H Sompolinsky *Phys. Rev. X* **11** 3 031059 (2021)
- [33] A Maloney, DA Roberts, J Sully A solvable model of neural scaling laws. arXiv preprint, [arXiv:2210.16859](https://arxiv.org/abs/2210.16859) (2022)
- [34] S Mei and A Montanari *Commun. Pure Appl. Math.* **75** 4 667 (2022)
- [35] M Mézard *J. Phys. A: Math. Gen.* **22** 12 2181 (1989)
- [36] M Mézard *Phys. A* **265** 352 (1999)
- [37] M Mézard, G Parisi, N Sourlas, G Toulouse and M A Virasoro *Phys. Rev. Lett.* **52** 1156 (1984)
- [38] M Mézard, G Parisi and M A Virasoro *Spin-Glass Theory and Beyond* (Singapore: World Scientific) (1987)
- [39] M Mézard, G Parisi and M A Virasoro *Europhys. Lett.* **1** 2 77 (1986)
- [40] M Mohri, A Rostamizadeh and A Talwalkar *Foundations of Machine Learning* (Cambridge: MIT Press) (2012)
- [41] E Mossel Deep learning and hierarchal generative models. arXiv preprint, [arXiv:1612.09057](https://arxiv.org/abs/1612.09057) (2016)
- [42] G Parisi *Phys. Rev. Lett.* **43** 23 1754 (1979)
- [43] G Parisi *Phys. Rev. Lett.* **50** 1946 (1983)
- [44] G Parisi Nobel lecture: multiple equilibria. arXiv preprint, [arXiv:2304.00580](https://arxiv.org/abs/2304.00580) (2023)
- [45] L Petrini, F Cagnetta, M Tomasini, A Favero, and M Wyart How deep neural networks learn compositional data: the random hierarchy model. [arXiv:2307.02129](https://arxiv.org/abs/2307.02129) (2023)
- [46] M Refinetti, A Inghrosso, and S Goldt Neural networks trained with sgd learn distributions of increasing complexity. In ICML 2023 (2023)
- [47] H S Seung, H Sompolinsky and N Tishby *Phys. Rev. A* **45** 8 6056 (1992)
- [48] D Sherrington and S Kirkpatrick *Phys. Rev. Lett.* **35** 26 1792 (1975)
- [49] J Sohl-Dickstein, E Weiss, N Maheswaranathan, S Ganguli Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning (2015)
- [50] H Sompolinsky and A Zippelius *Phys. Rev. Lett.* **47** 359 (1981)
- [51] Y Song, S Ermon Generative modeling by estimating gradients of the data distribution. Advances in Neural Information Processing Systems (2019)
- [52] Y Song, J Sohl-Dickstein, DP Kingma, A Kumar, S Ermon, B Poole Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations (2021)
- [53] S Spigler, M Geiger, and M Wyart Asymptotic learning curves of kernel methods: empirical data v.s. Teacher-Student paradigm. [arXiv:1905.10843](https://arxiv.org/abs/1905.10843) (2019)

- 
- [54] D J Thouless, P W Anderson and R G Palmer *Phil. Mag.* **35** 3 593 (1977)
- [55] V Vapnik *The Nature of Statistical Learning Theory* (USA: Springer Science & Business Media) (2013)
- [56] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, Ł Kaiser, I Polosukhin Attention is all you need. *Advances in neural information processing systems*, 30 (2017)
- [57] L Yang, Z Zhang, Y Song, S Hong, R Xu, Y Zhao, W Zhang, B Cui, MH Yang Diffusion models: a comprehensive survey of methods and applications. arXiv preprint, [arXiv:2209.00796](https://arxiv.org/abs/2209.00796) (2022)
- [58] L Zdeborová and F Krzakala Statistical physics of inference: thresholds and algorithms. arXiv preprint, [arXiv:1511.02476](https://arxiv.org/abs/1511.02476) (2015)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.