A Walk in the Statistical Mechanical Formulation of Neural Networks
Alternative Routes to Hebb Prescription

Elena Agliari[1], Adriano Barra[1], Andrea Galluzzi[2], Daniele Tantari[2] and Flavia Tavani[3]
[1]*Dipartimento di Fisica, Sapienza Università di Roma, Rome, Italy*
[2]*Dipartimento di Matematica, Sapienza Università di Roma, Rome, Italy*
[3]*Dipartimento SBAI (Ingegneria), Sapienza Università di Roma, Rome, Italy*

Keywords: Statistical Mechanics, Spin-glasses, Random Graphs.

Abstract: Neural networks are nowadays both powerful operational tools (e.g., for pattern recognition, data mining, error correction codes) and complex theoretical models on the focus of scientific investigation. As for the research branch, neural networks are handled and studied by psychologists, neurobiologists, engineers, mathematicians and theoretical physicists. In particular, in theoretical physics, the key instrument for the quantitative analysis of neural networks is statistical mechanics. From this perspective, here, we review attractor networks: starting from ferromagnets and spin-glass models, we discuss the underlying philosophy and we recover the strand paved by Hopfield, Amit-Gutfreund-Sompolinky. As a sideline, in this walk we derive an alternative (with respect to the original Hebb proposal) way to recover the Hebbian paradigm, stemming from mixing ferromagnets with spin-glasses. Further, as these notes are thought of for an Engineering audience, we highlight also the mappings between ferromagnets and operational amplifiers, hoping that such a bridge plays as a concrete prescription to capture the beauty of robotics from the statistical mechanical perspective.

1 INTRODUCTION

Neural networks are such a fascinating field of science that its development is the result of contributions and efforts from an incredibly large variety of scientists, ranging from *engineers* (mainly involved in electronics and robotics) (Hagan et al., 1996; Miller et al., 1995), *physicists* (mainly involved in statistical mechanics and stochastic processes) (Amit, 1992; Hertz and Palmer, 1991), and *mathematicians* (mainly working in logics and graph theory) (Coolen et al., 2005; Saad, 2009) to *(neuro) biologists* (Harris-Warrick, 1992; Rolls and Treves, 1998) and *(cognitive) psychologists* (Martindale, 1991; Domhoff, 2003).

Tracing the genesis and evolution of neural networks is very difficult, probably due to the broad meaning they have acquired along the years; scientists closer to the robotics branch often refer to the W. McCulloch and W. Pitts model of perceptron (McCulloch and Pitts, 1943), or the F. Rosenblatt version (Rosenblatt, 1958), while researchers closer to the neurobiology branch adopt D. Hebb's work as a starting point (Hebb, 1940). On the other hand, scientists involved in statistical mechanics, that joined the community in relatively recent times, usually refer to the seminal paper by Hopfield (Hopfield, 1982) or to the celebrated work by Amit Gutfreund Sompolinky (Amit, 1992), where the statistical mechanics analysis of the Hopfield model is effectively carried out.

Whatever the reference framework, at least 30 years elapsed since neural networks entered in the theoretical physics research and much of the former results can now be re-obtained or re-framed in modern approaches and made much closer to the engineering counterpart, as we want to highlight in the present work. In particular, we show that toy models for paramagnetic-ferromagnetic transition (Ellis, 2005) are natural prototypes for the autonomous storage/retrieval of information patterns and play as operational amplifiers in electronics. Then, we move further analyzing the capabilities of glassy systems (ensembles of ferromagnets and antiferromagnets) in storing/retrieving extensive numbers of patterns so to recover the Hebb rule for learning (Hebb, 1940) far from the original route contained in his milestone *The Organization of Behavior*. Finally, we will give prescription to map these glassy systems in ensembles of amplifiers and inverters (thus flip-flops) of the engineering counterpart so to offer a concrete bridge be-

Agliari E., Barra A., Galluzzi A., Tantari D. and Tavani F..
A Walk in the Statistical Mechanical Formulation of Neural Networks - Alternative Routes to Hebb Prescription.
DOI: 10.5220/0005077902100217
In *Proceedings of the International Conference on Neural Computation Theory and Applications* (NCTA-2014), pages 210-217
ISBN: 978-989-758-054-3

tween the two communities of theoretical physicists working with complex systems and engineers working with robotics and information processing.

As these notes are intended for non-theoretical-physicists, we believe that they can constitute a novel perspective on a by-now classical theme and that they could hopefully excite curiosity toward statistical mechanics in nearest neighbors scientists like engineers whom these proceedings are addressed to.

2 STATISTICAL MECHANICS IN A NUTSHELL

Hereafter we summarize the fundamental steps that led theoretical physicists towards artificial intelligence; despite this parenthesis may look rather distant from neural network scenarios, it actually allows us to outline and to historically justify the physicists perspective.

Statistical mechanics aroused in the last decades of the XIX century thanks to its founding fathers Ludwig Boltzmann, James Clarke Maxwell and Josiah Willard Gibbs (Kittel, 2004). Its “solely” scope (at that time) was to act as a theoretical ground of the already existing empirical thermodynamics, so to reconcile its noisy and irreversible behavior with a deterministic and time reversal microscopic dynamics. While trying to get rid of statistical mechanics in just a few words is almost meaningless, roughly speaking its functioning may be summarized via toy-examples as follows. Let us consider a very simple system, e.g. a perfect gas: its molecules obey a Newton-like microscopic dynamics (without friction -as we are at the molecular level- thus time-reversal as dissipative terms in differential equations capturing system’s evolution are coupled to odd derivatives) and, instead of focusing on each particular trajectory for characterizing the state of the system, we define order parameters (e.g. the density) in terms of microscopic variables (the particles belonging to the gas). By averaging their evolution over suitably probability measures, and imposing on these averages energy minimization and entropy maximization, it is possible to infer the macroscopic behavior in agreement with thermodynamics, hence bringing together the microscopic deterministic and time reversal mechanics with the macroscopic strong dictates stemmed by the second principle (i.e. arrow of time coded in the entropy growth). Despite famous attacks to Boltzmann theorem (e.g. by Zermelo or Poincaré) (Castiglione et al., 2012), statistical mechanics was immediately recognized as a deep and powerful bridge linking microscopic dynamics of a system’s constituents with (emergent) macroscopic properties shown by the system itself, as exemplified by the equation of state for *perfect gases* obtained by considering an Hamiltonian for a single particle accounting for the kinetic contribution only (Kittel, 2004).

One step forward beyond the perfect gas, Van der Waals and Maxwell in their pioneering works focused on *real gases* (Reichl and Prigogine, 1980), where particle interactions were finally considered by introducing a non-zero potential in the microscopic Hamiltonian describing the system. This extension implied fifty-years of deep changes in the theoretical-physics perspective in order to be able to face new classes of questions. The remarkable reward lies in a theory of phase transitions where the focus is no longer on details regarding the system constituents, but rather on the characteristics of their interactions. Indeed, phase transitions, namely abrupt changes in the macroscopic state of the whole system, are not due to the particular system considered, but are primarily due to the ability of its constituents to perceive interactions over the thermal noise. For instance, when considering a system made of by a large number of water molecules, whatever the level of resolution to describe the single molecule (ranging from classical to quantum), by properly varying the external tunable parameters (e.g. the temperature), this *system* eventually changes its state from liquid to vapor (or solid, depending on parameter values); of course, the same applies generally to liquids.

The fact that the macroscopic behavior of a system may spontaneously show *cooperative, emergent* properties, actually hidden in its microscopic description and not directly deducible when looking at its components alone, was definitely appealing in neuroscience. In fact, in the 70s neuronal dynamics along axons, from dendrites to synapses, was already rather clear (see e.g. the celebrated book by Tuckwell (Tuckwell, 2005)) and not too much intricate than circuits that may arise from basic human creativity: remarkably simpler than expected and certainly trivial with respect to overall cerebral functionalities like learning or computation, thus the aptness of a *thermodynamic formulation* of neural interactions -to *reveal* possible emergent capabilities- was immediately pointed out, despite the route was not clear yet.

Interestingly, a big step forward to this goal was prompted by problems stemmed from condensed matter. In fact, quickly theoretical physicists realized that the purely kinetic Hamiltonian, introduced for perfect gases (or Hamiltonian with mild potentials allowing for real gases), is no longer suitable for solids, where atoms do not move freely and the main energy contributions are from potentials. An ensemble of harmonic oscillators (mimicking atomic oscillations of the nuclei around their rest positions) was the first scenario for understanding condensed matter: however, as experimentally revealed by crystallography, nuclei are arranged according to regular lattices hence motivating mathematicians in study periodical structures to help physicists in this modeling, but merging statistical mechanics with lattice theories resulted soon in practically intractable models.

As a paradigmatic example, let us consider the one-dimensional Ising model, originally introduced to investigate magnetic properties of matter: the generic, out of N, nucleus labeled as i is schematically represented by a spin σ_i, which can assume only two values ($\sigma_i = -1$, spin down and $\sigma_i = +1$, spin up); nearest neighbor spins interact reciprocally through positive (i.e. ferromagnetic) interactions $J_{i,i+1} > 0$, hence the Hamiltonian of this system can be written as $H_N(\sigma) \propto -\sum_i^N J_{i,i+1}\sigma_i\sigma_{i+1} - h\sum_i^N \sigma_i$, where h tunes the external magnetic field and the minus sign in front of each term of the Hamiltonian ensures that spins try to align with the external field and to get parallel each other in order to fulfill the minimum energy principle. Clearly this model can trivially be extended to higher dimensions, however, due to prohibitive difficulties in facing the topological constraint of considering nearest neighbor interactions only, soon shortcuts were properly implemented to turn around this path. It is just due to an effective shortcut, namely the so called "mean field approximation", that statistical mechanics approached complex systems and, in particular, artificial intelligence.

3 THE ROUTE TO COMPLEXITY

As anticipated, the "mean field approximation" allows overcoming prohibitive technical difficulties owing to the underlying lattice structure. This consists in extending the sum on nearest neighbor couples (which are $O(N)$) to include all possible couples in the system (which are $O(N^2)$), properly rescaling the coupling ($J \rightarrow$J/N) in order to keep thermodynamical observables linearly extensive. If we consider a ferromagnet built of by N Ising spins $\sigma_i = \pm 1$ with $i \in (1, ..., N)$, we can then write

$$H_N(\sigma|J) = -\frac{1}{N}\sum_{i<j}^{N,N} J_{ij}\sigma_i\sigma_j \sim -\frac{1}{2N}\sum_{i,j}^{N,N} J_{ij}\sigma_i\sigma_j, \quad (1)$$

where in the last term we neglected the diagonal term ($i = j$) as it is irrelevant for large N. From a topological perspective the mean-field approximation equals to abandon the lattice structure in favor to a complete graph (see Fig. 1).When the coupling matrix has only

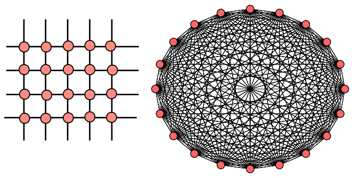

Figure 1: Example of regular lattice (left) and complete graph (right) with $N = 20$ nodes. In the former only nearest-neighbors are connected in such a way that the number of links scales linearly with N, while in the latter each node is connected with all the remaining $N-1$ in such a way that the number of links scales quadratically with N.

positive entries, e.g. $P(J_{ij}) = \delta(J_{ij} - J)$, this model is named Curie-Weiss model and acts as the simplest microscopic Hamiltonian able to describe the paramagnetic-ferromagnetic transitions experienced by materials when temperature is properly lowered. An external (magnetic) field h can be accounted for by adding in the Hamiltonian an extra term $\propto -h\sum_i^N \sigma_i$.

According to the principle of minimum energy, the two-body interaction appearing in the Hamiltonian in Eq. 1 tends to make spins parallel with each other and aligned with the external field if present. However, in the presence of noise (i.e. if temperature T is strictly positive), maximization of entropy must also be taken into account. When the noise level is much higher than the average energy (roughly, if $T \gg J$), noise and entropy-driven disorder prevail and spins are not able to "feel" reciprocally; as a result, they flip randomly and the system behaves as a *paramagnet*. Conversely, if noise is not too loud, spins start to interact possibly giving rise to a phase transition; as a result the system globally rearranges its structure orientating all the spins in the same direction, which is the one selected by the external field if present, thus we have a *ferromagnet*.

In the early '70 a scission occurred in the statistical mechanics community: on the one side "pure physicists" saw mean-field approximation as a merely bound to bypass in order to have satisfactory pictures of the structure of matter and they succeeded in working out iterative procedures to embed statistical mechanics in (quasi)-three-dimensional reticula, yielding to the *renormalization group* (Wilson, 1971): this proliferative branch gave then rise to superconductivity, superfluidity (Bean, 1962) and many-body problems in condensed matter (Bardeen et al., 1957).
Conversely, from the other side, the mean-field approximation acted as a breach in the wall of complex systems: a thermodynamical investigation of phenomena occurring on general structures lacking Euclidean metrics (whose subject largely covers neural networks too) was then possible.

4 TOWARD NEURAL NETWORKS

Hereafter we discuss how to approach neural networks from models mimicking ferromagnetic transition. In particular, we study the Curie-Weiss model and we show how it can store one pattern of information and then we bridge its input-output relation (called *self-consistency*) with the transfer function of an operational amplifier. Then, we notice that such a stored pattern has a very peculiar structure which is hardly *natural*, but we will overcome this (fake) flaw by introducing a gauge variant known as Mattis model. This scenario can be looked at as a primordial neural network and we discuss its connection with biological neurons and operational amplifiers. The successive step consists in extending, through elementary thoughts, this picture in order to include and store several patterns. In this way, we recover both the Hebb rule for synaptic plasticity and, as a corollary, the Hopfield model for neural networks too that will be further analyzed in terms of flip-flops and information storage.

The statistical mechanical analysis of the Curie-Weiss model (CW) can be summarized as follows: Starting from a microscopic formulation of the system, i.e. N spins labeled as $i, j, \ldots$, their pairwise couplings $J_{ij} \equiv J$, and possibly an external field h, we derive an explicit expression for its (macroscopic) free energy $A(\beta)$. The latter is the effective energy, namely the difference between the internal energy U, divided by the temperature $T = 1/\beta$, and the entropy S, namely $A(\beta) = S - \beta U$, in fact, S is the "penalty" to be paid to the Second Principle for using U at noise level β. We can therefore link macroscopic free energy with microscopic dynamics via the fundamental relation

$$A(\beta) = \lim_{N\to\infty} \frac{1}{N} \ln \sum_{\{\sigma\}}^{2^N} \exp\left[-\beta H_N(\sigma|J,h)\right], \quad (2)$$

where the sum is performed over the set $\{\sigma\}$ of all 2^N possible spin configurations, each weighted by the Boltzmann factor $\exp[-\beta H_N(\sigma|J,h)]$ that tests the likelihood of the related configuration. From expression (2), we can derive the whole thermodynamics and in particular phase-diagrams, that is, we are able to discern regions in the space of tunable parameters (e.g. temperature/noise level) where the system behaves as a paramagnet or as a ferromagnet.

Thermodynamical averages, denoted with the symbol $\langle . \rangle$, provide for a given observable the expected value, namely the value to be compared with measures in an experiment. For instance, for the magnetization $m(\sigma) \equiv \sum_{i=1}^N \sigma_i / N$ we have

$$\langle m(\beta) \rangle = \frac{\sum_\sigma m(\sigma) e^{-\beta H_N(\sigma|J)}}{\sum_\sigma e^{-\beta H_N(\sigma|J)}}. \quad (3)$$

When $\beta \to \infty$ the system is noiseless (zero temperature) hence spins feel reciprocally without errors and the system behaves ferromagnetically ($|\langle m \rangle| \to 1$), while when $\beta \to 0$ the system behaves completely random (infinite temperature), thus interactions can not be felt and the system is a paramagnet ($\langle m \rangle \to 0$). In between a phase transition happens.

In the Curie-Weiss model the magnetization works as *order parameter*: its thermodynamical average is zero when the system is in a paramagnetic (disordered) state ($\to \langle m \rangle = 0$), while it is different from zero in a ferromagnetic state (where it can be either positive or negative, depending on the sign of the external field). Dealing with order parameters allows us to avoid managing an extensive number of variables σ_i, which is practically impossible and, even more important, it is not strictly necessary.

Now, an explicit expression for the free energy in terms of $\langle m \rangle$ can be obtained carrying out summations in eq. 2 and taking the *thermodynamic limit* $N \to \infty$ as

$$A(\beta) = \ln 2 + \ln\cosh[\beta(J\langle m \rangle + h)] - \frac{\beta J}{2}\langle m \rangle^2. \quad (4)$$

In order to impose thermodynamical principles, i.e. energy minimization and entropy maximization, we need to find the extrema of this expression with respect to $\langle m \rangle$ requesting $\partial_{\langle m(\beta) \rangle} A(\beta) = 0$. The resulting expression is called the *self-consistency* and it reads as

$$\partial_{\langle m \rangle} A(\beta) = 0 \Rightarrow \langle m \rangle = \tanh[\beta(J\langle m \rangle + h)]. \quad (5)$$

This expression returns the average behavior of a spin in a magnetic field. In order to see that a phase transition between paramagnetic and ferromagnetic states actually exists, we can fix $h = 0$ and expand the r.h.s. of eq. 5 to get

$$\langle m \rangle \propto \pm\sqrt{\beta J - 1}. \quad (6)$$

Thus, while the noise level is higher than one ($\beta < \beta_c \equiv J^{-1}$ or $T > T_c \equiv J$) the only solution is $\langle m \rangle = 0$, while, as far as the noise is lowered below its critical threshold β_c, two different-from-zero branches of solutions appear for the magnetization and the system becomes a ferromagnet (see Fig.2 (left)). The branch effectively chosen by the system usually depends on the sign of the external field or boundary fluctuations: $\langle m \rangle > 0$ for $h > 0$ and vice versa for $h < 0$.

Clearly, the lowest energy minima correspond to the two configurations with all spins aligned, either

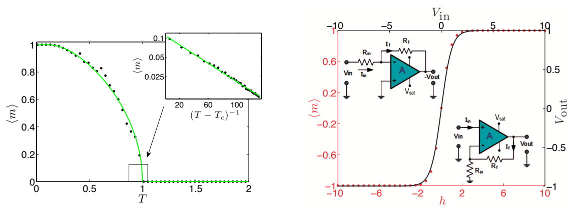

Figure 2: (left) Average magnetization $\langle m \rangle$ versus temperature T for a Curie-Weiss model in the absence of field ($h = 0$). The critical temperature $T_c = 1$ separates a magnetized region ($|\langle m \rangle| > 0$, only one branch shown) from a non-magnetized region ($\langle m \rangle = 0$). The box zooms over the critical region (notice the logarithmic scale) and highlights the power-law behavior $m \sim (T - T_c)^{\beta}$, where $\beta = 1/2$ is also referred to as critical exponent (see also eq. 6). Data shown here ($\bullet$) are obtained via Monte Carlo simulations for a system of $N = 10^5$ spins and compared with the theoretical curve (solid line). (right) Average magnetization $\langle m \rangle$ versus the external field h and response of a charging neuron (solid black line), compared with the transfer function of an operational amplifier (red bullets) (Tuckwell, 2005; Agliari et al., 2013). In the inset we show a schematic representation of an operational amplifier (upper) and of an inverter (lower).

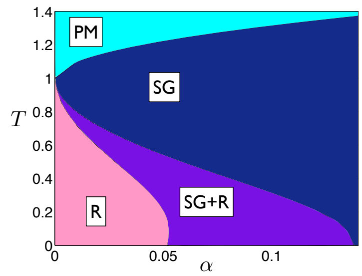

Figure 3: Phase diagram for the Hopfield model (Amit, 1992). According to the parameter setting, the system behaves as a paramagnet (PM), as a spin-glass (SG), or as an associative neural network able to perform information retrieval (R). The region labeled (SG+R) is a coexistence region where the system is glassy but still able to retrieve.

upwards ($\sigma_i = +1, \forall i$) or downwards ($\sigma_i = -1, \forall i$), these configurations being symmetric under spin-flip $\sigma_i \to -\sigma_i$. Therefore, the thermodynamics of the Curie-Weiss model is solved: energy minimization tends to align the spins (as the lowest energy states are the two ordered ones), however entropy maximization tends to randomize the spins (as the higher the entropy, the most disordered the states, with half spins up and half spins down): the interplay between the two principles is driven by the level of noise introduced in the system and this is in turn ruled by the tunable parameter $\beta \equiv 1/T$ as coded in the definition of free energy.

A crucial bridge between condensed matter and neural network could now be sighted: One could think at each spin as a basic neuron, retaining only its ability to spike such that $\sigma_i = +1$ and $\sigma_i = -1$ represent firing and quiescence, respectively, and associate to each equilibrium configuration of this spin system a *stored pattern* of information. The reward is that, in this way, the spontaneous (i.e. thermodynamical) tendency of the network to relax on free-energy minima can be related to the spontaneous retrieval of the stored pattern, such that the cognitive capability emerges as a natural consequence of physical principles: we well deepen this point along the whole paper.
Let us now tackle the problem by another perspective and highlight a structural/mathematical similarity in the world of electronics: the plan is to compare self-consistencies in statistical mechanics and transfer functions in electronics so to reach a unified description for these systems: keeping the symbols of Fig. 2 (insets in the right panel), where R_{in} stands for the input resistance while R_f represents the feed-back resistance, $i_+ = i_- = 0$ and assuming $R_{in} = 1\Omega$ -without loss of generality as only the ratio R_f/R_{in} matters- the following transfer function is achieved:

$$V_{out} = GV_{in} = (1 + R_f)V_{in}, \tag{7}$$

where $G = 1 + R_f$ is called *gain*, therefore as far as $0 > R_f > \infty$ (thus retro-action is present) the device is amplifying.
Let us emphasize deep structural analogies with the Curie-Weiss response to a magnetic field h: once fixed $\beta = 1$ for simplicity, expanding $\langle m \rangle = \tanh(J\langle m \rangle + h) \sim (1 + J)h$, we can compare term by term the two expression as

$$V_{out} = (1 + R_f)V_{in}, \tag{8}$$
$$\langle m \rangle = (1 + J)h. \tag{9}$$

We see that R_f plays as J, and, consistently, if R_f is absent the retroaction is lost in the op-amp and the gain is no longer possible; analogously if $J = 0$, spins do not mutually interact and no feed-back is allowed to drive the phase transition.

Actually, the Hamiltonian (1) would encode for a rather poor model of neural network as it would account for only two stored patterns, corresponding to the two possible minima, moreover, these ordered patterns, seen as information chains, have the lowest possible entropy and, for the Shannon-McMillan Theorem, in the large N limit, they will never be observed.

This criticism can be easily overcome thanks to the Mattis-gauge, namely a re-definition of the spins via $\sigma_i \to \xi_i^1 \sigma_i$, where $\xi_i^1 = \pm 1$ are random entries extracted with equal probability and kept fixed in the network (in statistical mechanics these are called *quenched* variables to stress that they do not contribute to thermalization, a terminology reminiscent of metallurgy (Mézard et al., 1987)). Fixing $J \equiv 1$ for simplicity, the Mattis Hamiltonian reads as

$$H_N^{Mattis}(\sigma|\xi) = -\frac{1}{2N}\sum_{i,j}^{N,N}\xi_i^1\xi_j^1\sigma_i\sigma_j - h\sum_i^N\xi_i^1\sigma_i. \quad (10)$$

The Mattis magnetization is defined as $m_1 = \sum_i \xi_i^1\sigma_i$. To inspect its lowest energy minima, we perform a comparison with the CW model: in terms of the (standard) magnetization, the Curie-Weiss model reads as $H_N^{CW} \sim -(N/2)m^2 - hm$ and, analogously we can write $H_N^{Mattis}(\sigma|\xi)$ in terms of Mattis magnetization as $H_N^{Mattis} \sim -(N/2)m_1^2 - hm_1$. It is then evident that, in the low noise limit (namely where collective properties may emerge), as the minimum of free energy is achieved in the Curie-Weiss model for $\langle m\rangle \to \pm 1$, the same holds in the Mattis model for $\langle m_1\rangle \to \pm 1$. However, this implies that now spins tend to align parallel (or antiparallel) to the vector ξ^1, hence if the latter is, say, $\xi^1 = (+1,-1,-1,-1,+1,+1)$ in a model with $N = 6$, the equilibrium configurations of the network will be $\sigma = (+1,-1,-1,-1,+1,+1)$ and $\sigma = (-1,+1,+1,+1,-1,-1)$, the latter due to the gauge symmetry $\sigma_i \to -\sigma_i$ enjoyed by the Hamiltonian. Thus, the network relaxes autonomously to a state where some of its neurons are firing while others are quiescent, according to the *stored pattern* ξ^1. Note that, as the entries of the vectors ξ are chosen randomly ± 1 with equal probability, the retrieval of free energy minimum now corresponds to a spin configuration which is also the most entropic for the Shannon-McMillan argument, thus both the most likely and the most difficult to handle (as its information compression is no longer possible).

Two remarks are in order now. On the one side, according to the self-consistency equation (5) and as shown in Fig. 2 (right), $\langle m\rangle$ versus h displays the typical graded/sigmoidal response of a charging neuron (Tuckwell, 2005), and one would be tempted to call the spins σ neurons. On the other side, it is definitely inconvenient to build a network via N spins/neurons, which are further meant to be diverging (i.e. $N \to \infty$) in order to handle one stored pattern of information only. Along the theoretical physics route overcoming this limitation is quite natural (and provides the first derivation of the Hebbian prescription in this paper): If we want a network able to cope with P patterns, the starting Hamiltonian should have simply the sum over these P previously stored patterns, namely

$$H_N(\sigma|\xi) = -\frac{1}{2N}\sum_{i,j=1}^{N,N}\left(\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu\right)\sigma_i\sigma_j, \quad (11)$$

where we neglect the external field ($h = 0$) for simplicity. As we will see in the next section, this Hamiltonian constitutes indeed the Hopfield model, namely the harmonic oscillator of neural networks, whose coupling matrix is called *Hebb matrix* as encodes the Hebb prescription for neural organization (Amit, 1992).

Despite the extension to the case $P > 1$ is formally straightforward, the investigation of the system as P grows becomes by far more tricky. Indeed, neural networks belong to the so-called "complex systems" realm. We propose that complex behaviors can be distinguished by simple behaviors as for the latter the number of free-energy minima of the system *does not scale* with the volume N, while for complex systems the number of free-energy minima *does scale* with the volume according to a proper function of N. For instance, the Curie-Weiss/Mattis model has two minima only, whatever N (even if $N \to \infty$), and it constitutes the paradigmatic example for a simple system. As a counterpart, the prototype of complex system is the Sherrington-Kirkpatrick model (SK), originally introduced in condensed matter to describe the peculiar behaviors exhibited by real glasses (Hertz and Palmer, 1991; Mézard et al., 1987). This model has an amount of minima that scales $\propto \exp(cN)$ with $c \neq f(N)$, and its Hamiltonian reads as

$$H_N^{SK}(\sigma|J) = \frac{1}{\sqrt{N}}\sum_{i<j}^{N,N} J_{ij}\sigma_i\sigma_j, \quad (12)$$

where, crucially, coupling are Gaussian distributed as $P(J_{ij}) \equiv \mathcal{N}[0,1]$. This implies that links can be either positive (hence favoring parallel spin configuration) as well as negative (hence favoring anti-parallel spin configuration), thus, in the large N limit, with large probability, spins will receive conflicting signals and we speak about "frustrated networks". Indeed *frustration*, the hallmark of complexity, is fundamental in order to split the phase space in several disconnected zones, i.e. in order to have several minima, or several stored patterns in neural network language. This mirrors a clear request also in electronics, namely the need for inverting amplifiers too.

The mean-field statistical mechanics for the low-noise behavior of spin-glasses has been first described by Parisi and it predicts a hierarchical organization of states and a relaxational dynamics spread over many timescales (for which we refer to specific textbooks (Mézard et al., 1987)). Here we just need to know that their natural order parameter is no longer the magnetization (as these systems do not magnetize), but the *overlap* q_{ab}, as we are explaining. Spin glasses are balanced ensembles of ferromagnets and antiferromagnets (this can also be seen mathematically as $P(J)$ is symmetric around zero) and, as a result, $\langle m\rangle$ is always equal to zero, on the other hand, a comparison between two realizations of the system (pertaining to the same coupling set) is meaningful because at large temperatures it is expected to be zero, as everything is uncorrelated, but at low temperature their overlap is strictly non-zero as spins freeze in disordered but correlated states. More precisely, given two "replicas" of the system, labeled as a and b, their overlap q_{ab} can be defined as the scalar product between the related spin configurations, namely as $q_{ab} = (1/N)\sum_i^N \sigma_i^a \sigma_i^b$, thus the mean-field spin glass has a completely random paramagnetic phase, with $\langle q \rangle \equiv 0$ and a "glassy phase" with $\langle q \rangle > 0$ split by a phase transition at $\beta_c = T_c = 1$.

The Sherrington-Kirkpatrick model displays a large number of minima as expected for a cognitive system, yet it is not suitable to act as a cognitive system because its states are too "disordered". We look for an Hamiltonian whose minima are not purely random like those in SK, as they must represent ordered stored patterns (hence like the CW ones), but the amount of these minima must be possibly extensive in the number of spins/neurons N (as in the SK and at contrary with CW), hence we need to retain a "ferromagnetic flavor" within a "glassy panorama": we need *something in between*.

Remarkably, the Hopfield model defined by the Hamiltonian (11) lies exactly in between a Curie-Weiss model and a Sherrington-Kirkpatrick model. Let us see why: When $P = 1$ the Hopfield model recovers the Mattis model, which is nothing but a gauge-transformed Curie-Weiss model. Conversely, when $P \to \infty$, $(1/\sqrt{N})\sum_\mu^P \xi_i^\mu \xi_j^\mu \to \mathcal{N}[0,1]$, by the standard central limit theorem, and the Hopfield model recovers the Sherrington-Kirkpatrick one. In between these two limits the system behaves as an associative network (Barra et al., 2012).
Such a crossover between CW (or Mattis) and SK models, requires for its investigation both the P Mattis magnetization $\langle m_\mu \rangle$, $\mu = (1,...,P)$ (for quantifying retrieval of the whole stored patterns, that is the *vocabulary*), and the two-replica overlaps $\langle q_{ab} \rangle$ (to control the glassyness growth if the vocabulary gets enlarged), as well as a tunable parameter measuring the ratio between the stored patterns and the amount of available neurons, namely $\alpha = \lim_{N\to\infty} P/N$, also referred to as *network capacity*.

As far as P scales sub-linearly with N, i.e. in the low storage regime defined by $\alpha = 0$, the phase diagram is ruled by the noise level β only: for $\beta < \beta_c$ the system is a paramagnet, with $\langle m_\mu \rangle = 0$ and $\langle q_{ab} \rangle = 0$, while for $\beta > \beta_c$ the system performs as an attractor network, with $\langle m_\mu \rangle \neq 0$ for a given μ (selected by the external field) and $\langle q_{ab} \rangle = 0$. In this regime no dangerous glassy phase is lurking, yet the model is able to store only a tiny amount of patterns as the capacity is sub-linear with the network volume N.
Conversely, when P scales linearly with N, i.e. in the high-storage regime defined by $\alpha > 0$, the phase diagram lives in the α, β plane (see Fig. 3).When α is small enough the system is expected to behave similarly to $\alpha = 0$ hence as an associative network (with a particular Mattis magnetization positive but with also the two-replica overlap slightly positive as the glassy nature is intrinsic for $\alpha > 0$). For α large enough ($\alpha > \alpha_c(\beta), \alpha_c(\beta \to \infty) \sim 0.14$) however, the Hopfield model collapses on the Sherrington-Kirkpatrick model as expected, hence with the Mattis magnetizations brutally reduced to zero and the two-replica overlap close to one. The transition to the spin-glass phase is often called "blackout scenario" in neural network community.

5 CONCLUSIONS

We conclude this survey on the statistical mechanical approach to neural networks with a remark about possible perspectives: we started this historical tour highlighting how, thanks to the mean-field paradigm, engineering (e.g. robotics, automation) and neurobiology have been tightly connected from a theoretical physics perspective. However, as statistical mechanics is starting to access techniques to tackle complexity hidden even in non-mean-field networks (e.g. as in the hierarchical graphs, where thermodynamics for the glassy scenario is almost complete (Castellana et al., 2010)), we will probably witness another split in this smaller community of theoretical physicists working in spontaneous computational capability research: from one side continuing to refine techniques and models meant for artificial systems, well lying in high-dimensional/mean-field topologies, and from the other beginning to develop ideas, models and techniques meant for biological systems only, strictly defined in finite-dimensional spaces or, even worst, embedded on fractal supports.

This work was supported by Gruppo Nazionale per la Fisica Matematica (GNFM), Istituto Nazionale d'Alta Matematica (INdAM).

REFERENCES

Agliari, E., Barra, A., Burioni, R., Di Biasio, A., and Uguzzoni, G. (2013). Collective behaviours: from biochemical kinetics to electronic circuits. *Scientific Reports*, 3.

Amit, D. J. (1992). *Modeling brain function*. Cambridge University Press.

Bardeen, J., Cooper, L. N., and Schrieffer, J. R. (1957). Theory of superconductivity. *Physical Review*, 108:1175.

Barra, A., Genovese, G., Guerra, F., and Tantari, D. (2012). How glassy are neural networks?. *Journal of Statistical Mechanics: Theory and Experiment*, P07009.

Bean, C. P. (1962). Magnetization of hard superconductors. *Physical Review Letters*, 8:250.

Castellana, M., Decelle, A., Franz, S., Mezard, M., and Parisi, G. (2010). The hierarchical random energy model. *Physical Review Letters*, 104:127206.

Castiglione, P., Falcioni, M., Lesne, A., and Vulpiani, A. (2012). *Chaos and coarse graining in statistical mechanics,*. Cambridge University Press.

Coolen, A. C. C., Kühn, R., and Sollich, P. (2005). *Theory of neural information processing systems*. Oxford University Press.

Domhoff, G. W. (2003). *Neural networks, cognitive development, and content analysis*. American Psychological Association.

Ellis, R. (2005). *Entropy, large deviations, and statistical mechanics.*, volume 1431. Taylor & Francis.

Hagan, M. T., Demuth, H. B., and Beale, M. H. (1996). *Neural network design*. Pws Pub.,, Boston.

Harris-Warrick, R. M., editor (1992). *Dynamic biological networks*. MIT press.

Hebb, D. O. (1940). The organization of behavior: A neuropsychological theory. *Psychology Press*.

Hertz, John, A. K. and Palmer, R. (1991). *Introduction to the theory of neural networks*. Lecture Notes.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. A. Sc.*, 79(8):2554–2558.

Kittel, C. (2004). *Elementary statistical physics*. Courier Dover Publications,.

Martindale, C. (1991). *Cognitive psychology: A neural-network approach*. Thomson Brooks/Cole Publishing Co.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5.4:115–133.

Mézard, M., Parisi, G., and Virasoro, M. A. (1987). *Spin glass theory and beyond*, volume 9. World scientific, Singapore.

Miller, W. T., Werbos, P. J., and Sutton, R. S., editors (1995). *Neural networks for control*. MIT press.

Reichl, L. E. and Prigogine, I. (1980). *A modern course in statistical physics*, volume 71. University of Texas press, Austin.

Rolls, E. T. and Treves, A. (1998). *Neural networks and brain function*.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Saad, D., editor (2009). *On-line learning in neural networks*, volume 17. Cambridge University Press.

Tuckwell, H. C. (2005). *Introduction to theoretical neurobiology.*, volume 8. Cambridge University Press.

Wilson, K. G. (1971). Renormalization group and critical phenomena. *Physical Review B*, 4:3174.