THERMODYNAMIC FORMALISM OF NEURAL COMPUTING[1]

Dimitri Petritis
Institut de Recherche Mathématique
Université de Rennes I and cnrs ura 305
Campus de Beaulieu
F - 35042 Rennes Cedex
petritis@levy.univ-rennes1.fr

15 May 1995

Abstract

Neural networks are systems of interconnected processors mimicking some of the brain functions. After a rapid overview of neural computing, the thermodynamic formalism of the learning procedure is introduced. Besides its use in introducing efficient stochastic learning algorithms, it gives an insight in terms of information theory. Main emphasis is given in the information restitution process; stochastic evolution is used as the starting point for introducing statistical mechanics of associative memory. Instead of formulating problems in their most general setting, it is preferred stating precise results on specific models. In this report are mainly presented those features that are relevant when the neural net becomes very large. A survey of the most recent results is given and the main open problems are pointed out.

[1] Work partially supported by eu network chrx-ct93-0411. Lectures given at the IV$^{\text{th}}$ summer school on "Statistical mechanics and co-operative systems" held in Santiago de Chile, 12–16 December 1994.
1991 Mathematics Subject Classification: 82C32, 82B44, 82B20, 60K35
1986 Physics and Astronomy Classification Scheme: 05.20, 75.40
Key words and phrases: Neural networks, thermodynamic formalism, Gibbs measures, Hopfield model, self-averaging.

1 Introduction and motivation

Comparison of the computational efficiency between the most powerful computer and the brain of the most rudimentary animal shows a really overwhelming advantage to the biological "computer" for tasks involving learning and generalisation [22]. Therefore, a considerable effort has been made to understand the brain functions and possibly mimic some of them for computational purposes. In this report, we only focus on the memory function and compare the main features between computer and brain memories. Thus,

- computer memory is localised: damage on a memory cell of the computer destroys the information contained in it; cerebral memory is diffused: even extended damage on some parts of the brain (due to trauma or tumour) does not significantly affect the contained information.
- magnetic memory is addressed: one has to know the exact location of the memory cell to recall the information contained in it; brain memory is associative: it is pointless to remember that the last friend you met on your way was the 587th person you were acquainted with; instead, it is preferable to recall her voice or the shape of her face.
- computer memory is permanent up to the moment the memory cell containing it is erased; on the contrary, there are two main types of brain memory: short term memory, erased within few minutes, and long term one, engraved up to the moment of the death, with the whole intermediate spectrum. Besides, it is very difficult to force a well memorised fact to be forgotten!
- computer memory is immediately updated in contrary to the brain memory that is progressive, an information needing repetitions in order to become permanently memorised.

Researchers from various disciplines — mathematics, computer sciences, physics, biology, psychology, linguistics, neurophysiology — are interested in the general study of brain functions. However, the epistemological backgrounds and *rationale* of these scientists are quite different. Those coming from physics or mathematics use an analytic method to describe a formidably complex system — the neural system — in terms of fairly simple, often simplist, constituting *interacting* units — the neurones. Based on their experience on phase transitions and statistical mechanics, they expect that the large scale behaviour of the system is independent of the details of the individual units so that brain functions can be expressed in terms of a small number of characteristics of the individual neurones. Those coming from medical or behavioural sciences are willing to understand how the real brain of mammals works. Finally those coming from computer sciences are interested in using the information about the brain functioning to construct more powerful computers [74].

In this survey no such issues are treated. Instead, a utilitarian point of view is adopted; the neural nets are defined as mathematical models and the consequences and implications

of this definition are explained in a deductive form. Thus, mainly the mathematical aspects of the subject are treated. Although the heuristic results obtained by physicists are omnipresent and act as a *Leitmotiv* in the sequel, only rigorously established results are presented here; the reason is that there exist two excellent surveys [5, 66] on the physical results and conjectures on the subject. Moreover, not all the mathematical results are presented; a selection — dictated by personal preferences of the author — of the material is performed; the reader must always keep in mind that this text is the written version of a series of lectures taught within a finite time to an audience of real (summer) students; therefore her indulgence for omissions is implored.

Neural nets are arrays of simple processors mimicking some of characteristics of brain memory [5, 38, 39, 45, 53, 70, 95]. Several systems carry the name *neural net* nowadays; to be systematic, a system must have the following ingredients to be recognised as a neural net:

- a simple oriented graph $G = (V, E)$ where V is the set of vertices (sites), called *neurones*, and $E \subset V \times V$ is the set of edges, called *synapses*.
- a set $S \subseteq \mathbb{R}^d$ of possible *states* of every neurone. In most cases, the set S will be chosen as the binary set $S = \{-1, 1\}$ but more complicated situations may occur.
- a *configuration space*, $X = \{x : V \to S\} = S^V$, containing all the microscopic states of the system.
- a family of real variables, $J = (J_{ij})_{(ij)\in E}$, indexed by the synapses, called *synaptic efficiencies*.
- a family of real variables $= (w_i)_{i\in V}$, indexed by the neurones, called *activation thresholds*.
- a family of *post-synaptic potentials*, indexed by the neurones, defined by
$$h_i = \sum_{j\in V:(ji)\in E} x_j J_{ji} - w_i.$$
- a family of *transfer functions* $f = (f_i)_{i\in V}$, indexed by the synapses, that serves to update the value of the configuration at every neurone. This can be done either in a deterministic way — and in that case the transfer function, $f_i : \mathbb{R} \to S$, assigns a new state to the neurone i given the post-synaptic potential, h_i, that excites it by $x_i = f_i(h_i)$ —, or in a stochastic way — and in that latter case the transfer function $f_i : S \times \mathbb{R} \to [0, 1]$ assigns a value to the conditional probability $\mathbb{P}(x_i = s|h_i = \eta)$ for $s \in S$ and $\eta \in \mathbb{R}$, through $\mathbb{P}(x_i = s|h_i = \eta) = f_i(s, \eta)$.

Definition 1.1 We call *neural net* the system (G, X, J, w, h, f), *i.e.* an oriented graph G, a configuration space X, two families of real variables — the synaptic efficiencies J and the activation thresholds w — a family of post-synaptic potentials h and a family of transfer functions f. According to the mode of transfer, the net is termed deterministic or stochastic.

Remark: The graph G being simple and oriented, there exists a natural order induced by the orientation. For finite graphs G, we can stratify the set of vertices in the following manner:

$$V_0 = \{i \in V : \forall j \in V, (ji) \notin E\};$$

this set represents the *sensor neurones* that receive the *external stimuli*. The intermediate strata are defined recursively, by

$$V_k = \{i \in V \setminus (V_0 \cup V_1 \cup \cdots \cup V_{k-1}) : \exists j \in V_{k-1} \text{ such that } (ji) \in E\}.$$

The last layer, $L = \sup\{k : V \setminus \cup_{i=0}^{k} V_i) \neq \emptyset\} + 1$, gives rise to the stratum V_L of *motor neurones* that communicate to the external world the result of the computation. This stratification of the vertex set induces a natural stratification of the configuration space $X = \oplus_{l=0}^{L} X_l$, with $X_l = \{x : V_l \to S\} = S^{V_l}$.

The configuration over the set V_0 must be fixed *ad hoc* but the configuration over all the other strata is determined by the neural network updating rules. This updating can be done either in a *synchronous* way, when whole batches of neurones change simultaneously their internal configuration according to the post synaptic potential they receive from their "parent" neurones, or in an *asynchronous* way, when to each neurone is attached an internal clock — independent of the clock of the other neurones — that commands the moment of updating. In the case of a synchronous net, the time evolution can be studied as a discrete time dynamical system; for asynchronous nets, continuous time evolution is more appropriate.

Given a configuration $x \in X_0$ at the sensor layer, the net returns a configuration $y \in X_L$ at the motor layer, implementing thus a mapping $F : X_0 \to X_L$ that is completely determined by $x \mapsto y = F(x)$ in terms of the set of parameters $J = (J_{ij})_{(ij)\in E}$ and $w = (w_i)_{i\in V}$. It will be convenient in the sequel to consider the set of all possible parameters J and w as a space Θ, generic points, $\theta = (J, w) \in \Theta$, of which are meant to represent a given realisation of the network. This space is termed *control space* and it will be eventually equipped with a probability measure. The choice of the θ, fixing the parameters of the network, controls (defines) completely the map F. To stress this control, we write, when necessary, F_θ for this map.

At this level of generality, it is difficult to implement the network, to understand its functioning, and to decide whether it is advantageous to use a neural computer versus a conventional one. We must therefore specify the network more precisely: this will be done by studying particular examples.

2 Examples of neural networks

A whole spectrum of neural networks is introduced; they are classified according to the nature of their state space, their architecture, and their transfer function.

Since implementation on a digital computer proceeds always by discretisation, very often, it is enough to consider binary networks, *i.e.* $S = \{-1, 1\}$. These networks have a configuration space reminiscent of the Ising configuration space in statistical mechanics. However, some more general single-neurone internal states space are used. For instance, q-states Potts neural nets [33] have $S = \{0, \cdots, q-1\}$, XY-neural nets [89] have $S = \mathbb{T}^1$, and so on.

The second characteristic serving to classify the nets is their architecture. This is defined mainly by the edge structure of the underlying graph. However, architecture can also be a dynamical characteristic. As a matter of fact, edges serve to index the synaptic efficiencies. Now, if for a given graph $G = (V, E)$, the synaptic efficiencies $J_{ij} \not\equiv 0$ only for all $\{ij\} \in E' \subset E$, it is the set E' that defines the architecture and not the set E. In the same spirit, the architecture can be even a random characteristic as it may happen in the case of randomly diluted nets [15], where each synaptic efficiency is multiplied by a random variable that can take values 0 or 1 independently on every edge of the graph, leading thus to a bond percolation cluster sub-graph of E.

The nature of the transfer function must also be taken into account. In view of practical applications, it is convenient to be able to parallelise the computations; therefore it is natural to consider the same transfer function function all over the network. For deterministic systems, we can choose the non-linear function $f : \mathbb{R} \to S$ defined by

$$f(s) = \mathrm{sgn}(s) = \begin{cases} 1 & \text{if} \quad s \geq 0 \\ -1 & \text{if} \quad s < 0. \end{cases}$$

In several applications some continuity properties are required; in that case instead of the sharp sgn function a smoothed sigmoidal version can be used, for instance $f(s) = \tanh(\beta s)$, for some real parameter β.

Finally, the synchronisation must be specified to decide whether the updating follows a synchronous or an asynchronous udpating schedule. In case of synchronous updating and for some architectures, nets may have different evolution depending on their parallel or sequential updating [63, 79, 80, 103].

Having these characteristics in mind we present some of the most commonly studied neural nets.

2.1 The neural net of McCulloch and Pitts

This model [65], introduced in 1943, is composed by a semi-infinite repetition of N neurones, *i.e.* its vertex set is $V = \{1, \ldots, N\} \times \mathbb{N}$ and its space state $S = \{-1, 1\}$ is binary. The particular structure of the underlying graph allows to represent each vertex $v \in V$ as $v = (i, t)$ with $i \in \{1, \ldots, N\}$ and $t \in \mathbb{N}$. The latter index is interpreted as time. If $v = (i, t)$ and $v' = (j, t+1)$, the corresponding synaptic efficiency $J_{v,v'}$ is denoted $J_{ij}^{(t+1)}$ and the post-synaptic potential, $h_{v'} \equiv h_j(t+1)$, is expressed in terms of the configurations

by

$$h_j(t+1) = \sum_{i=1}^{N} x_i(t) J_{ij}^{(t+1)} - w_j^{(t+1)}.$$

The computation is then expressed in terms of the discrete time dynamics

$$x_j(t+1) = \mathrm{sgn}(h_j(t+1)).$$

Sequential updating is used to induce a dynamical evolution $X^{(t)} \to X^{(t+1)}$ where $X^{(t)} = S^{V_t}$ and V_t is the set of neurones involved at time t. The stratum S^{V_t} of the configuration space being isomorphic to S^N for every t, one time step of the network is a mapping $T : S^N \to S^N$ and the evolution can be regarded as a trivial (*i.e.* deterministic) Markov chain on S^N defined for $a, b \in S^N$ by

$$p_{ab} = \mathbb{P}(Y_{n+1} = b | Y_n = a) = \begin{cases} 1 & \text{if } b = Ta \\ 0 & \text{otherwise.} \end{cases}$$

The reason for introducing such a trivial probabilistic object is that it can be easily generalised to tackle stochastic dynamics introduced in subsequent sections. When $N \to \infty$, the previous evolution is a discrete dynamical system $[0,1] \to [0,1]$. We are interested in the asymptotic behaviour of T^t when $t \to \infty$; (for the mathematical treatment of such evolutions the reader may consult [23] for instance). It has been shown that the McCulloch and Pitts network is computationally equivalent to a Turing machine. It is not however evident that it offers a more efficient alternative to the computation of general functions than a universal computer.

2.2 The simple perceptron

This is the most elementary neural net, introduced [86] in 1962. It is a binary, *i.e.* $S = \{0,1\}$, synchronous net over a finite bipartite graph whose vertex set is $V_0 \oplus V_1$ with $V_0 = \{1, \cdots, N\}$ and $V_1 = \{1\}$, *i.e.* there are N sensor neurones and one motor neurone. The edge set is $E = \{(i,j) : i \in V_0, j \in V_1\}$ and the transfer function is plainly $f = \frac{1+\mathrm{sgn}}{2}$. Therefore, the simple perceptron implements a Boolean function of N inputs and one output.

A fundamental question, for using neural computing, is whether all Boolean functions can be implemented by a single perceptron *i.e.* whether, for every Boolean function with N entries, $G_k : S^N \to S$, with $k = 1, \cdots, 2^{2^N}$, is it possible to chose the control parameters $\theta^{(k)}$ (synaptic efficiencies and activation thresholds) of simple perceptron so that

$$G_k \equiv F_{\theta^{(k)}}, \quad \forall k = 1, \cdots, 2^{2^N}.$$

The answer to this question is negative as it can be shown for the exclusive or (`XOR`) Boolean function of two entries. In fact, for the simple perceptron, the input configuration space X_0 has a vector space structure and this endows the control space $\Theta \equiv X_0'$ with the dual linear structure. In other words, a choice θ of control parameters defines a linear

functional $\theta \in \Theta \equiv X_0'$ and the post-synaptic potential is the action of this functional to the entry configuration[2]

$$h = \sum_{i=1}^{N} J_i x_i = \langle\, \theta, x \,\rangle.$$

Definition 2.1 Let A and B be two subsets of the space X_0. We say they are *linearly separable* if there exists a linear functional $\theta \in X_0'$ such that

$$\forall x \in A, \forall y \in B, \quad \langle\, \theta, x \,\rangle < \langle\, \theta, y \,\rangle.$$

If no such linear functional exists, the sets are called non-separable.

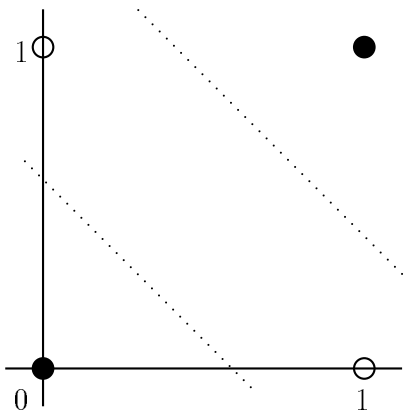

Figure 1: The figure gives the truth table of the `XOR` function. Full blob (or number 0) stands for `false` and empty blob (or number 1) for `true`. It is evident that there is no single straight line splitting the plane into two regions containing single colour points; two such lines are necessary. The truth space of the `XOR` function is not linearly separable.

It is immediate to see (figure 1) that the truth table of the Boolean function `XOR` with two entries splits the space X_0 into two regions needing two linear functionals to be separated instead of just one. As a consequence, this function cannot be implemented by a single perceptron. This remark was at the origin of the oblivion into which felt the neural computing for more than two decades.

2.3 The multilayered perceptron

The multilayered perceptron [88] is a binary neural net with sharp transfer function. It is composed by a finite simple oriented graph of processors arranged in various layers

[2]The activation threshold can be incorporated in this writing by adding a $N+1$ input fixed to the value 1.

$l = 0, \ldots, L$; each layer is composed by N_l neurones so that V_l is isomorphic to $\{1, \ldots, N_l\}$ and the configuration space, X, is stratified, $X = \otimes_{l=0}^{L} X_l$, where

$$X_l = \{x : V_l \cong \{1, \ldots, N_l\} \to S\} = S^{N_l}.$$

Only edges connecting processors of a given layer, l, with processors of the next layer, $l+1$, appear and we denote by $J_{ij}^{(l+1)}$ the synaptic efficiency between the site i of the l-th later with the j site of the $l+1$-th layer. Thus the post-synaptic potential reads

$$h_j^{(l+1)} = \sum_{i \in V_l} x_i J_{ij}^{(l+1)} - w_j^{(l+1)}$$

and the deterministic dynamics

$$x_j^{(l+1)} = \mathrm{sgn}(h_j^{(l+1)}), \quad \text{for} \quad j \in V_{l+1}.$$

The interest of the multilayered perceptron stems on the fact that it can implement all Boolean function with an arbitrary number of entries. More precisely, it shown (see [70] for instance) the following

Theorem 2.2 *Let $F : \{-1, 1\}^N \to \{-1, 1\}$ be a Boolean function with N entries and one output. Then, there exists a two layered network with binary neurones, with a layer of N sensor neurones, a hidden layer of 2^N neurones, and a layer of one motor neurone that implements the function F without error.*

Remark: What remains an open question however is whether there is an optimal size for the intermediate layer not saturating the bound required by the previous theorem. For specific Boolean functions the answer is affirmative; for instance, the function `XOR` with two entries can be implemented on a two layered network with only two intermediate neurones instead of the four required by the existence theorem. It is not clear whether this bound is really non saturated for a general Boolean function with an arbitrary number of inputs.

2.4 Fully connected committee machine

This network was introduced in [72]. It is a special case of a two-layered perceptron with binary neurones and sharp transfer functions. The input layer has N sensor neurones, the intermediate layer has K hidden neurones and the output layer has one motor neurone. There is a first class edge connecting every input neurone to every hidden neurone and a second class edge connecting every hidden neurone to the output neurone. The first class edges carry synaptic efficiencies J_{ij} with $i = 1, \cdots, N$ and $j = 1, \cdots, K$ so that

$$x_j = \mathrm{sgn}(\sum_{i=1}^{N} x_i J_{ij}), \quad i = 1, \cdots, N \quad j = 1, \cdots, K$$

and the second class of edges carry synaptic efficiencies that are all equal to one so that

$$x = \mathrm{sgn}\left(\sum_{j=1}^{K} \mathrm{sgn}(\sum_{i=1}^{N} x_i J_{ij})\right).$$

Thus, a fully connected committee machine [1, 91] is a fully connected simple perceptron of N entries and K outputs feeding a K entries majority rule voting machine[3].

2.5 XY-networks

This is a continuous state space network with $S = \mathbb{T}^1$ introduced in [55]. The vertex set is $V = \{1, \cdots, N\} \times \mathbb{N}$. Thus, as it is the case for the McCulloch and Pitts network, sites have a spatial and a temporal component, namely $v = (i, t)$, with $i \in \{1, \cdots, N\}$ and $t \in \mathbb{N}$. Edges connect sites of a given layer with all sites of the next layer. Each neurone is attached an activation threshold $w_i(t)$ and each edge a synaptic efficiency $J_{ij}^{(t+1)}$. The novel feature of this network is the way the updating is performed:

$$x_i(t+1) = \sum_{j=1}^{N} J_{ij}^{(t+1)} \sin(x_i(t) - x_j(t)) - w_i^{(t+1)} \mod 2\pi.$$

Usually, the activation threshold of this model is a Gaussian random variable independent from site to site.

This network has also a continuous time version [89] for asynchronous udpating, leading in an underlying space-time structure of the form $V = \{1, \cdots, N\} \times \mathbb{R}^+$. In this case the updating is through the stochastic differential equation

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{N} J_{ij}(t) \sin(x_i(t) - x_j(t)) - w_i(t) \mod 2\pi.$$

Notice however that in this case the model resembles more to a field-theoretical model and less to a statistical mechanical one. In the usual case of random activation threshold, $w_i(t)$ is a white noise generalised random process, independent for different i's.

2.6 Short range finite-dimensional network

For layered networks, it is convenient to interpret a given layer as the spatial extent of the network and the passage from a layer to the next one as a time evolution. The common feature of all the nets introduced so far is that all neurones of a given layer intervene to define the state of a single neurone of the next layer; in this sense the previous models are long-range — as a matter of fact mean-field models (see [82] for instance for definitions) — in the statistical mechanics terminology.

[3] To avoid any indeterminacy, we can chose K to be odd.

Another possibility should be to define a layered network $V = \oplus_{l\in\mathbb{N}} V_l$ where each layer V_l is isomorphic to a finite subset of a regular D-dimensional lattice,

$$V_l \cong [-N, N]^D \cap \mathbb{Z}^D.$$

The novel feature is the "short range" edge structure. In this model, a pair of vertices (v, v'), where $v = (i, t)$ and $v' = (i', t')$), with $i, i' \in [-N, N]^D \cap \mathbb{Z}^D$ and $t, t' \in \mathbb{N}$, belongs to the edge set E if, and only if, $t' = t+1$ and $|i - i'| = 1$, where $|\cdot|$ denotes the Euclidean distance in $\mathbb{Z}^D$. The simplest network has binary neurones but more complicated nets can be defined. The edge set indexes a family of synaptic efficiencies and the (synchronous) update is performed according to the usual formula. Such short range models are studied in [73] or [19].

2.7 Randomly diluted networks

These networks can be defined as a modification of any of the previously defined nets. Edges of the underlying graph index not only a family of synaptic efficiencies but also a family of independent identically distributed random variables $(c_e)_{e\in E}$, taking values in $\{0, 1\}$. The effective synaptic efficiency is $J_e c_e$. Therefore, the edge set that really contributes to the network architecture is the set $E_1 = \{e \in E : c_e = 1\} \subset E$. Since this procedure is equivalent in erasing some edges, the resulting network is termed randomly diluted.

3 Learning algorithms

We have seen that a neural net on a finite simple graph is the implementation of a mapping $F : X_0 \to X_L$ depending on the particular realisation $\theta \in \Theta$ of the control parameters. To stress out this dependence, we shall use in the sequel the notation F_θ for the mapping implemented by the realisation θ.

In this section, we address the inverse problem; namely the problem of choosing the realisation of the control parameters so that the network implements a given mapping, provided the implementation is possible. The direct problem will be discussed in the following sections.

3.1 Supervised learning

The most convenient description is the statistical one. The space Θ contains all possible realisations of the network. In the absence of any additional information, the parameters can have arbitrary values. In general, we can assume that in such a situation there is an

a priori measure μ_0 on Θ, that is reasonable to choose non-atomic and having support on the whole space Θ.

Supervised learning will be interpreted as a modification of this measure μ_0 in such a way that it will become concentrated on smaller and smaller sets [101]. To be more specific, we stick to the deterministic multilayered perceptron with binary neurones and constant transfer function $f = \mathrm{sgn}$. For a given realisation θ, we denote, as usual, by F_θ the mapping implemented by the network. We identify in the sequel any mapping $g : X_0 \to X_L$ with its graph

$$g \equiv \{\Xi = (x, y) \in X_0 \times X_L : \Xi = (x, g(x))\}.$$

Definition 3.1 A *training set* for the mapping $g : X_0 \to X_L$ is a finite subset, L_g, of the graph of g, *i.e.*

$$L_g = \{\Xi^\alpha = (x^\alpha, y^\alpha) : y^\alpha = g(x^\alpha)\}_{\alpha=1,\ldots,A}.$$

We shall say that the network has been totally trained by the training set L_g if the control parameters have been adjusted in such a manner that[4] $F_\theta \mid_{L_g} = g \mid_{L_g}$.

Such a total training is however very demanding in terms of time; moreover, it reduces the generalisation capabilities of the network. It is much more efficient to allow for a small number of errors. To be more specific, denote by d a natural distance on the configuration space stratum X_L, for instance the Hamming distance, defined for every two configurations x and y as the number of sites where they differ

$$d(x, y) = \sum_{i=1}^{N_L} \frac{(x_i - y_i)^2}{4},$$

and define a map $\mathcal{H}_A : \Theta \to \mathbb{R}^+$ by

$$\mathcal{H}_A(\theta) = \sum_{\alpha=1}^{A} d(y^\alpha, F_\theta(x^\alpha)).$$

The function is called *total learning error* and attains its minimal value, 0, when the net is totally trained on L_g. Otherwise it takes a positive value that counts the number of errors between the graphs of g and of F_θ for the specific realisation of the control parameters. Any sensible learning algorithm can be viewed as an algorithm searching for the minimum of $\mathcal{H}_A$.

3.2 Deterministic learning algorithms

Deterministic algorithms are totally specified sequences of control parameters $(\theta_n)_{n\in\mathbb{N}}$ exploring a subset of Θ. It is expected that such sequences converge towards the global

[4] A slightly abusive notation is used here; restriction of a function on L_g actually meaning restriction to the set of the first co-ordinates of points composing L_g.

minimum of $\mathcal{H}_A$, *i.e.* that

$$\lim_{n \to \infty} \theta_n = \overline{\theta} = \arg\min_{\theta \in \Theta} \mathcal{H}_A(\theta).$$

However, for such a convergence to hold, some additional conditions are needed. Typically, convexity and differentiability of $\mathcal{H}_A$ are required. Differentiability can be replaced by continuity and subdifferentiability; on the contrary, convexity is a very important condition.

3.2.1 Gradient and subgradient algorithms and error back-propagation

Let $\mathcal{H} : \Theta \to \mathbb{R}^+$ be a numerical function, defined on a finite-dimensional vector space Θ, which we want to minimise. Suppose moreover that $\mathcal{H}$ is differentiable. Obviously, if $\overline{\theta} = \arg\min_{\theta \in \Theta} \mathcal{H}_A(\theta)$, then $\mathcal{H}'(\overline{\theta}) = 0$. The *gradient algorithm* reads [8]

Algorithm 3.2 (Gradient minimisation algorithm)
FIX some numerical sequence $(\delta_n)_{n \in \mathbb{N}}$ of positive numbers.
INITIALISE $n \leftarrow 0$
CHOOSE some arbitrary $\theta_0 \in \Theta$.
REPEAT UNTIL $\mathcal{H}'(\theta_n) = 0$
{ $\theta_n \leftarrow \theta_n - \delta_n \frac{\mathcal{H}'(\theta_n)}{\|\mathcal{H}'(\theta_n)\|}$
$n \leftarrow n + 1.$
}

This algorithm converges in general to local minima. We have however the following

Theorem 3.3 *Let the sequence $(\delta_n)_{n \in \mathbb{N}}$ of the previous algorithm be chosen so that*

1. $\lim_{n \to \infty} \delta_n = 0$,
2. $\sum_{n \in \mathbb{N}} \delta_n = +\infty$,

and assume that $\mathcal{H} : \Theta \to \mathbb{R}$ is a convex, differentiable, bounded from below function. Then the sequence

$$y_k = \min_{n=0,\ldots,k} \mathcal{H}(\theta_n),$$

where (θ_n) is the sequence defined in the previous algorithm, converges to the infimum of $\mathcal{H}$, namely,

$$\lim_{k \to \infty} y_k = \inf_{\theta \in \Theta} \mathcal{H}(\theta).$$

Now, the differentiability of the function$\mathcal{H}$ is a very stringent condition; very often it can be verified for a distance used to define $\mathcal{H}$ and can fail for a slight modification of the distance function. It is therefore wishable to get rid of this condition and to replace it by less restrictive and more stable conditions of continuity and subdifferentiablity. Recall that gradient, when it exits, is a linear form on the tangent space. According to [85], define then

Definition 3.4 Let $\mathcal{H} : \Theta \to \mathbb{R}$ be a convex function. A vector θ^* is called *subgradient* of $\mathcal{H}$ at θ if

$$\mathcal{H}(\theta') \geq \mathcal{H}(\theta) + \langle\, \theta^*, \theta' - \theta \,\rangle, \quad \forall \theta' \in \Theta.$$

The set of all sugradients of $\mathcal{H}$ at θ is called the *subdifferential* at θ and is denoted $\partial\mathcal{H}(\theta)$.

For continuous functions, the minimisation algorithm becomes

Algorithm 3.5 (Subgradient minimisation algorithm)
FIX some numerical sequence $(\delta_n)_{n\in\mathbb{N}}$ of positive numbers.
INITIALISE $n \leftarrow 0$
CHOOSE some arbitrary $\theta_0 \in \mathrm{Dom}(\mathcal{H})$ and some subgradient $p_0 \in \partial\mathcal{H}(\theta_0)$.
REPEAT UNTIL $p_n = 0$
{ CHOOSE some $p_n \in \partial\mathcal{H}(\theta_n)$
$\theta_n \leftarrow \theta_n - \delta_n \frac{p_n}{\|p_n\|}$
$n \leftarrow n + 1$.
}

The convergence of this algorithm is guaranteed by the following

Theorem 3.6 *Let the sequence $(\delta_n)_{n\in\mathbb{N}}$ of the previous algorithm be chosen so that*

1. $\lim_{n\to\infty} \delta_n = 0$, *and*
2. $\sum_{n\in\mathbb{N}} \delta_n = +\infty$.

Assume that $\mathcal{H} : \Theta \to \mathbb{R} \cup \{+\infty\}$ is a convex, bounded from below function and that the interior of $\mathrm{Dom}(\mathcal{H})$ *is non empty. Then the sequence*

$$y_k = \min_{n=0,\ldots,k} \mathcal{H}(\theta_n),$$

where (θ_n) is the sequence defined in the previous algorithm, converges to the infimum of $\mathcal{H}$, namely,

$$\lim_{k\to\infty} y_k = \inf_{\theta\in\Theta} \mathcal{H}(\theta).$$

Proof: See [8] for instance. □

Application of these minimisation algorithms gives rise to the so called *error back propagation adaptation scheme*. To illustrate the method, consider a $L+1$-layered perceptron with N neurones at every layer and a smooth transfer function f. Recall that we want to minimise

$$\begin{aligned}\mathcal{H}_A(\theta) &= \sum_{\alpha=1}^{A} d\left(y^\alpha, F_\theta(x^\alpha)\right) \\ &= \frac{1}{4}\sum_{\alpha=1}^{A}\sum_{i=1}^{N}\left(y_i^\alpha - F_\theta(x^\alpha)_i\right)^2.\end{aligned}$$

Assume to simplify notation that the activation thresholds are identically vanishing all over the network. Starting from a given initial realisation of the control parameters,

$$\theta = (J_{ij}^{(l)})_{i,j=1,\ldots,N}^{l=0,\ldots,L},$$

we wish to follow the gradient algorithm, that is to modify synaptic efficiences according to the formula

$$J_{ij}^{(l)} \leftarrow J_{ij}^{(l)} - \delta_1 \frac{H_{ij}^{(l)}}{|H_{ij}^{(l)}|}$$

where

$$H_{ij}^{(l)} = \frac{\partial \mathcal{H}_A(\theta)}{\partial J_{ij}^{(l)}}.$$

The practical problem to solve therefore is to compute the partial derivative in $H_{ij}^{(l)}$. To update the control parameters, it proves convenient to start from the last layer L and continue backwards to the 0-th layer, hence the name error back propagation, namely

$$\begin{aligned}H_{ij}^{(L)} &= \frac{\partial \mathcal{H}_A(\theta)}{\partial J_{ij}^{(L)}} \\ &= \frac{1}{4}\sum_{\alpha=1}^{A}\frac{\partial}{\partial J_{ij}^{(l)}}\sum_{k=1}^{N}\left(y_k^\alpha - F_\theta(x^\alpha)_k\right)^2 \\ &= -\frac{1}{2}\sum_{\alpha=1}^{A}\left(y_j^\alpha - f(h_j^{\alpha,(L)})\right) f'(h_j^{\alpha,(L)}) x_i^{\alpha,(L-1)},\end{aligned}$$

where

$$h_j^{\alpha,(L)} = \sum_{k=1}^{N} x_k^{\alpha,(L-1)} J_{kj}^{(L)}$$

is the post-synaptic potential due to the 0-th layer configuration x^α. After the modification on $J_{ij}^{(L)}$ is completed for the L-th layer, the procedure is restarted with the layer $L-1$ and then again up to the first layer. It is thus established that gradient descent search leads to the error back propagation learning algorithm [45, 53].

3.2.2 Newton algorithm and progressive learning

First recall the principle of Newton's method for the search of the zero of a smooth function $\mathcal{H} : \Theta \to \mathbb{R}$, under the assumption that $\mathcal{H}'(\theta)$ is invertible. The methods reads

Algorithm 3.7 (Newton's one-dimensional method)
INITIALISE $n \leftarrow 0$
CHOOSE $\delta > 0$ and $\theta_0 \in \Theta$.
REPEAT until $\mathcal{H}(\theta_n) = 0$
{ $\theta_n \leftarrow \theta_n - \delta \frac{\mathcal{H}(\theta_n)}{\mathcal{H}'(\theta_n)}$
$n \leftarrow n + 1$.
}

Proof: It is an elementary exercise (see exercise 6, page 380 of [57] for instance) to show that under C^2 and boundedness conditions, the above sequence converges to a zero, *i.e.*

$$\lim_{n \to \infty} \theta_n = \overline{\theta}$$

with

$$\mathcal{H}(\overline{\theta}) = 0.$$

□

The method can be easily generalised to the multi-dimensional case with only notational complications (the derivative is now a linear operator on the tangent space).

Turn now to the specific problem we have to solve, namely adjust the control parameters θ so that the net becomes totally trained on some training set. In other words, we have to find a realisation $\overline{\theta}$ such that $\mathcal{H}_A(\overline{\theta}) = 0$. Now,

$$\begin{aligned} \mathcal{H}_A(\theta) &= \sum_{\alpha=1}^{A} d\left(y^\alpha, F_\theta(x^\alpha)\right) \\ &= \frac{1}{4} \sum_{\alpha=1}^{A} \sum_{i=1}^{N} \left(y_i^\alpha - F_\theta(x^\alpha)_i\right)^2, \end{aligned}$$

and since every term in the above sum is positive, it corresponds to an additional constraint on the control parameters. Starting from an arbitrary realisation $\theta_0 = (J_{ij}^l)$, change the parameters by $\Delta^1 J_{ij}^l$, where

$$\Delta^1 J_{ij}^l = \frac{\delta}{4} \left[\frac{\partial}{\partial J_{ij}^l} \sum_m \left(y_m^1 - F_\theta(x^1)_m\right)^2 \right]^{-1} \sum_{i=1}^{N} \left(y_i^1 - F_\theta(x^1)_i\right)^2,$$

for $i, j = 1, \ldots, N$ and $l = 0, \ldots, L$. After this change, the parameters are "closer" to those corresponding to a net having learnt the first example. Now, repeat the same procedure with the other examples

$$\Delta^\alpha J_{ij}^l = \frac{\delta}{4} \left[\frac{\partial}{\partial J_{ij}^l} \sum_m \left(y_m^\alpha - F_\theta(x^\alpha)_m\right)^2 \right]^{-1} \sum_{i=1}^{N} \left(y_i^\alpha - F_\theta(x^\alpha)_i\right)^2,$$

for $i, j = 1, \ldots, N$ and $l = 0, \ldots, L$ and for $\alpha = 2, \ldots, A$. When all the examples are scanned consider the realisation

$$\theta_1 = (J_{ij}^l + \sum_{\alpha=1}^{A} \Delta^\alpha J_{ij}^l)_{ij}^l$$

as the new starting point and iterate. This algorithm corresponds to a *progressive learning* procedure.

3.3 A digression: need for new algorithms

We have seen that the two previous deterministic algorithms converge under rather stringent conditions, convexity and continuity for the subgradient algorithm, derivability and inversibility of the derivative operator for the Newton's algorithm. For small nets we can hope or even explicitly check that such conditions hold. For large nets however, it is unreasonable even to think that such conditions can be valid. Experience from the area of spin glasses and numerical evidence show that they almost surely fail.

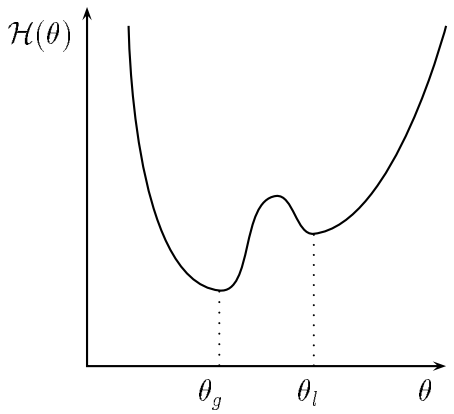

Figure 2: When the function to minimise is not convex, depending on the initial point θ_0, the algorithm will eventually converge either to the local minimum θ_l or the global one θ_g.

To be more specific, consider the trivial counter-example, depicted in figure 2, of a smooth function, bounded from below, but not convex. It is immediate to see that depending on the starting point for the sequence

$$\theta_{n+1} = \theta_n - \delta_n \frac{\mathcal{H}'(\theta_n)}{\|\mathcal{H}'(\theta_n)\|}$$

the algorithm converges to a global or a local minimum. Intuitively, the sequence θ_n can be thought as the positions of a ball without kinetic energy falling with friction into the

potential well defined by the function $\mathcal{H}$. The ball gets trapped to the first local minimum encountered. As a matter of fact, this trapping phenomenon is a common feature of all deterministic algorithms. For realistic functions $\mathcal{H}$ we need therefore some new algorithms, improved by stochastic considerations, like *stochastic gradient* or *simulated annealing*.

3.3.1 Stochastic gradient algorithms

For simple cases as the one exemplified above, an alternative is provided by the stochastic gradient algorithm: it is a stochastic process $(X_n)_{n\geq 0}$ with values in the space Θ, defined by the sequence

$$X_{n+1} = X_n - \delta_{n+1}\mathcal{H}'(X_n) + \epsilon_{n+1}\xi_{n+1}$$

where $(\xi_n)_{n\geq 1}$ is a sequence of independent identically distributed variables and (δ_n) and (ϵ_n) are two deterministic (or previsible with respect to the natural filtration $\sigma(\xi_1, \ldots, \xi_n)$) sequences tending slowly to zero. When the speed of convergence to zero of these sequences is properly chosen, the stochastic process (X_n) converges to a random variable X_∞, distributed according to a law charging with large probability small intervals containing in their interior the absolute minimum of $\mathcal{H}$. We don't wish to give further details on the convergence criteria since even this algorithm has a limited field of applications in small nets. It is only useful to remark that the white noise perturbation ξ_n to the dynamical system corresponds to random kicking the ball permitting thus to get out of the local minima. These kicks have to eventually vanish however since otherwise they could take the ball outside the basin of attraction of the global minimum and thus prevent the process from converging. This precise balance between the "downwards" driving term δ_n and the random kicking term ϵ_n gives essentially the condition of convergence (see [30] for precise statements).

3.3.2 Simulated annealing

The stochastic minimisation algorithm known as *simulated annealing* proceeds [7, 41] by changing the *a priori* measure μ_0 in a way exponentially suppressing the sets of parameters leading to a large learning error [81]. This is achieved by introducing a positive parameter β, interpreted as the inverse temperature, and defining

$$\mu_{A,\beta}(d\theta) = \frac{\exp(-\beta\mathcal{H}_A(\theta))}{\mathcal{Z}_A(\beta)}\mu_0(d\theta),$$

where

$$\mathcal{Z}_A(\beta) = \int_\Theta \exp(-\beta\mathcal{H}_A(\theta))\mu_0(d\theta).$$

Denote by $\Theta^A_{\min} = \{\theta \in \Theta : \mathcal{H}_A(\theta) = 0\} \subset \Theta$ the set composed from the network realisations minimising the total training error. When $\beta \to \infty$, it is intuitively clear that $\mu_{A,\beta}$ charges solely $\Theta_{\min}$. Of course, taking the limit $\beta \to \infty$ at this stage removes any computational advantage of the formalism since we recover the minimisation problem we

had started with. Instead, it is much more efficient to use *simulating annealing algorithm* for attaining the minimum.

In mathematical terms, simulating annealing is the choice of a *cooling schedule* — *i.e.* a monotonically diverging sequence $(\beta_n)_{n\geq 1}$ of inverse temperatures $\beta_n \uparrow \infty$ — and the construction of an *irreducible inhomogeneous Markov chain* on Θ with transition probability kernel $p_{\beta_n}(\cdot,\cdot) : \Theta \times \mathcal{B}(\Theta) \to [0,1]$, indexed by the sequence $(\beta_n)_{n\geq 1}$. The first requirement is that for fixed n, the measure μ_{A,β_n} must be an invariant measure for the Markov evolution *i.e.* it must be the left normalised eigenvector corresponding to the eigenvalue one for the Markov operator

$$\int_\Theta \mu_{A,\beta_n}(d\theta_1) p_{\beta_n}(\theta_1,\cdot) = \mu_{A,\beta_n}(\cdot).$$

This can be effectively implemented by standard algorithms like Metropolis, Kawasaki, *etc.* for any fixed β_n. The second requirement concerns the speed of cooling; provided that β_n diverges not very fast when $n \to \infty$, the inhomogeneous Markov chain, having transition kernel p_{β_n} at time n, converges to a measure $\mu_{A,\infty}$ charging only the network realisations minimising the total training error, $\operatorname{supp} \mu_{A,\infty} = \Theta^A_{\min}$. This result is precisely stated in the following

Theorem 3.8 (Simulated annealing) *Let $(Y_n)_{n\geq 0}$ be an inhomogeneous irreducible Markov chain on Θ defined by the transition probability kernel p_{β_n},*

$$\mathbb{P}(Y_{n+1} \in d\theta' | Y_n = \theta) = p_{\beta_n}(\theta, d\theta').$$

Then,

$$\lim_{n\to\infty} \mathbb{P}(Y_n \in \Theta^A_{\min}) = 1$$

if, and only if, the cooling schedule (β_n) verifies

$$\sum_{n=1}^{\infty} \exp(-\beta_n D) = \infty,$$

where D stands for the maximal depth of local minima of $\mathcal{H}_A$.

Proof: See [9]. □

The previous theorem establishes the algorithmic implementability of such a minimisation scheme; moreover, it gives the optimal speed of the cooling schedules since the condition $\sum_{n=1}^{\infty} \exp(-\beta_n D) = \infty$ is satisfied provided $\lim \frac{\log n}{\beta_n} = c$ with $c \geq D$. In practice however such logarithmic cooling schedule proves very time consuming. To the extent of author's knowledge, *all* numerical results for large systems based on simulated annealing that are published in the literature are obtained with exponentially fast cooling schedules!

3.4 Thermodynamic formalism of learning

Beyond the practical algorithmic reasons, this (thermodynamic) formalism allows a thorough understanding of the learning procedure in terms of information theory.

To be more specific and to avoid unnecessary complications, assume that all the neurones are binary, all layers are finite, and, moreover, the parameter set is discrete [95, 101]. Now the finiteness of the sensor and motor layers together with the binary nature of the neurones implies that the set of all possible mappings $\mathcal{M} = \{f : X_0 \to X_L\}$ is discrete and finite. The *a priori* measure μ_0 on Θ induces a measure on $\mathcal{M}$, denoted by the same symbol, by

$$\mu_0(f) = \int \mu_0(d\theta) \mathbb{1}_{\{F_\theta = f\}}, \quad \text{for} \quad f \in \mathcal{M}.$$

Given an arbitrary measure ν on Θ, define its entropy by

$$S(\nu) = -\sum_{f \in \mathcal{M}} \nu(f) \log \nu(f),$$

with the convention $0 \log 0 = 0$. It can be shown (exercise) that $0 \leq S(\nu) \leq \log \operatorname{card} \mathcal{M}$. As usual in information theory, the entropy of the measure ν can be interpreted [48] as the richness of ν or equivalently the *computational diversity* of the network architecture.

Similarly, given two arbitrary measures μ and ν with $\mu \ll \nu$ define the relative entropy by

$$S(\mu|\nu) = \sum_{f \in \mathcal{M}} \nu(f) \log \frac{\mu(f)}{\nu(f)}.$$

Denote finally by $\mathcal{G}_A(\beta)$ the Gibbs free energy, defined by

$$\mathcal{G}_A(\beta) = -\frac{1}{\beta} \log \mathcal{Z}_A(\beta).$$

A straightforward computation of the mean learning error leads to the formula

$$\frac{\partial(\beta \mathcal{G}_A(\beta))}{\partial \beta} = \mathbb{E}\mathcal{H}_A \equiv \int_\Theta \mathcal{H}_A(\theta) \mu_{A,\beta}(d\theta) = \sum_{f \in \mathcal{M}} \mu_{A,\beta}(f) \sum_{\alpha=1}^{A} d(y^\alpha, f(x^\alpha)),$$

identifying thus the average total training error with the internal energy of the corresponding thermodynamic system. Using the trivial observation that on the set $\{F_\theta = f\}$, the learning error reads $\mathcal{H}_A(\theta) = \sum_\alpha d(y^\alpha, f(x^\alpha))$, we obtain the standard relation of thermodynamics

$$\begin{aligned} S_A &\equiv S(\mu_A|\mu_0) \\ &= \sum_{f \in \mathcal{M}} \mu_A(f) \log \frac{\mu_A(f)}{\mu_0(f)} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{f\in\mathcal{M}} \mu_A(f)\left[\log\int_\Theta \exp(-\beta\mathcal{H}_A(\theta))\mathbb{1}_{\{F_\theta=f\}}\mu_0(d\theta) - \log\mathcal{Z}_A - \log\mu_0(f)\right]\\
&= \sum_{f\in\mathcal{M}} \mu_A(f)\left[-\beta\sum_{\alpha=1}^{A} d(y^\alpha, f(x^\alpha)) + \log\mu_0(f) - \log\mu_0(f) - \log\mathcal{Z}_A\right]\\
&= -\beta\mathbb{E}\mathcal{H}_A + \beta\mathcal{G}_A\\
&= \beta(\mathcal{G}_A - \mathbb{E}\mathcal{H}_A).
\end{aligned}
$$

Thermodynamically, this formula relates the entropy with the free and internal energies. From the information point of view, the meaning of the last relation is also clear. First observe that $S_0 = 0$ and that $S_{A+1} \geq S_A$. Relative entropy being interpreted as information gain, this monotonicity implies that, starting from *tabula rasa*, learning new examples from the training set increases the information. Moreover, we compute

$$\frac{\partial S_A}{\partial \mathbb{E}H_A} = -\beta$$

meaning that information increases through a minimisation of the learning error.

The above arguments are strictly valid for finite networks and for training sets that remain small compared to the size of the net. It is intuitively clear that when the training set starts increasing without bounds some *saturation* must occur since, otherwise, an infinite quantity of information should be stored into a finite system! Several numerical studies have been done on this saturation phenomenon and heuristic computations based on replica trick or annealed approximation showed a clear transition from a memory regime to a saturation regime [54, 96]. However a rigorous mathematical treatment of this phenomenon is still an open problem.

4 Neural network as associative memory

To fix ideas, we consider a McCulloch-Pitts network with N neurones at every layer and m particular fixed configurations of X_N, denoted $\xi^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)$ for $\mu = 1, \ldots, m$, called *patterns*. We also use the symbol ξ_i to denote the family of variables $\xi_i = (\xi_1^\mu, \ldots, \xi_N^\mu)$ for $i = 1, \ldots, N$, so that upper indices number different patterns and lower indices number different sites.

4.1 Choice of synaptic efficiences

We wish the net to memorise the patterns and to be able to recall them when *a clue* is presented to it. Now synaptic efficiencies J_{ij} must be *local*, *i.e.* they must depend only on ξ_i and ξ_j. This is a very important general principle; it excludes the possibility that the synapse connecting neurone i with neurone j needs some global information beyond that contained in ξ_i and ξ_j.

A reasonable and convenient additional assumption, but no so crucial as locality, is the *exchange-symmetry*, *i.e.* the constraint $J_{ij} = J_{ji}$. This allows actually to express the network evolution dynamics in terms of *Hamiltonian dynamics*. The most general form [43] of synaptic efficiencies satisfying locality and symmetry is in terms of a *symmetric synaptic kernel* $Q : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ by

$$J_{ij} = \frac{1}{N} Q(\xi_i; \xi_j), \quad \text{for} \quad i, j = 1, \ldots, N.$$

Several forms have been proposed for the synaptic kernel:

1. *Hebb's rule [42]:* the synaptic efficiencies are given by the formula

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{m} \xi_i^\mu \xi_j^\mu = \frac{1}{N} \xi_i \cdot \xi_j,$$

   where $\cdot$ denotes the scalar product of $\mathbb{R}^m$.

2. *Clipped synapses:* we choose here

$$J_{ij} = \frac{1}{N} \phi(\xi_i \cdot \xi_j),$$

   where $\phi(x) = \mathrm{sgn}(x)$.

3. *Truncated synapses:* the choice here is given by

$$J_{ij} = \frac{1}{N} \phi_a(\xi_i \cdot \xi_j),$$

   where, for some positive constant $a \leq N$ the truncating function ϕ_a is defined by

$$\phi_a(x) = \begin{cases} a & \text{for} \quad x > a \\ x & \text{for} \quad |x| \leq a \\ -a & \text{for} \quad x < -a. \end{cases}$$

4. *Inversion-symmetric synapses [43]:* for binary neurones, and for a given site i, the possible values of the vector ξ_i are the extremal points of the hypercube $[-1, 1]^m$ of $\mathbb{R}^m$, namely the set $\{-1, 1\}^m$. Denote by G_m the Abelian group generated by the m inversions

$$\begin{aligned} g_\mu \xi_i &= g_\mu(\xi_i^1, \ldots, \xi_i^{\mu-1}, \xi_i^\mu, \xi_i^{\mu+1}, \ldots, \xi_i^m) \\ &= (\xi_i^1, \ldots, \xi_i^{\mu-1}, -\xi_i^\mu, \xi_i^{\mu+1}, \ldots, \xi_i^m) \quad \text{for} \quad \mu = 1, \ldots, m. \end{aligned}$$

   The group G_m contains 2^m elements and $g^2 = e$ for all $g \in G_m$. We say that a synaptic kernel is *totally inversion-symmetric* if for every $x, y \in \mathbb{R}^m$ and for every $g \in G_m$, it verifies

$$Q(gx, gy) = Q(x, y).$$

   This symmetry provides a nice simplification since the 2^m characters of the group G_m are then eigenvectors of the kernel Q and can therefore serve as a basis for the spectral decomposition of Q from which interesting results can be easily obtained [43, 44, 29].

4.2 Characterisation of patterns

We are interested in obtaining *generic* results holding for a vast class of memorised patterns and not for some very specific ones in the limit when the network becomes very large: we don't expect the obtained results to hold for *all* patterns. We must therefore be able to characterise simply the set of patterns for which the net functions as an associative memory. This can be achieved by defining the patterns as random variables over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and asking whether the results hold almost surely, or in the mean, or with large probability.

Several choices have been used to characterise the possible patterns: independent, correlated, or Gibbsian. In the sequel, we discuss briefly the various possibilities.

4.2.1 Independent patterns

The variables $(\xi_i^\mu)_{i=1,\dots,N}^{\mu=1,\dots,m}$ are independent, identically distributed random variables. For binary neurones we can use random variables with values in $\{-1, 1\}$ but more general value-spaces do not alter the lines of reasoning. This choice has the advantage of being simple to realise and highly non trivial rigorous results can be obtained [13, 17, 28, 38, 49, 50, 61, 71, 83, 99, 103]...

Epistemologically and philosophically it is however questionable whether this choice offers a good modelling of reality. On the other hand, several argue that if we are able to obtain results in that case, we can expect that the net will in fact perform better. Actually, the case of independent patterns is thought as the worst case; experience shows that it is much easier for the human brain to recall a well structured information than a collection of unrelated random facts. This anthropic way of thinking is however unreliable since in the absence of any closed theory of neural computing, it is not granted that neural nets perform in the same way as human brain: this is the fact we wish to establish. Epistemologically and logically it is not acceptable to include such an hypothesis as postulate.

Nevertheless, it is worth exploring the case of independent random variables because we can hope obtaining a mathematically closed theory within a reasonably remote future that could serve as a first approximation of a more general and realistic theory.

4.2.2 Spatially or semantically correlated patterns

At the neurophysiological level, many examples of *spatial correlation* are known in the context of visual information processing [27]. *Semantic correlations*, on the other hand, occur when patterns are splitted into categories and subcategories for the purpose of classification and biological evidence of this phenomenon is given in [68]. Therefore, an intensive effort to model such phenomena was made [40, 69, 100].

For our purposes, instead of considering patterns that are independent and identically distributed symmetric random variables with variance one, we consider families of random variables $(\xi_i^\mu)_{i=1,\dots,N}^{\mu=1,\dots,m}$ over an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a more complicated dependence structure. We always consider variables with $\mathbb{E}\xi_i^\mu = 0$ and $\mathbb{E}[(\xi_i^\mu)^2] = 1$. In the case of spatially correlated patterns, we impose

$$\mathbb{E}(\xi_i^\mu \xi_j^\nu) = \delta^{\mu\nu} C_{ij},$$

where $(C_{ij})_{ij}$ is a $N \times N$ non-trivial matrix of spatial correlations. For semantic correlations, we impose

$$\mathbb{E}(\xi_i^\mu \xi_j^\nu) = \delta_{ij} \tilde{C}^{\mu\nu},$$

where $(\tilde{C}^{\mu\nu})^{\mu\nu}$ is a $m \times m$ non-trivial matrix of semantic correlations.

Several interesting, but not rigorous, results concerning the storage capacity of such models are obtained in [69, 100] where, by use of replica trick, it is shown that the critical capacity for storing correlated patterns exceeds the capacity of storing independent ones. This is a very appealing characteristic of the replica trick calculations and tends to moderate the somehow severe epistemological criticism made in the previous subsesction. It should be interesting to be able to obtain rigorous results for such correlated patterns.

4.2.3 Gibbsian patterns

The next step is to introduce mixed semantic and spatial correlations. To keep the models tractable, some particular form of correlation must be chosen. In that direction, Schlüter and Wagner [92] introduced Gibbsian structure for the patterns. More specifically, the vertex set indexing the configurations (hence the patterns) is now a finite subset, Λ_n, of the D-dimensional lattice $\mathbb{Z}^D$, *i.e.* we have a family of random variables $(\xi_i^\mu)_{i\in\Lambda_n}^{\mu=1,\dots,m}$, over an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, taking values in $\{-1, 1\}$, with $\Lambda_n = [-n, n]^D \cap \mathbb{Z}^D$. We denote by κ the product measure

$$\kappa = (\delta_{-1} + \delta_1)^{\mathbb{Z}^D}$$

on $\{-1, 1\}^{\mathbb{Z}^D}$ and impose a Gibbsian distribution to the random variables ξ given by

$$g_{n,m,\lambda}(\xi \in d\eta) = \frac{\exp(\frac{\lambda}{\sqrt{m}} \sum_{\mu=1}^{m} \sum_{\langle i,j \rangle} \eta_i^\mu \eta_j^\mu)}{z_m(\lambda)} \prod_{i\in\Lambda_n} \prod_{\mu=1}^{m} \kappa(d\eta_i^\mu),$$

and

$$z_m(\lambda) = \int_{\{-1,1\}^{m|\Lambda_n|}} \exp(\frac{\lambda}{\sqrt{m}} \sum_{\mu=1}^{m} \sum_{\langle i,j \rangle} \eta_i^\mu \eta_j^\mu) \prod_{i\in\Lambda_n} \prod_{\mu=1}^{m} \kappa(d\eta_i^\mu),$$

where $\langle i, j \rangle$ means that the sum extends over pairs i, j of sites of Λ_n that are nearest neighbours.

In a very interesting pre-print, Nishimori, Whyte, and Sherrington [73] study the phase diagram of a short-range finite-dimensional network having the short-range 'Hebbian' form synaptic efficiencies, namely

$$J_{ij} = \begin{cases} \frac{1}{N} \sum_{\mu=1}^{m} \xi_i^\mu \xi_j^\mu & \text{if } |i-j| = 1 \\ 0 & \text{otherwise,} \end{cases}$$

in which the memorised patterns are distributed according to the Gibbs measure $g_{n,m,\lambda}$.

4.3 Sequential Hamiltonian dynamics for the Pastur-Figotin-Hopfield network

Among the neural networks, a special *rôle* is played by the Pastur-Figotin (improperly called Hopfield) network [75, 76, 46]. Consider a McCulloch and Pitts network with N binary neurones at every layer and consider m particular fixed configurations of X_N, denoted $\xi^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)$ for $\mu = 1, \ldots, m$. These configurations are called *patterns*. Fix the synaptic efficiencies by the Hebb's rule [42]

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{m} \xi_i^\mu \xi_j^\mu$$

and let all the threshold w_i vanish. The evolution of the network proceeds as usual, through[5]

$$x_i(t+1) = \operatorname{sgn}\left(\frac{1}{N} \sum_{\mu=1}^{m} \sum_{1 \le i < j \le N} \xi_i^\mu \xi_j^\mu x_j(t) \right).$$

For reasons that will become clear in the next section, we also introduce the so called *Hopfield Hamiltonian*

$$H_N(x) = \frac{2}{N} \sum_{\mu=1}^{m} \sum_{1 \le i < j \le N} \xi_i^\mu \xi_j^\mu x_i x_j.$$

It is immediate to verify that if $m = 1$, the pattern $\xi = \xi^1 = (\xi_1^1, \cdots, \xi_N^1)$ is a fixed point of the network dynamics (check!). What is more important however is the functioning of the network as an associative memory; consider in fact the simplest case $m = 1$, and an initial configuration

$$x_i(0) = \begin{cases} \xi_i & \text{for } i \in A \\ -\xi_i & \text{for } i \notin A, \end{cases}$$

where $A \subseteq \{1, \ldots, N\}$ with $\operatorname{card} A < \frac{N}{2}$.

It is immediate to show that $x_i(1) = \xi_i$, $\forall i$, recovering thus the memorised pattern ξ, by presenting to the network a clue, even substantially differing from the memorised

[5] Notice that the factor $\frac{1}{N}$ in this formula is completely irrelevant as far as evolution is concerned. It is there merely for later thermodynamical considerations.

pattern. The situation is of course more complicated in the case of $m > 1$ but proceeds essentially in the same lines. Neural networks being usually very large, important effort is made in the understanding the limit $N \to \infty$ and $m \to \infty$ simultaneously. It turns out that when the number of patterns grows too rapidly with respect to the size of the net, a memory saturation phenomenon appears and the network is unable to recall memorised patterns any more.

The most important results, summarised in the sequel, are obtained for a family (ξ_i^μ) of independent Bernoulli variables.

4.3.1 Asymptotic stability and attraction

We consider a Hopfield network whose size and number of memorised patterns tend eventually to infinity. We have the following

Theorem 4.1 *Let $m = \frac{N}{\gamma \log N}$. Then*

1. *if $\gamma > 6$, then, for $N \to \infty$, the m original patterns are almost surely stable,* i.e.

$$\mathbb{P}[\liminf_N (\cap_{\nu=1}^m \{T\xi^\nu = \xi^\nu\})] = 1$$

2. *if $\gamma > 4$, then*

$$\mathbb{P}[\cap_{\nu=1}^m \{T\xi^\nu = \xi^\nu\}] = 1 - R_N,$$

with $\lim_{N\to\infty} R_N = 0$.

Proof: Fix some pattern ν. Updating the i-th component, involves the estimation of

$$\operatorname{sgn}\left(\sum_{j\neq i} \xi_i^\nu (\xi_j^\nu)^2 + \sum_{\mu\neq\nu} \xi_i^\mu \sum_{j\neq i} \xi_j^\mu \xi_j^\nu\right) = \xi_i^\nu \operatorname{sgn}\left(N - 1 + \xi_i^\nu \sum_{\mu\neq\nu} \xi_i^\mu \sum_{j\neq i} \xi_j^\mu \xi_j^\nu\right).$$

Therefore the pattern ξ^ν is stable over the site i on the event $A^c(N, m, i, \nu)$, where

$$A(N, m, i, \nu) = \{N - 1 + \xi_i^\nu \sum_{\mu\neq\nu} \xi_i^\mu \sum_{j\neq i} \xi_j^\mu \xi_j^\nu \leq 0\}.$$

Random variables ξ_i^μ being bounded, they posses exponential moments; hence the result is obtained by use of Markov inequality. Notice that

$$\begin{aligned}
\mathbb{P}[\cup_{\nu=1}^m \cup_{i=1}^N A(N, m, i, \nu)] &\leq \sum_{\nu=1}^m \sum_{i=1}^N \mathbb{P}[A(N, m, i, \nu)] \\
&\leq \sum_{\nu=1}^m \sum_{i=1}^N \inf_{t>0} \exp[-t(N-1)] \mathbb{E} \exp(-t\xi_i^\nu \sum_{\mu\neq\nu} \xi_i^\mu \sum_{j\neq i} \xi_j^\mu \xi_j^\nu).
\end{aligned}$$

Computing the expectation is performed inductively by first computing conditional expectation with respect to σ-fields generated from adequate subsets of random variables ξ. Introduce in fact the following σ-fields:

$$\mathcal{F}_j^\mu = \sigma\{(\xi_k^\lambda)_{k=1,\cdots,N}^{\lambda=1,\cdots,m;\lambda\neq\mu}\}$$

and

$$\mathcal{F}_j^\mu = \sigma\{(\xi_k^\lambda)_{k=1,\cdots,N}^{\lambda=1,\cdots,m;\lambda\neq\mu}; (\xi_k^\mu)_{k=1,\cdots,N;k\neq j}\}$$

We get

$$\begin{aligned}
\mathbb{E}\exp(-t\xi_i^\nu \sum_{\mu\neq\nu}\xi_i^\mu \sum_{j\neq i}\xi_j^\mu\xi_j^\nu) &= \mathbb{E}\left(\mathbb{E}(\prod_{\mu\neq\nu}\exp(-t\xi_i^\nu\xi_i^\mu\sum_{j\neq i}\xi_j^\mu\xi_j^\nu)|\mathcal{F}^\nu)\right)\\
&= \mathbb{E}\left(\prod_{\mu\neq\nu}\mathbb{E}(\exp(-t\xi_i^\nu\xi_i^\mu\sum_{j\neq i}\xi_j^\mu\xi_j^\nu)|\mathcal{F}^\nu)\right)\\
&= \mathbb{E}\left(\prod_{\mu\neq\nu}\mathbb{E}\left(\mathbb{E}(\prod_{j\neq i}\exp(-t\xi_i^\nu\xi_i^\mu\ \xi_j^\mu\xi_j^\nu|\mathcal{F}_i^\nu)|\mathcal{F}^\nu\right)\right)\\
&= \mathbb{E}\left(\prod_{\mu\neq\nu}\mathbb{E}\left(\prod_{j\neq i}\mathbb{E}(\exp(-t\xi_i^\nu\xi_i^\mu\ \xi_j^\mu\xi_j^\nu|\mathcal{F}_i^\nu)|\mathcal{F}^\nu\right)\right)\\
&= (\cosh t)^{(N-1)(m-1)}.
\end{aligned}$$

Now $\cosh t \leq \exp(t^2/2)$; hence

$$\begin{aligned}
\mathbb{P}[\cup_{\nu=1}^m\cup_{i=1}^N A(N,m,i,\nu)] &\leq \sum_{\nu=1}^m\sum_{i=1}\mathbb{P}[A(N,m,i,\nu)]\\
&\leq \sum_{\nu=1}^m\sum_{i=1}^N\inf_{t>0}\exp[-t(N-1)]\exp[\frac{t^2}{2}(N-1)(m-1)]\\
&= mN\inf_{t>0}\exp[-t(N-1)+\frac{t^2}{2}(N-1)(m-1)]\\
&= mN\exp\left(-\frac{N-1}{2(m-1)}\right).
\end{aligned}$$

The first part of the proof is completed by use of Borel-Cantelli lemma since the choice of $\gamma > 6$ leads to a convergent series $\sum_{N\geq 2}\frac{1}{N(\log N)^2}$. For the weak part of this theorem to be true, it is enough that $\lim_{N\to\infty} mN\exp(-\frac{N-1}{2(m-1)}) = 0$; this happens in fact for $\gamma > 4$. □

The weak part of the previous theorem was formulated in [62] and proved in [64] and together with the almost sure version, in [103]; it guarantees that the original patterns are stable under sequential evolution, provided that their number m grows slightly sub linearly in N. This stability is of course a necessary condition for the network to behave as an associative memory. However, attraction of configurations towards stable patterns is also needed. This is provided by the following

Theorem 4.2 *Let $\rho \in [0, 1/2[$ and $m = (1-2\rho)^2 \frac{N}{\gamma \log N}$. For $\nu = 1, \ldots, m$, let $x^\nu(0)$ be a configuration on the Hamming sphere centred at ξ^ν and of radius ρN. Then,*

1. *if $\gamma > 6$,*
$$\mathbb{P}[\liminf_N (\cap_{\nu=1}^m \{T x^\nu(0) = \xi^\nu\})] = 1$$
2. *if $\gamma > 4$,*
$$\mathbb{P}[\cap_{\nu=1}^m \{T x^\nu(0) = \xi^\nu\}] = 1 - R_N,$$
with $\lim_{N \to \infty} R_N = 0$.

Proof: The proof is developed along the same lines as the proof of the previous theorem by remarking that

$$\mathbb{P}[\{(T x^\nu(0))_i = -\xi_i^\nu\}] \leq \exp(-\frac{(N(1-2\rho)-1)^2}{(m-1)(N-1)} \equiv K_{N,m}(\rho)$$

where $K_{N,m}(\rho)$ is an increasing function of ρ. Observe also that

$$\mathbb{P}[\cup_{\nu=1}^m \{T x^\nu(0) = \xi^\nu\}] \leq mN K_{N,m}(\rho)$$

and conclude by optimising on the parameters m for the Borel-Cantelli lemma to apply. □

Remark: Results of the previous kind were first established in [52] and a complete proof was given in [103]. The natural question one can ask is whether *all* the vectors of the Hamming spheres $\cup_{\nu=1}^m \mathcal{S}(\xi^\nu, \rho N)$ *do* converge directly to the original patterns. It is possible to choose some particular vectors, containing only $\mathcal{O}(\sqrt{N})$ errors that do not converge to the original pattern, providing thus a negative answer to the previous question. Notice however that these vectors are explicitly constructed not to converge and hence they are not in contradiction with the almost sure result stated in the theorem.

It is quite remarkable that these results can be extended to other initial configurations, namely random mosaics. These are configurations coinciding with a given original pattern on some sites, with a different original pattern on some other sites and so on. It is even allowed to have an independent random noise on some sites. A theorem of attraction similar to the previous one can be shown for such random mosaics [104]. Namely, it is shown the

Theorem 4.3 *Let $(\xi_i^{m+1})_{i=1,\cdots,N}$ be a collection of N random variables, mutually independent and independent of the original variables $(\xi_i^\mu)_{i=1,\cdots,N}^{\mu=1,\cdots,m}$ having the same distribution with the original variables. Let $n_0, n_1, \cdots, n_{m+1}$ be an increasing family of positive integers with $0 = n_0 \leq n_1 \leq \cdots \leq n_{m+1} = N$ and $(I_\lambda)_{\lambda=1,\cdots,m+1}$ be a partition of the set of indices $\{1, \cdots, N\}$ with $\mathrm{card}(I_\lambda) = n_\lambda - n_{\lambda-1}$. Let*

$$x^\nu(0) = \begin{cases} \xi_i^\nu & \text{for } i \in I_1 \\ \xi_i^{a^\nu(\mu)} & \text{for } i \in I_\mu, \quad \mu = 2, \cdots, m \\ \xi_i^{\mu+1} & \text{for } i \in I_{\mu+1}, \end{cases}$$

where a^ν is a transposition of the index set $\{1, \cdots, N\}$ with $a^\nu(1) = \nu$ and $a^\nu(\nu) = 1$. Let $m = \frac{N}{\gamma \log N}$, $n_1 = \gamma_1 N$, and $n_m = \gamma_m N$, with $\gamma > 0$ and $0 < \gamma_1 \leq \gamma_m \leq 1$ and let finally ϵ be arbitrarily small. Then

1. *If $\gamma_1 > \frac{1}{2}(\gamma_m + \epsilon)$ and $\gamma > \frac{6}{(2\gamma_1 - \gamma_m - \epsilon)^2}$ then*

$$\mathbb{P}\left(\liminf_N [\cap_{\nu=1}^m \{Tx^\nu(0) = \xi^\nu\}]\right) = 1.$$

2. *If $\gamma_1 > \frac{1}{2}(\gamma_m + \epsilon)$ and $\gamma > \frac{4}{(2\gamma_1 - \gamma_m - \epsilon)^2}$ then*

$$\mathbb{P}\left(\cap_{\nu=1}^m \{Tx^\nu(0) = \xi^\nu\}\right) = 1 - R_N,$$

with $\lim_{N \to \infty} R_N = 0$.

However, all these are only partial results. It should be interesting to be able to prove attraction after a long number of time steps, or even asymptotically in time, but this remains still a challenging open problem.

4.3.2 Fluctuations of the Hamiltonian and existence of local minima

It can easily be shown that the Hopfield Hamiltonian decreases in time under the network evolution (exercise !). Moreover, the stable configurations are local minima of the Hamiltonian. It is therefore interesting to study the structure of local minima. A first result in this direction has been proven by Newman in 1988 by use of large deviation techniques [71]. Namely, he showed the following

Theorem 4.4 *There exists a number $\alpha_c > 0$ such that for all $\alpha \leq \alpha_c$ and $m = \alpha N$, there exist $\delta \in]0, 1/2[$ and $\epsilon > 0$ such that*

$$\mathbb{P}[\liminf_N (\cap_{\nu=1}^m \cap_{y \in \mathcal{S}(\xi^\mu, \delta N)} \{H_{N,m}(y) > H_{N,m}(\xi^\mu) + \epsilon N\})] = 1$$

where $\mathcal{S}(\xi^\mu, \delta N)$ is the Hamming sphere centred ξ^μ at and of radius δN.

Moreover, Newman obtained the numerical value $\alpha_c > 0.056$. On the other hand, Amit and his collaborators [5, 6], based on numerical simulations and non rigorous computations, predict $\alpha_c = 0.14$. In a more recent work [60, 61], Loukianova obtained $\alpha_c > 0.071$; she used large deviation techniques similar to that of Newman, but a finer decomposition of the space and then of non-uniform estimates. We give here this more recent proof.

Proof: Denote by $\mathcal{S}(x, r)$ the sphere centred at the configuration x with (Hamming) radius r and define $H_N(x, r) = \min_{x' \in \mathcal{S}(x,r)} H_N(x')$.

For $\delta \in]0,1/2[$ and $\epsilon > 0$, denote by $A(N,m,\delta,\epsilon)$ the event that each pattern is surrounded by an energy barrier of height ϵN on the sphere of radius $[\delta N]$:

$$A(N,m,\delta,\epsilon) = \cap_{\mu=1}^{m} \{h_N(\xi^\mu, [\delta N]) > H_N(\xi^\mu) + \epsilon N\}.$$

Put $M = m-1$ and define the random variables

$$W_N^\mu = \frac{1}{\sqrt{N}}(\xi^\mu, \xi^m).$$

These random variables are independent and identically distributed; their law weakly converges, when $N \to \infty$, towards a standard normal law. Define, for $\eta > 0$,

$$\Omega_\eta = \{\frac{1}{M}\sum_{\mu=1}^{M}(W_N^\mu)^2 \in]1-\eta, 1+\eta[\}.$$

Obviously,

$$\mathbb{P}[A^c(N,m,\delta,\epsilon)] \leq \mathbb{P}(\Omega_\eta^c) + \mathbb{P}[A^c(N,m,\delta,\epsilon) \cap \Omega_\eta].$$

For $J \subseteq \{1,\ldots,N\}$ and $x \in X_N$, denote by x_J the configuration that differs from x exactly on the co-ordinates J. It is immediate then to see that $x_J \in \mathcal{S}(x,|J|)$. Thus

$$\begin{aligned}
\mathbb{P}[A^c(N,m,\delta,\epsilon) \cap \Omega_\eta] &= \mathbb{P}[\cup_\mu \cup_{J:|J|=[\delta N]} \{H(\xi_J^\mu) - H(\xi^\mu) \leq \epsilon N\} \cap \Omega_\eta] \\
&\leq \sum_\mu \sum_{J:|J|=[\delta N]} \mathbb{P}[\{H(\xi_J^\mu) - H(\xi^\mu) \leq \epsilon N\} \cap \Omega_\eta] \\
&\leq m C_N^{[\delta N]} \mathbb{P}[\{H(\xi^\mu_{\{1,\ldots,[\delta N]\}}) - H(\xi^\mu) \leq \epsilon N\} \cap \Omega_\eta].
\end{aligned}$$

Now remark that the variables appearing in the last event can be expressed in terms of the variables W^μ by

$$H(\xi^\mu_{\{1,\ldots,[\delta N]\}}) - H(\xi^\mu) = 4[\delta N]\left(N - [\delta N] - 4\sqrt{[\delta N](N-[\delta N])}\sum_{\mu=1}^{M}\overline{W}_N^\mu \tilde{W}_N^\mu\right),$$

where

$$\overline{W}_N^\mu = \frac{1}{\sqrt{[\delta N]}}\sum_{i=1}^{[\delta N]}\xi_i^\mu \xi_i^\mu$$

and

$$\tilde{W}_N^\mu = \frac{1}{\sqrt{N-[\delta N]}}\sum_{i=1+[\delta N]}^{N}\xi_i^\mu \xi_i^\mu.$$

Introduce the quantities

$$q_N(M,\eta,C) = \mathbb{P}[\{\frac{1}{M}\sum_{\mu=1}^{M}\overline{W}_N^\mu \tilde{W}_N^\mu < C\} \cap \Omega_\eta]$$

and

$$p_N(M,\eta) = \mathbb{P}(\Omega_\eta^c)$$

to obtain

$$\mathbb{P}[A^c(N,m,\delta,\epsilon)] \leq p_N(M,\eta) + C_N^{[\delta N]} q_N(M,\eta,\frac{-\sqrt{\delta(1-\delta)}}{\alpha}+\epsilon'),$$

for all $\epsilon' > \frac{\epsilon}{4\alpha\sqrt{\delta(1-\delta)}}$ and N sufficiently large.

Let

$$\Lambda(t,r) = \lim_{M\to\infty} \frac{1}{M} \log \mathbb{E} \exp\{t\sum_{\mu=1}^{M}(W_N^\mu)^2 + r\sum_{\mu=1}^{M} \overline{W}_N^\mu \tilde{W}_N^\mu\}$$

and

$$\Lambda^*(x,y) = \sup_{t,r}(xt+yr-\Lambda(t,r))$$

its Legendre dual. We have then the standard large deviation estimate

$$\limsup_N \frac{1}{N}\log q_N(M,\eta,\frac{-\sqrt{\delta(1-\delta)}}{\alpha}+\epsilon') \leq -\inf_{1-\eta\leq x\leq 1+\eta} \alpha\Lambda^*(x,\frac{-\sqrt{\delta(1-\delta)}}{\alpha}+\epsilon')$$

and

$$\limsup_N \frac{1}{N}\log p_N(M,\eta) \leq -\frac{\alpha}{2}(\eta-\log(1+\eta)).$$

Denoting $\mathcal{I}(\delta) = -\delta\log\delta - (1-\delta)\log(1-\delta)$, Stirling formula guarantees that $\lim \frac{1}{N}\log C_N^{[\delta N]} = \mathcal{I}(\delta)$.

Now $\limsup_N \frac{1}{N}\log\mathbb{P}[A^c(N,m,\delta,\epsilon) < -K(\alpha,\delta,\epsilon)$, with $K(\alpha,\delta,\epsilon) > 0$ provided that we can choose $\alpha \in]0,\infty[$ and $\delta \in]0,1/2[$ such that, for $\eta > 0$ and $\epsilon > 0$, we have simultaneously

$$-\mathcal{I}(\delta) + \inf_{1-\eta\leq x\leq 1+\eta} \alpha\Lambda^*(x,\frac{-\sqrt{\delta(1-\delta)}}{\alpha}+\epsilon') > 0$$

and

$$\eta - \log(1+\eta) > 0.$$

It is shown in [61] that this is possible; numerical estimation gives then $\alpha^* \simeq 0.0712$ and for the corresponding $\delta \simeq 0.009$. □

Remark: The estimate for α^* has been further improved in [99] to the value $\alpha^* \simeq 0.085$. Obtaining rigorously a finite upper bound of α_c still remains an open problem.

Another result along the same lines was also proved in [61]

Theorem 4.5 *For given integers N and m, and $\delta > 0$, let*

$C(N,m\delta) = \{\exists\mu \in \{1,\cdots,m\} : \exists x \in \mathcal{B}(\xi^\mu,[\delta N])$ *such that x is a local minimum of H}*,

where $\mathcal{B}(\xi^\mu,[\delta N])$ is the Hamming ball around the configuration ξ^μ of radius $[\delta N]$, the symbol $[\cdot]$ denoting the integer part. Suppose that $\lim_{N\to\infty}\frac{m}{N} = \alpha > 0$. Then there exists a positive constant $\delta(\alpha)$, depending on α, such that

$$\liminf_{N\to\infty} \frac{1}{N}\log\mathbb{P}[C(N,m,\delta)] \leq -K(\alpha,\delta),$$

with $K(\alpha,\delta) > 0$ *and* $\liminf_{\alpha \to \infty} \delta(\alpha) \geq \delta_* \simeq 0.05$.

Remark: Notice that the set $C(N, m\delta)$ can be rewritten as

$$C(N, m\delta) = \cup_{\mu=1}^{m} \cup_{k=1}^{[\delta N]} \cup_{\substack{J \subseteq \{1,\ldots,N\} \\ |J|=k}} \{\xi_J^\mu \text{ is a local minimum of } H\},$$

where

$$\{\xi_J^\mu \text{ is a local minimum of } H\} = \{\forall i \in J : H(\xi^\mu_{J\setminus\{i\}}) > H(\xi^\mu_J) \cap \{\forall i \notin J : H(\xi^\mu_{J\cup\{i\}}) > H(\xi^\mu_J)\}.$$

Thus the previous theorem establishes that there are no local minima in small balls around each pattern.

With a view to numerical simulations, it is very interesting to know whether near a given configuration $x \in X_N$ there are many Hamming neighbours having lower energy than x. This question can be formalised more precisely in terms of the notion of escape direction. For a given configuration x and a given site i, denote as usual by $x_{\{i\}}$ the configuration that differs from x only at site i. Thus $x_{\{i\}}$ is a nearest Hamming neighbour of x.

Definition 4.6 A site $i \in \{1, \ldots, N\}$ is an *escape direction* for the configuration x if

$$H_{N,m}(x) > H_{N,m}(x_{\{i\}}).$$

We denote by

$$\mathcal{N}_N(x) = \text{card}\{i \in \{1, \ldots, N\} : i \text{ is an escape direction of } x\}.$$

It can the be proved [61] the following

Theorem 4.7 *Let* $\lim \frac{m}{N} = \alpha > 0$. *There exist two strictly positive parameters* $\gamma_*(\alpha)$ *and* $\gamma^*(\alpha)$ *such that, for every* $\mu = 1, \ldots, m$,

$$\gamma_*(\alpha) \leq \liminf_N \frac{\mathcal{N}_N(\xi^\mu)}{N} \leq \limsup_N \frac{\mathcal{N}_N(\xi^\mu)}{N} \leq \gamma^*(\alpha).$$

Numerical values for the parameters $\gamma_*(\alpha)$ and $\gamma^*(\alpha)$ can be explicitly obtained as functions of α. It is argued that the critical storage capacity should have an interpretation also in terms of the asymptotic number of escape directions.

5 Thermodynamics of the restitution process

5.1 Main definitions

We again limit ourselves to the Hopfield model of binary neurones, sequentially updated. Now however, we allow for a stochastic evolution, *i.e.* instead of rigidly imposing the evolution

$$x_i(t+1) = \mathrm{sgn}(\frac{1}{N}\sum_{\mu=1}^{m}\sum_{1\leq i<j\leq N}\xi_i^\mu\xi_j^\mu x_j(t)),$$

we introduce the $2^N \times 2^N$ transition matrices

$$p_i^{(\beta)}(x,y) = \begin{cases} \frac{\exp(\beta h_i(x)y_i)}{\exp(\beta h_i(x))+\exp(-\beta h_i(x))} & \text{if } y_j = x_j \text{ for } j \neq i \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, N$, and $x, y \in \{-1, 1\}^N$ where

$$h_i(x) = \frac{1}{N}\sum_{\mu=1}^{m}\sum_{j:j\neq i}\xi_i^\mu\xi_j^\mu x_j.$$

These transition matrices allow the definition of a Markov chain $(Y_n)_{n\geq 0}$ on $\{-1,1\}^N$ by

$$\mathbb{P}(Y_{n+1} = y | Y_n = x) = p^{(\beta)}(x,y),$$

where

$$p^{(\beta)}(x,y) = \sum_{z_1,\ldots,z_N} p_1^{(\beta)}(x,z_1)\cdots p_N^{(\beta)}(z_N,y).$$

As it was the case for the stochastic learning procedure, the stochastic evolution tends to the deterministic one if one let $\beta \to \infty$. Gibbs formalism arises now naturally. Define the finite volume Gibbs measure as a probability on the configuration space given by

$$\gamma_{N,m,\beta}(\cdot) = \frac{1}{Z_{N,m,\beta}}\sum_x \exp(-\beta H_{N,m}(x))\mathbb{1}_{(\cdot)}(x),$$

where $Z_{N,m,\beta}$ is the partition function.

It is then easily shown the following

Proposition 5.1 *The finite volume Gibbs measure $\gamma_{N,m,\beta}$ is the unique invariant measure associated with the Markov chain of transition probability p.*

Proof: Exercise! □

The specific free energy is defined, as usual, by

$$f_{N,m,\beta} = -\frac{1}{\beta N}\log Z_{N,m,\beta}$$

and since the Hamiltonian depends on the patterns ξ, so does the partition function and hence the free energy.

To obtain generic results, we can again choose the variables ξ to be Bernoulli independent random variables. Therefore the thermodynamic functions in finite volume are random variables as it is the case in spin glass systems [82]. The same questions can be asked, namely whether the thermodynamic variables tend to a non random limit when $N \to \infty$, or whether quenched, $f_{N,m,\beta}$, and annealed,

$$\overline{f}_{N,m,\beta} = -\frac{1}{\beta N} \log \mathbb{E} Z_{N,m,\beta},$$

free energies coincide in the infinite volume limit. These questions are partially answered in the following sections.

5.2 Behaviour of the annealed free energy

The asymptotic behaviour of the annealed free energy is established in the high and low temperature regimes [103].

Theorem 5.2 *Let* $\alpha = \lim \frac{m}{N}$. *Then*

1. *if* $\beta < 1$,
$$\lim_{\substack{N \to \infty \\ \alpha \geq 0}} \overline{f}_{N,m,\beta} = -\frac{\log 2}{\beta} + \frac{\alpha}{2\beta} \log(1-\beta) + \frac{\alpha}{2}$$

2. *if* $\beta > 1$,
$$\lim_{m,N \to \infty} \overline{f}_{N,m,\beta} = -\infty$$
and
$$\lim_{\substack{N \to \infty \\ m \text{ finite}}} \overline{f}_{N,m,\beta} = -\frac{\log 2}{\beta} + \frac{m}{\beta} \inf_{x \in \mathbb{R}} \{\frac{x^2}{2} - \log \cosh(\sqrt{\beta} x)\}.$$

Proof: The main idea of the proof is to introduce auxiliary Gaussian variables to linearise the quadratic form in the exponent of the partition function

$$Z_{N,m,\beta} = \exp(-\frac{\beta m}{2}) \sum_{x \in X_N} \prod_{\mu=1}^{m} \exp(\frac{\beta}{2N} (\xi^\mu, x)_N^2)$$

where $(\cdot,\cdot)_k$ denotes the scalar product of $\mathbb{R}^k$. Using the identity

$$\exp(\frac{x^2}{2a}) = \sqrt{\frac{a}{2\pi}} \int_{\mathbb{R}} \exp(-\frac{a}{2} t^2 + tx) dt,$$

we write

$$Z_{N,m,\beta} = \exp(-\frac{\beta m}{2})(\frac{N}{2\pi})^{m/2} \int_{\mathbb{R}^m} dr_1 \cdots dr_m \sum_{x \in X_N} \exp(-\frac{N}{2}\sum_{\mu=1}^{m} r_\mu^2 + \sqrt{\beta}\sum_{\mu=1}^{m} r_\mu \sum_{i=1}^{N} x_i \xi_i^\mu).$$

But now the integration over x is trivially performed and by Fubini's theorem we obtain

$$\mathbb{E} Z_{N,m,\beta} = 2^N \exp(-\frac{\beta m}{2}) \left[(\frac{N}{2\pi})^{1/2} \int_{\mathbb{R}} dr \exp(-\frac{Nr^2}{2} + N \log\cosh(\sqrt{\beta} r)) \right]^m .$$

Using standard Laplace's method, we obtain, asymptotically in N,

$$\overline{f}_{N,m,\beta} = -\frac{\log 2}{\beta} + \frac{m}{2N} - \frac{m}{2\beta N} - \frac{m}{2\beta N} \log \frac{N}{2\pi} - \frac{m}{\beta} \inf_{x \in \mathbb{R}} \{ \frac{x^2}{2} - \log\cosh(\sqrt{\beta} x) \} + R_N,$$

with $\lim_N R_N = 0$.

Now, for $\beta > 1$

$$\inf_{x \in \mathbb{R}} \{ \frac{x^2}{2} - \log\cosh(\sqrt{\beta} x) \} = c_\beta < 0$$

and when $\lim_N m = \infty$ the annealed free energy diverges. On the contrary, for $\beta < 1$,

$$\inf_{x \in \mathbb{R}} \{ \frac{x^2}{2} - \log\cosh(\sqrt{\beta} x) \} = 0.$$

We split then the term $-\frac{Nr^2}{2}$ appearing in the exponential into $-\frac{Nr^2(1-\beta)}{2} - \frac{Nr^2\beta}{2}$, absorb $(1-\beta)$ in the integration variable, carry out the integration, and conclude. □

5.3 Behaviour of the quenched free energy

The next point is to establish the existence of the limit of the quenched free energy. This can be proved only for a region of the parameters.

Theorem 5.3 *Let $\alpha \in [0,1[$ and $\delta \in]0,1[$ such that $\delta - 4(\sqrt{\alpha}(1-\delta) > 0$. Let*

$$\begin{aligned} \Phi_\delta :]0,\infty[\times \mathbb{R} &\to \mathbb{R} \\ (\beta, h) &\mapsto -\frac{1}{\beta} \log\cosh(\beta h) + \frac{1-\delta}{2} h^2, \end{aligned}$$

and define two functions

$$\Psi^1(\beta, \alpha, \delta) = \min_{h \in \mathbb{R}} \Phi_\delta(\beta, h) - \frac{\log 2}{\beta} + \frac{\alpha}{2} + \frac{\alpha}{\beta} \log(\delta - 4\sqrt{\alpha}(1-\delta)),$$

and

$$\Psi^2(\beta, \alpha) = \min_{h \in \mathbb{R}} \Phi_0(\beta, h) - \frac{\log 2}{\beta} + \frac{\alpha}{2}.$$

Then, for every $\beta > 0$,

1. For $\mathbb{P}$-almost every $\omega \in \Omega$,

$$\Psi^1(\beta,\alpha,\delta) \leq \liminf_{\substack{N\to\infty \\ \frac{m}{N}\to\alpha}} f_{N,m,\beta} \leq \limsup_{\substack{N\to\infty \\ \frac{m}{N}\to\alpha}} f_{N,m,\beta} \leq \Psi^2(\beta,\alpha),$$

2.

$$\Psi^1(\beta,\alpha,\delta) \leq \liminf_{\substack{N\to\infty \\ \frac{m}{N}\to\alpha}} \mathbb{E} f_{N,m,\beta} \leq \limsup_{\substack{N\to\infty \\ \frac{m}{N}\to\alpha}} \mathbb{E} f_{N,m,\beta} \leq \Psi^2(\beta,\alpha),$$

In particular, when $\alpha = 0$, the free energy converges almost surely to a constant (the Curie-Weiss free energy) as it can be seen from the previous theorem by taking $\alpha = 0$ and letting, by continuity, $\delta \to 0$.

The weaker assertion (2) of the previous theorem — the majorisation and minoration of $\mathbb{E} f_{N,m,\beta}$ — and the idea of its proof were formulated in [49]. In the present form it was completely proved in [18, 103]. The complete proof is quite complicated and needs several intermediate steps for which we direct the interested reader to the original papers. Notice also that the existence of the quenched free energy, when $\alpha \neq 0$, remains an open problem in spite of the continuing efforts. This is a general problem not only for Hopfield model but also for the frustrated mean-field spin-glasses (see [82] for instance).

5.4 Self-averaging

Self-averaging is a property of some macroscopic quantities of disordered systems. Here we only stick to the self-averaging property of the quenched free energy.

It is remarkable that free energy of the Hopfield model has the *weak self-averaging property*; it is shown that the quenched free energy is very "close" to its average. Thus although the existence of the average it is not known, it can be proven that the free energy satisfies a very strong concentration property.

This result has a long history. Using martingale difference methods introduced by Girko for the study of spectral properties of random matrices in [35] (see [37] for a more recent and more easily accessible document), Pastur and Shcherbina obtain, in [77], a weak self-averaging property for the quenched free energy of the Sherrington-Kirkpatrick spin glass. Then, their method has been applied to the Hopfield model by Vermet, in [102], and then this result has been improved by Shcherbina and Tirozzi, in [93], and then in [78]. Following the same lines of reasoning but using a much more careful estimation of the terms appearing in the expansions, Bovier, Gayrard, and Picco [18] obtained the strongest formulation of weak self-averaging. Their theorem essentially proves the almost sure convergence of a rate function that is related to a constrained free energy. Stating precisely their result at this stage should need the introduction of various notions and notations that will be naturally introduced in the next chapter. Therefore, formulation

of this result is postponed. Instead, it is preferred to state and prove here a weaker [102] self-averaging result that has the advantage of illustrating the Girko's method quite transparently.

Theorem 5.4 *Let $\beta > 0$ and $m \leq \gamma(N)\sqrt{N}$, with $\lim_N \gamma(N) = 0$. Then,*

$$\lim_{N \to \infty; m \to \infty} \mathbb{E}\left((f_{N,m,\beta} - \mathbb{E}f_{N,m,\beta})^2\right) = 0.$$

Proof: Introduce an auxiliary parameter $t \in [0,1]$ and fix some site k with $1 \leq k \leq N$. Instead of the original Hopfield Hamiltonian

$$H_{N,m}(x) = -\frac{1}{2N} \sum_{\mu=1}^{m} \sum_{i,j=1,\ldots,N; i \neq j} \xi_i^\mu \xi_j^\mu x_i x_j,$$

consider the modified Hamiltonian

$$\tilde{H}_{N,m}(x;k,t) = -\frac{1}{2N} \sum_{\mu=1}^{m} \sum_{i \neq j; i,j \neq k} \xi_i^\mu \xi_j^\mu x_i x_j - \frac{t}{N} \sum_{\mu=1}^{m} \sum_{i=1,\ldots,N; i \neq k} \xi_i^\mu \xi_k^\mu x_i x_k.$$

Obviously, $\tilde{H}_{N,m}(x;k,1) = H_{N,m}(x)$.

With this modified Hamiltonian, define the modified partition function

$$\tilde{Z}_{N,m,\beta}(k,t) = \sum_x \exp(-\beta \tilde{H}_{N,m}(x;k,t)),$$

the modified finite volume, quenched, specific, free energy

$$\tilde{f}_{N,m,\beta}(k,t) = -\frac{1}{\beta N} \log \tilde{Z}_{N,m,\beta}(k,t),$$

and the modified finite volume Gibbs measure

$$\tilde{\gamma}_{N,m,\beta}(\cdot;k,t) = \frac{1}{\tilde{Z}_{N,m,\beta}(k,t)} \sum_x \exp(-\beta \tilde{H}_{N,m}(x;k,t)) \mathbb{1}_{(\cdot)}(x).$$

Following Girko, introduce now the increasing sequence of σ-algebras

$$\mathcal{F}_k = \sigma(\xi_1, \ldots, \xi_k), \quad k = 0, \ldots, N,$$

with $\mathcal{F}_0 = \{\emptyset, \Omega\}$. It is immediate that

$$\begin{aligned} f_{N,m,\beta} - \mathbb{E}f_{N,m,\beta} &= \sum_{k=1}^{N} [\mathbb{E}(f_{N,m,\beta}|\mathcal{F}_k) - \mathbb{E}(f_{N,m,\beta}|\mathcal{F}_{k-1})] \\ &= [\mathbb{E}(\tilde{f}_{N,m,\beta}(k,1)|\mathcal{F}_k) - \mathbb{E}(\tilde{f}_{N,m,\beta}(k,1)|\mathcal{F}_{k-1})] \end{aligned}$$

because $f_{N,m,\beta}$ is measurable with respect to $\mathcal{F}_N$. Moreover, the terms of this sum are orthogonal for different k's. Now, use the trivial identity, valid for any $g \in C^1([0,1])$,

$$g(1) = g(0) + \int_0^1 \frac{dg}{dt}(t)dt.$$

Applying this identity in the summands above we remark that

$$\mathbb{E}(\tilde{f}_{N,m,\beta}(k,0)|\mathcal{F}_k) - \mathbb{E}(\tilde{f}_{N,m,\beta}(k,0)|\mathcal{F}_{k-1}) = 0$$

because the part of $\tilde{f}_{N,m,\beta}(k,t)$ that is measurable with respect to $\mathcal{F}_k$ without being measurable with respect to $\mathcal{F}_{k-1}$, is precisely the term

$$-\frac{t}{N}\sum_{\mu=1}^{m}\sum_{i=1,\ldots,N;i\neq k}\xi_i^\mu\xi_k^\mu x_i x_k$$

which vanishes when $t = 0$. It only remains the derivative part yielding

$$f_{N,m,\beta} - \mathbb{E}f_{N,m,\beta} = \frac{1}{N}\sum_{k=1}^{N}\int_0^1\sum_{\mu=1}^{m}\sum_{i;i\neq k}[\mathbb{E}(\xi_i^\mu\xi_k^\mu\tilde{\gamma}_{N,m,\beta}(x_ix_k;k,t)|\mathcal{F}_{k-1}) - \xi_k^\mu\mathbb{E}(\xi_i^\mu\tilde{\gamma}_{N,m,\beta}(x_ix_k;k,t)|\mathcal{F}_k)]dt.$$

Bounding trivially

$$|\tilde{\gamma}_{N,m,\beta}(x_ix_k;k,t)| \leq 1,$$

we get from the orthogonality of the martingale differences that

$$\mathbb{E}\left((f_{N,m,\beta} - \mathbb{E}f_{N,m,\beta})^2\right) \leq \frac{4m^2}{N},$$

hence the theorem. □

Remark: It is worth noticing that to improve the above result, the martingale difference part of the proof remains unchanged. Only some technical additional work is needed to improve the trivial bound $|\tilde{\gamma}_{N,m,\beta}(x_ix_k;k,t)| \leq 1$ used to estimate the summands. This can be achieved by using combinatorial and spectral estimates [93, 103].

At this place, a special mention is needed for the profound results of Talagrand [99], interested in precise estimates valid not only asymptotically but at every N. Based on techniques of concentration of measures [58, 97, 98], he establishes, among other things, a series of results on concentration properties of the free energy.

Finally in [90], using previously developed methods [2], it is shown that $Z_{N,m,\beta}/\mathbb{E}Z_{N,m,\beta}$ converges, in the high temperature regime, to a log-normal random variable when $m = \alpha N$, with $\alpha > 0$. Although I believe this result is undoubtedly correct, the proof given in [90], especially the proof of the technical result 2.17′ given in their appendix — on which the estimates of the whole paper rely —, is in my opinion incomplete. More recently, Talagrand [99] proves a stronger result, namely

Theorem 5.5 *Let $\alpha = m/N$ and $\beta(1+\sqrt{\alpha}) < 1$. Then there is a positive constant K depending on α and β such that, for every $u > 0$,*

$$\mathbb{P}\left(f_{N,m,\beta} \geq +\frac{m}{2\beta N}(\log(1-\beta) - u) \right) \leq K \exp(-\frac{u^2}{K}).$$

This theorem is valid at every N and not only asymptotically.

6 Gibbs states of neural networks

6.1 Extension of measures

Extension of a measure defined on a finite-dimensional space to a measure on an infinite dimensional one is a basic problem in probability theory since the mere existence of stochastic processes lies on such a construction. The first result towards this direction was the celebrated Kolmogorov's extension theorem [51]. It is instructive to recall the precise statement of Kolmogorov's theorem and to introduce some notation that will be useful in the sequel.

6.1.1 Fixing marginals: Kolmogorov's extension and unicity

As usual, the one site state space is a probability space $(S, \mathcal{S}, \kappa)$ that will be assumed *compact* and metrisable. For a vertex set V (that is a denumerable discrete set and assumed embedable in a Lipschitz way into $\mathbb{R}^d$ for some d), the configuration space is given by $X = \{x : V \to S\} = S^V$. For a finite subset $\Lambda \subset V$, we denote $X_\Lambda = \{x : \Lambda \to S\} = S^\Lambda$; this space will be identified with the set of restrictions x_Λ of configurations on Λ. The natural σ-algebra on X_Λ will be $\mathcal{F}_\Lambda = \mathcal{S}^\Lambda$ and product measure will be denoted κ^Λ. We use the symbol $\mathcal{F}$ for $\mathcal{F}^V$. The problem of extension of measure is the possibility of equipping the measurable space $(X, \mathcal{F})$ with a probability $\mathbb{P}$ whose finite-dimensional marginals are fixed.

Theorem 6.1 (Kolmogorov) *Let $(\Lambda_n)_{n\geq 1}$ be an increasing sequence of finite subsets of V and $(X_{\Lambda_n}, \mathcal{F}_{\Lambda_n})$ be the corresponding sequence of measurable configuration spaces. Suppose that on each $\mathcal{F}_{\Lambda_n}$ is defined a probability $\mathbb{P}_{\Lambda_n}$ such that the family $(\mathbb{P}_{\Lambda_n})_n$ verifies the following compatibility condition:*

$$\textit{If } \Lambda_n \subseteq \Lambda_{n'} \textit{ then } \mathbb{P}_{\Lambda_n}(F) = \mathbb{P}_{\Lambda_{n'}}(F \times S^{\Lambda_{n'} \setminus \Lambda_n}), \quad \forall F \in \mathcal{F}_{\Lambda_n}$$

(the measure $\mathbb{P}_{\Lambda_n}$ is the marginal of $\mathbb{P}_{\Lambda_{n'}}$ on X_{Λ_n}). Then there exists a unique *probability measure $\mathbb{P}$ on $(X, \mathcal{F})$ such that*

$$\mathbb{P}(F \times S^{\Lambda_n^c}) = \mathbb{P}_{\Lambda_n}(F), \quad \forall F \in \mathcal{F}_{\Lambda_n}$$

(the finite-dimensional marginals of $\mathbb{P}$ coincide with $\mathbb{P}_{\Lambda_n}$).

This theorem, a corner-stone in the theory of stochastic processes, is of restricted use in the context of statistical mechanics because it excludes the possibility of phase transitions.

6.1.2 Fixing conditional expectations: the DLR construction and phase transition

The DLR construction, after the names of Dobrushin [26] and Lanford and Ruelle [56], is reminiscent of the Kolmogorov's construction. Instead of fixing finite-dimensional marginals however, we fix conditional probabilities with respect to fixed boundary conditions. The DLR construction can be seen as a generalisation of the existence of an invariant probability for a Markov chain. Here however, the ordering of the indexing set, played by "time" for Markov chains, is played by an ordering by inclusions of subsets of finite subsets of V. The Markovian character of the process is guaranteed by the existence of a genuine Hamiltonian as it is explained below.

Notations 6.2 We use the symbol $A \subset\subset V$ to denote that A is a *finite* subset of V.

Definition 6.3 Let $(\Phi_A)_{A\subset\subset V}$ be a family of mappings $\Phi_A : X \to \mathbb{R}$, indexed by the finite subsets of V such that Φ_A is $\mathcal{F}_A$-measurable for every $A \subset\subset V$ (the mapping Φ_A depends only on the configurations restricted over A). Then the family $(\Phi_A)_{A\subset\subset V}$ is called a family of *interaction potential*.

Interaction potentials introduce a coupling between configurations over different sites that allow to construct measures that are not plainly product measures. However, to keep an overall spatial Markovian structure, this interaction must be moderate.

Definition 6.4 If, for every $\Lambda \subset\subset V$ and every $x \in X$, the sum

$$\sum_{A\subset\subset V; A\cap\Lambda\neq\emptyset} \Phi_A(x)$$

exists, then it is called (genuine) *interaction Hamiltonian* and is denoted by $H_\Lambda(x)$. If this sum does not exist but there is an increasing and diverging to $+\infty$ sequence of positive numbers $(a_\Lambda)_{\Lambda\subset\subset V}$ such that

$$\frac{1}{a_\Lambda} \sum_{A\subset\Lambda} \Phi_A(x)$$

is extensive, then the latter sum is called *mean-field interaction Hamiltonian* and is denoted by the same symbol $H_\Lambda(x)$.

Remark: In the above definition, *extensive* means that, asymptotically and in some sense not to be farther precised here, the mean-field Hamiltonian $H_\Lambda(x)$ behaves like $|\Lambda|$.

Example 6.5 Let $S = \{-1, 1\}$. For a positive parameter J and a real parameter h, we define the interaction potentials

$$\Phi_A(x) = \begin{cases} -Jx_ix_j & \text{if } A = \{i,j\} \text{ and } |i-j| = 1 \\ -hx_i & \text{if } A = \{i\} \\ 0 & \text{otherwise} \end{cases}$$

and

$$\Psi_A(x) = \begin{cases} -Jx_ix_j & \text{if } A = \{i,j\} \\ -hx_i & \text{if } A = \{i\} \\ 0 & \text{otherwise.} \end{cases}$$

The potential Φ gives rise to a genuine Hamiltonian and defines the ferromagnetic Ising model with external field; the potential Ψ gives rise to mean-field Hamiltonian with $a_\Lambda = |\Lambda|$ and defines the Curie-Weiss model with external field.

We assume henceforth in this subsection that the model is defined by a genuine interaction Hamiltonian and that for every $\beta > 0$ and every y_{Λ^c}, the integral

$$Z_{\Lambda^c,\beta}(y_{\Lambda^c}) = \int_{X_\Lambda} \exp\left(-\beta H_\Lambda(x_\Lambda y_{\Lambda^c})\right) \kappa^\Lambda(dx)$$

exists. Here $x_\Lambda y_{\Lambda^c}$ denotes the concatenation of two restricted configurations, that is a configuration $z \in X$ such that

$$z_i = \begin{cases} x_i & \text{if } i \in \Lambda \\ y_i & \text{if } i \notin \Lambda \end{cases}$$

The configuration y_{Λ^c} is a boundary condition and the quantity $Z_{\Lambda^c,\beta}(y_{\Lambda^c})$ is called the *finite-volume partition function* with boundary condition y_{Λ^c}.

Definition 6.6 A probability defined on $(X, \mathcal{F})$ for every measurable set $F \in \mathcal{F}$ and any boundary condition $y_{\Lambda^c} \in X_{\Lambda^c}$ by

$$\gamma_{\Lambda,\beta}(F|y_{\Lambda^c}) = \frac{1}{Z_{\Lambda^c,\beta}(y_{\Lambda^c})} \int_{X_\Lambda} \exp\left(-\beta H_\Lambda(x_\Lambda y_{\Lambda^c})\right) \mathbb{1}_F(x_\Lambda y_{\Lambda^c}) \kappa^\Lambda(dx)$$

is called a *finite-dimensional Gibbs' specification* for the boundary condition $y_{\Lambda^c} \in X_{\Lambda^c}$.

Remark: The Gibbs' specification is a Markovian kernel of conditional probabilities with respect to the ordering defined by the inclusions.

We are seeking for measures γ on $(X, \mathcal{F})$ having as finite-dimensional conditional laws the Gibbs' specifications.

Definition 6.7 (DLR equation) Let

$$\mathcal{G} = \{\gamma \in \mathcal{M}_1(X, \mathcal{F}) : \forall F \in \mathcal{F}, \gamma\text{-a.e. } y \in X, \forall \Lambda \subset\subset V; \gamma(F|y_{\Lambda^c}) = \gamma_{\Lambda,\beta}(F|y_{\Lambda^c})\}.$$

The set $\mathcal{G}$ is called set of *Gibbs measures* specified by the family of kernels $(\gamma_{\Lambda,\beta}(\cdot|y_{\Lambda^c}))_{\Lambda \subset\subset V}$.

Remark: Contrary to the Kolmogorov's theorem, the DLR equation is less restrictive: the set $\mathcal{G}$ can be empty, can be a singleton, or it can have several elements (in fact infinitely many). For models encountered in statistical mechanics for which the DLR construction is possible, the set $\mathcal{G}$ is not empty. The passage from the regime of unique Gibbs measure to the regime where many solutions exist is called a *phase transition*.

The general structure of the set $\mathcal{G}$ is a difficult problem and is the object of study of a whole discipline, the equilibrium statistical mechanics, lying beyond the scope of the present review. The interested reader can profitably consult complete treaties on the topic, [34, 87, 94]. The only thing that will be mentioned here is that the set $\mathcal{G}$ has a simplicial structure. The extremal points of this set, called *pure states*, are attained through special choices of the boundary conditions; in general, the limiting Gibbs measure is a convex combination of pure states.

We end however this subsection by recalling once more that the DLR construction is possible only for models defined by a genuine interaction Hamiltonian excluding all mean-field neural network models. For instance, the only neural network model from the ones presented in this report for which DLR construction is possible is the one called "short-range finite-dimensional network", introduced in section 2.6.

6.1.3 Weak limiting procedure

For models defined by genuine Hamiltonians, still another construction is possible that is shown to be equivalent to the DLR construction. This is the *weak-limiting* procedure, defined briefly below.

Consider a system with genuine finite-volume Hamiltonian H_Λ. Fix some arbitrary configuration $y \in X$ and define the relative Hamiltonian $H_\Lambda(\cdot|y) : X \to \mathbb{R}$ by

$$H_\Lambda(x|y) = H_\Lambda(x_\Lambda y_{\Lambda^c}).$$

Denote by $\sigma_\Lambda : X \to X_\Lambda$ the *canonical projection* such that $x \mapsto \sigma_\Lambda(x) = x_\Lambda \in X_\Lambda$. Every one site space S (viewed as a fibre for the construction of the configuration space as a fibre bundle over the base V), can be equipped with the discrete topology τ_0. Thus, X can be equipped with the product topology $\tau = \prod_{i \in V} \tau_i$; with this topology, the canonical projections are continuous functions.

Now the Borel σ-algebra $\mathcal{B}(X)$, rendering measurable the open sets for the topology τ of X, coincide with the σ-algebra $\mathcal{F}$ generated by the collection $\mathcal{C}$ of *cylinder sets* $C_{\Lambda_n}(F_n)$, where for a given $n \in \mathbb{N}$ and a given increasing sequence of volumes Λ_n, we define the cylinder sets by

$$C_{\Lambda_n}(F_n) = \{x \in X : \sigma_{\Lambda_n}(x) \in F_n \quad \text{for} \quad F_n \in \mathcal{F}_{\Lambda_n}\}.$$

Define also, for a fixed configuration $y \in X$, the space

$$X_\Lambda(y) = \{x \in X : x_i = y_i \quad \text{for} \quad i \in \Lambda^c\}.$$

With the definitions introduced so far, it is possible to introduce a sequence of probability measures on $(X, \mathcal{B}(X))$ by

$$\mu_{n,\beta,y}(A) = \gamma_{\Lambda_n,\beta,y_{\Lambda_n^c}}(\sigma_{\Lambda_n}(A \cap X_{\Lambda_n}(y))), \quad \forall A \in \mathcal{B}(X),$$

where $\gamma_{\Lambda_n,\beta,y_{\Lambda_n^c}}$ is a finite-volume Gibbs measure defined by

$$\gamma_{\Lambda,\beta,y_{\Lambda^c}}(x_\Lambda) = \frac{\exp(-\beta H_\Lambda(x_\Lambda|y_{\Lambda^c})}{\sum_{x_\Lambda \in X_\Lambda} \exp(-\beta H_\Lambda(x_\Lambda|y_{\Lambda^c})}.$$

Notice that $\gamma_{\Lambda_n,\beta,y_{\Lambda_n^c}}$ is very reminiscent of the Gibbs specification introduced in the previous section but it does not exactly coincides with it since it is not defined on the whole configuration space X but only the subspace X_Λ.

The space S being compact for the discrete topology, the same holds true, by virtue of the Tychonov's theorem, for the space X for the product topology τ. Moreover, the topology τ can be easily metrised so that X is a compact metrisable space, hence complete and separable, what technically is called a *Polish space*. By Riesz-Markov theorem, there exists a bijection between probability measures μ on $(X, \mathcal{B}(X))$ and positive normalisable linear functionals on the Banach space $C(X)$ of continuous bounded real functions on X. Thus the set$\mathcal{M}_1(X)$ of probability measures on X is identified with a subset of the dual $C^*(X)$ of $C(X)$. Now the topology τ defines continuity on $C(X)$; it induces therefore a weak-* on $C^*(X)$, called the *vague topology* on $\mathcal{M}_1(X)$. (Recall [12] that a sequence of probability measures (μ_n) of $\mathcal{M}_1(X)$ converges weakly to μ, denoted $\mu_n \Rightarrow \mu$ if, and only if, $\lim_{n \to \infty} \int f d\mu_n \to \int f d\mu$, for all $f \in C(X)$.) We have the following

Theorem 6.8 *The space $\mathcal{M}_1(X)$ is compact for the topology of weak convergence. Moreover, $\mu_n \Rightarrow \mu$ if, and only if, $\mu_n(C) \to \mu(C)$ for every cylinder set C of X.*

The previous theorem guarantees that all sequences of finite volume measures have at least one accumulation point. We thus define

Definition 6.9 A probability μ on $(X, \mathcal{B}(X))$ is a Gibbs measure if μ belongs to the closed convex envelope of the set of accumulation points (for the weak topology) of the sequence of measures $(\mu_{n,\beta,y}(\cdot))_n$ defined above.

This construction gives rise to Gibbs measures that are convex combinations of extremal (pure) DLR Gibbs states [67], defined through fixed boundary conditions. The natural question is how to choose *extremal* states through this weak limiting procedure. This can be achieved by perturbing the original Hamiltonian by additional terms that vanish eventually, after the infinite-volume limit is taken. To fix ideas, consider the standard Ising model Hamiltonian

$$H_\Lambda(x) = -J \sum_{i \in \Lambda} \sum_{j \in \mathbb{Z}^d; |i-j|=1} x_i x_j.$$

This Hamiltonian has two minima[6], denoted respectively + and −, defined so that

$$\begin{aligned} x = + &\quad \Leftrightarrow \quad x_i = +1, \forall i \in \Lambda \\ x = - &\quad \Leftrightarrow \quad x_i = -1, \forall i \in \Lambda \end{aligned}$$

The well known Peierls argument establishes the existence of a phase transition (existence of at least two different infinite volume Gibbs measures) provided that $d \geq 2$. Consider now a Hamiltonian perturbed by an additional term

$$H_\Lambda^h(x) = -J \sum_{i \in \Lambda} \sum_{j \in \mathbb{Z}^d; |i-j|=1} x_i x_j - h \sum_{i \in \Lambda} x_i.$$

It is clear that if $h \neq 0$, one of the two ground energy configurations is favoured and thus the external field lifts the degeneracy. As a result, H_Λ^h, with $h \neq 0$ gives rise to a *unique* Gibbs measure. Taking now the limit $h \downarrow 0$ *after the infinite volume limit is performed*, the extremal Gibbs measure corresponding to the + boundary condition DLR Gibbs state is obtained. Similarly by taking $h \uparrow 0$, the DLR Gibbs state with − boundary condition is chosen.

This observation is quite general and allows to choose extremal Gibbs measures. When there are q minimising configurations, $q-1$ external fields are needed to lift the degeneracy.

6.2 Induced measures

Even the weak limiting procedure introduced in the last subsection needs the description of the model by a genuine Hamiltonian and thus fails for mean-field models. One method to circumvent this difficulty is to use null boundary conditions for mean-field models as it was done in [4] for the Curie-Weiss random field model. Another possibility is to reduce the problem to the equivalent problem over magnetisation and follow a precise weak limiting procedure introduced in [18] we describe briefly below.

We have already seen that $\gamma_{N,m,\beta}$, the finite volume Gibbs measure, is the unique invariant probability of stochastic evolution. However, the model is a long range system and this destroys its spatial Markovian properties and makes impossible the DLR construction.

Starting from the (slightly modified) Hopfield Hamiltonian

$$H_{N,m}(x) = -\frac{1}{2N} \sum_{i,j} \sum_{\mu=1}^{m} \xi_i^\mu \xi_j^\mu x_i x_j,$$

[6] This is true *stricto sensu* only in the infinite-volume limit. Now, in the infinite-volume limit the Hamiltonian diverges. The precise statement is that both $H_\Lambda(+)$ and $H_\Lambda(-)$ are of the order $-J|\Lambda|(1 + \mathcal{O}(\frac{|\partial\Lambda|}{|\Lambda|}))$

we introduce, for every $\eta \in \mathbb{N}$, the finite volume Gibbs measure in an external field h by

$$\gamma^{\eta,h}_{N,m,\beta}(x) = \frac{1}{Z^{\eta,h}_{N,m,\beta}} \exp(-\beta H_{N,m}(x) + \beta h \sum_i \xi^\eta_i x_i),$$

where $Z^{\eta,h}_{N,m,\beta}$ is the new normalising factor.

The Hamiltonian can also be expressed in terms of overlap parameters

$$v^\mu_N(x) = \frac{1}{N} \sum_i \xi^\mu_i x_i, \quad \text{for} \quad \mu = 1, \ldots, m,$$

by writing

$$H_{N,m}(x) = -\frac{N}{2} \sum_{\mu=1}^{m} (v^\mu_N(x))^2 = -\frac{N}{2} \|v_N(x)\|_2^2.$$

Therefore, the Hamiltonian depends on the configurations x only through the overlap parameters v. It is convenient to introduce *induced measures* on $\mathbb{R}^m$, instead of ordinary Gibbs measures, by

$$g^{\eta,h}_{N,m,\beta}(v) = \gamma^{\eta,h}_{N,m,\beta}(\{v(x) = v\})$$

for $v = (v^1, \ldots, v^m)$. For $\delta > 0$, write $a(\delta, \beta)$ for the largest solution of the equation $\delta a = \tanh(\beta a)$ and denote by $\| \cdot \|$ the ℓ^2 norm on $\mathbb{R}^{\mathbb{N}}$. Let $\lim \frac{m}{N} = \alpha$. For fixed β, for $\nu \in \mathbb{N}$, and $S \in \{-1, 1\}$, we define the ball

$$B^{(\nu,s)}_\rho = \{x \in \mathbb{R}^{\mathbb{N}} : \|x - sa(1 - 2\sqrt{\alpha}, \beta)e^\nu\| \leq \rho\},$$

where e^ν denotes the ν-th unit vector in $\mathbb{R}^{\mathbb{N}}$. With this notation, it is proven in [18] the following

Theorem 6.10 *There exists $\alpha_0 > 0$ such that, for every $\alpha \leq \alpha_0$ and every $\beta > 1 + 3\sqrt{\alpha}$, if $\rho^2 > C[a(1 - 2\sqrt{\alpha}, \beta)]^{3/2} \alpha^{1/8} |\log \alpha|^{1/4}$, almost surely,*

$$\lim_{h \downarrow 0} \lim_{N \uparrow \infty} g^{\eta,h}_{N,m,\beta}(B^{(\eta,+1)}_\rho) = 1.$$

So, this theorem guarantees that by using a small external field, eventually vanishing to zero, we can force the induced measure to be concentrated on a ball of overlap parameters slightly differing from a selected direction η. This result is the first rigorous indication that stochastic dynamics can be used to recover associatively the stored patterns in a large network. However, important problems remain open concerning the dynamical evolution of the net. For instance, it is not yet rigorously established whether a simulated annealing algorithm can be used to explore the stored patterns. Although such a possibility is expected, it remains to know how the various parameters have to be adjusted for such an algorithm to converge.

6.3 Self-averaging revisited

We are now able to state the result of [18] concerning the weak self-averaging property of the rate function. Denote

$$\phi_{N,\beta,\rho}(\tilde{v}) = -\frac{1}{\beta N} \log g_{N,\beta,h=0}(\|v_\Lambda - \tilde{v}\|_2^2 \leq \rho).$$

The function $\phi_{N,\beta,\rho}$ is the large deviation rate function governing the exponential convergence to zero of the corresponding probability. It is closely related to the free-energy since the latter can be expressed in a similar way where the restricting event $\|v_\Lambda - \tilde{v}\|_2^2 \leq \rho)$ is replace by the whole space.

It is proven in [18] the following

Theorem 6.11 *Assume that* $\lim_{N\to\infty} \frac{m}{N} = \alpha > 0$. *Let* $\rho < 1$ *and* $\|\tilde{v}\|_2$ *be bounded. Then, for every* $n \in \mathbb{N}$, *there exists* $t_n < \infty$ *such that* $\forall t \geq t_n$, *and for* N *large enough,*

$$\mathbb{P}(|\phi_{N,\beta,\rho}(\tilde{v}) - \mathbb{E}\phi_{N,\beta,\rho}(\tilde{v})| \geq \frac{t(\log N)^{3/2}}{\sqrt{N}}) \leq \frac{1}{N^n}.$$

The proof of this theorem consists in writing $\phi_{N,\beta,\rho}(\tilde{v}) - \mathbb{E}\phi_{N,\beta,\rho}(\tilde{v})$ as a martingale difference, following the ideas of Girko explained in the previous chapter, and then use precise estimates for these martingales differences. The proof, although quite straightforward, is somewhat technical to be reproduced *in extenso* here and the interested reader is directed to the original publication.

7 Conclusion

In this review, the profound relation existing between neural networks and statistical mechanics is shown. Due to the limited space and time available, only a selection of mathematical results is presented here.

Neural nets have found an extended field of applications in constructing engineering devices where they are used as an alternative to universal computers to perform tasks of categorisation, pattern recognition, forecasting, and so on. All these fascinating applications are missing from the present report. One may consult the books [53, 70, 45] to get a flavour of possible applications and the specialised journals like *Network, IEEE Neural Networks, IEEE Patterns Analysis and Machine Intelligence, ...* for some more finalised applications.

Biologists and neurophysiologists also use neural nets modelling to explain the functioning of the brain. Experiments in neurophysiology identify new characteristics of neural

behaviour and these are incorporated into more and more sophisticated neural models. All these results are missing from this review.

From the moment that the connection of neural networks with statistical mechanics was established, the subject became a branch of theoretical physics. Many interesting and intuitively appealing results were obtained by the physical community. Based on heuristic approaches like the replica trick, many qualitatively convincing results have been obtained. All these results are also missing from this survey. The interested reader is directed to [66, 25, 32, 31, 69, 54] for the most important of them.

The reader could expect therefore that all the mathematical aspects would be presented here. She will be disappointed: only results establishing that learning process is equivalent to an information increasing (entropy decreasing) process and that the restitution process is a Markov process converging to an invariant measure interpreted as the Gibbs measure of statistical mechanics is presented. Interesting issues as those describing dynamical evolution of neural nets with stochastic differential equations [10, 47] or with discrete time evolutions [11, 21] are omitted. But also mathematical results connected to more specific issues of neural networks such as biomathematical modelling of neural functions [24] or connections with graph theory are absent from this report.

The main reason for these omissions is that the subject is so vast that without limiting oneself in some clearly circumscribed region their is a risk of endless ramification. The second reason is that even in the restricted domain of thermodynamic formalism examined here, there are still numerous open problems. The known mathematical results only partially explain the numerically observed phenomena. Many questions pointed out in the main text — like the storage capacity, the stability and convergence under several steps of the dynamics, the study of systems with dependent patterns, etc. — remain unanswered for the moment. The mathematical methods presented here offer may be a good starting point to tackle these questions but new bright ideas are also certainly needed.

Acknowledgements: The author wishes to thank the Departamento de Ingeniería Matematica de l'Universidad de Chile for its kind invitation to give this series of lectures. He also acknowledges support from the EU network CHRX-CT93-0411 that produced a significant amount of the information contained here and favoured its dissemination. He wishes also to express his acknowledgements to Anton Bovier and Franck Vermet for their careful reading of the manuscript and their comments.

References

[1] L F Abbott, Learning in neural network memories, *Network* **1**, 105–122 (1990).

[2] M Aizenman, J L Lebowitz, D Ruelle, Some rigorous results on the Sherrington-Kirkpatrick model, *Commun. Math. Phys.*, **112**, 3–20 (1987).

[3] S Albeverio, B Tirozzi, B Zegarlinski, Rigorous results for the free energy in the Hopfield model, *Commun. Math. Phys.*, **150**, 337–373 (1992).

[4] J M G Amaro de Matos, A E Patrick, V A Zagrebnov, Random infinite volume Gibbs states for the Curie-Weiss random field Ising model, *J. Stat. Phys.*, **66**, 139–164 (1992).

[5] D J Amit, *Modelling brain function*, Cambridge University Press, Cambridge (1989).

[6] D J Amit, G Gutfreund, H Sompolinsky, Statistical mechanisms of neural networks near saturation, *Ann. Phys.*, **173**, 30–67 (1987).

[7] S Anily, A Federgruen, Simulated annealing methods with general acceptance probabilities, *J. Appl. Prob.* **24**, 657–667 (1968).

[8] J P Aubin, *Mathematical methods for neural networks*, lecture notes of a COMETT graduate module held in Les Houches, 16–29 March (1992).

[9] R Azencott, Simulated annealing, *Séminaire Bourbaki No. 697*, 1–15 (1988).

[10] G Ben Arous, A Guionnet, *Large deviations for Langevin spin glass dynamics*, preprint DMI, École Normale Supérieure (1995).

[11] O Bernier, Stochastic analysis of the dynamics of a general class of synchronous neural networks, *J. Phys. A: Math. Gen.*, **26**, 6879–6892 (1993).

[12] P Billingsley, *Convergence of probability measures*, Wiley, New York (1968).

[13] A Bovier, Self-averaging in a class of generalised Hopfield models, *J. Phys. A: Math. Gen.* **27**, 7069–7077 (1994).

[14] A Bovier, V Gayrard, Rigorous bounds on the storage capacity of the dilute Hopfield model, *J. Stat. Phys.* **69**, 597–627 (1992).

[15] A Bovier, V Gayrard, Rigorous results on the thermodynamics of the dilute Hopfield model, *J. Stat. Phys.*, **72**, 79–112 (1993).

[16] A Bovier, V Gayrard, *An almost sure large deviation principle for the Hopfield model*, Weierstraß IAAS preprint (1995).

[17] A Bovier, V Gayrard, P Picco, Gibbs states of the Hopfield model in the regime of perfect memory, *Prob. Th. Rel. Fields*, **100**, 329–363 (1994).

[18] A Bovier, V Gayrard, P Picco, Gibbs states of the Hopfield model with extensively many patterns, *J. Stat. Phys.*, **79**, 395–414 (1995).

[19] A Bovier, V Gayrard, P Picco, *Large deviation principles for the Hopfield and the Kac-Hopfield model*, Weirestraß IAAS preprint (1994).

[20] C van den Broeck, Statistical physics of learning from examples: a brief introduction, *Acta Phys. Polon. B*, **25**, 903–923 (1994).

[21] M Cassandro, A Galves, E Olivieri, M E Vares, Metastable behaviour of stochastic dynamics: a pathwise approach, *J. Stat. Phys.*, **35**, 603–??? (1984).

[22] J-P Changeux, *L'homme neuronal*, Fayard, Paris (1983).

[23] P Collet, J-P Eckmann, *Iterated maps on the interval as dynamical systems*, Birkhäuser, Basel (1980).

[24] M Cottrell, Mathematical analysis of a neural network with inhibitory coupling, *Stoch. Proc. Appl.*, **40**, 103–126 (1992).

[25] B Derrida, R B Griffiths, A Prügel-Bennett, Finite-size effects and bounds for perceptron models, *J. Phys. A: Math. Gen.*, **24**, 4907–4940 (1991).

[26] R L Dobrushin, The description of a random field by means of conditional probabilities and condition of its regularities, *Th. Prob. Appl.*, **13**, 458–486 (1968).

[27] C Fassnacht, A Zippelius, A recognition and categorisation in a structured neural network with attractor dynamics, *Network*, **2**, 63–84 (199?).

[28] P A Ferrari, S Martínez, P Picco, A lower bound for the memory capacity in the Potts-Hopfield model, *J. Stat. Phys.*, **66**, 1643–1652 (1992).

[29] R Folk, A Kartashov, P Lisoněk, P Paule, Symmetries in neural networks: a linear group action approach, *J. Phys. A: Math. Gen.*, **26**, 3159–3164 (1993).

[30] M I Freidlin, A D Wentzell, *Random perturbations of dynamical systems*, Springer-Verlag, Berlin (1984).

[31] E Gardner, The space of interactions in neural network models, *J. Phys. A: Math. Gen.*, **21**, 257–270 (1987).

[32] E Gardner, B Derrida, Optimal storage properties of neural network models, *J. Phys. A: Math. Gen.*, **21**, 271–284 (1988).

[33] V Gayrard, The thermodynamic limit of the Potts-Hopfield model for infinitely many patterns, *J. Stat. Phys.*, **68**, 977–1011 (1992).

[34] H-O Georgii, *Gibbs measures and phase transitions*, Walter de Gruyter, Berlin (1988).

[35] V L Girko, *Random matrices*, Vishcha Shkola, Izdat. Kiev Univ., Kiev (1975).

[36] V L Girko, Limit theorems for maximal and minimal eignevalues of random matrices, *Th. Prob. Appli.*, **35**, 680–695 (1988).

[37] V L Girko, *Theory of random determinants*, Kluwer, Dodrecht (1990).

[38] E Goles, S Martínez, *Neural and automata networks*, Kluwer Academic Publ., Dodrecht (1990).

[39] E Goles, S Martínez, *Statistical Physics, automata networks, and dynamical systems*, Kluwer Academic Publ., Dodrecht (1992).

[40] M Griniasty, M V Tsodyks, D J Amit, *Conversion of temporal correlations between stimuli to spatial correlations between attractors*, preprint Università di Roma 1 (1992).

[41] H Haario, E Saksman, Simulated annealing process in general state space, *Adv. Appl. Prob.*, **23**, 866–893 (1991).

[42] D Hebb, *The organisation of behaviour: a neurolophysiological theory*, Wiley, New York (1949).

[43] J L van Hemmen, D Grensing, A Huber, R Kühn, Nonlinear neural networks I: general theory, *J. Stat. Phys.*, **50**, 231–257 (1988).

[44] J L van Hemmen, D Grensing, A Huber, R Kühn, Nonlinear neural networks II: information processing, *J. Stat. Phys.*, **50**, 259–293 (1988).

[45] J Hertz, A Krogh, R Palmer, *Introduction to the theory of neural computation,* Addison-Wesley, Redwood City CA (1991).

[46] J J Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci.*, **79**, 2554–2558 (1982).

[47] G Kallianpur, *Weak convergence of stochastic neuronal models*, in Stochastic methods in Biology, Nagoya (1985), M Kimura, G Kallianpur, and T Hida (eds.).

[48] A I Khinchin, *Mathematical foundations of information theory*, Dover, New York (1957).

[49] H Koch, A free energy bound for the Hopfield model, *J. Phys. A: Math. Gen.*, **26**, L353–L355 (1993).

[50] H Koch, J Piasko, Some rigorous results on the Hopfield neural network model, *J. Stat. Phys.*, **55**, 903–928 (1989).

[51] A N Kolmogorov, *Foundations of probability theory*, Chelsea Publishing, New York (1950).

[52] J Komlós, R Paturi, Convergence results in a autoassociative memory model, *Neural Networks*, **1**, 239–250 (1988).

[53] B Kosko, *Neural networks and fuzzy systems*, Prentice-Hall, London (1992).

[54] W Krauth, M Mézard, Storage capacity of memory networks with binary couplings, *J. Physique*, **50**, 3057–3066 (1989).

[55] Y Kuramoto, *Chemical oscillation, waves, and turbulence*, Springer-Verlag, Berlin (1984).

[56] O E Lanford III, D Ruelle, Observables at infinity and states with short range correlations in statistical mechanics, *Commun. Math. Phys.* **13**, 194-215 (1969).

[57] S Lang, *Real and functional analysis*, Springer-Verlag, Berlin (1993).

[58] M Ledoux, M Talagrand, *Probability in Banach spaces*, Springer-Verlag, Berlin (1991).

[59] W Little, G Shaw, Analytic study of the memory storage capacity of a neural network, *Math. Biosc.*, **39**, 281–290 (1978).

[60] D Loukianova, Capacité de mémoire dans le modèle de Hopfield, *C. R. Acad. Sci. Paris*, **318**, 157–160 (1994).

[61] D Loukianova, *Étude rigoureuse du modèle de mémoire associative*, PhD Thesis, Université de Paris VII, presented on 2 December (1994).

[62] R MacEliece, E Posner, E Rodemich, S Venkatesh, The capacity of the Hopfield associative memory, *IEEE Trans. Inf. Theory*, **33**, 461–482 (1987).

[63] C Maes, V Zagrebnov, On the parallel dynamics of a multi-layered perceptron, (1991).

[64] S Martínez, *Introduction to neural networks: storage capacity and optimisation*, proceedings CIMPA school on "Dynamical and disordered systems", Temuco (1992).

[65] W S McCulloch, W Pitts, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.* **5**, 115–133 (1943).

[66] M Mézard, G Parisi, M A Virasoro, *Spin-glass theory and beyond*, World scientific, Singapore (1988).

[67] R Minlos, Gibbs' limit distribution, *Funct. Anal. Appl.*, **2**, 60–73; **3**, 40–53 (1967).

[68] Y Miyashita, Neuronal correlate of visual associative long-term memory in the primate temporal cortex, *Nature*, **335**, 817–819 (1988).

[69] R Monasson, Properties of neural network storing spatially correlated patterns, *J. Phys. A: Math. Gen.*, **25**, 3701–3720 (1992).

[70] B Müller, J Reinhardt, *Neural Networks*, Springer-Verlag, Berlin (1990).

[71] C Newman, Memory capacity in neural networks, *Neural Networks*, **1**, 223–238 (1988).

[72] N J Nilson, *Learning machines*, McGraw-Hill, New York (1965).

[73] H Nishimori, W Whyte, D Sherrington, *Finite-dimensional neural networks storing structured patterns*, preprint University of Oxford (1994).

[74] G Parisi, *Attractor neural networks*, preprint (1994) available from `cond-mat@babbage.sissa.it` under reference `9412030`.

[75] L A Pastur, A L Figotin, Exactly solvable model of a spin glass, *Sov. J. Low Temp. Phys.*, **3**, 378–383 (1977).

[76] L A Pastur, A L Figotin, On the theory of disordered spin systems, *Theor. Math. Phys.*, **35**, 404–414 (1978).

[77] L A Pastur, M Shcherbina, Absence of self-averaging of the order parameter in the Sherrington-Kirkpatrick model, *J. Stat. Phys.*, **62**, 1–19 (1991).

[78] L A Pastur, M Shcherbina, B Tirozzi, The replica symmetric solution without replica trick for the Hopfield model, *J. Stat. Phys.*, **74**, 1161–1183 (1994).

[79] A E Patrick, V A Zagrebnov, Parallel dynamics for an extremely diluted neural network, *J. Phys. A: Math. Gen.*, **23**, L1323–L1329 (1990).

[80] A E Patrick, V A Zagrebnov, A probabilistic approach to parallel dynamics for the Little Hopfield model, *J. Phys. A: Math. Gen.*, **24**, 3413–3426 (1991).

[81] D Petritis, *Simulations numériques Monte Carlo*, preprint Université de Rennes I, to be published by Masson, Paris (1995).

[82] D Petritis, *Equilibrium statistical mechanics of frustrated disordered systems: a survey of mathematical results*, preprint Université de Rennes I (1994).

[83] P Picco, *Artificial neural networks*, preprint CNRS Marseille (1995).

[84] G Radons, H G Schuster, D Werner, Fractal measures and diffusion as results of learning in neural networks, *Phys. Lett. A*, **174**, 293–297 (1993).

[85] R T Rockafellar, *Convex analysis*, Princeton Univ. Press, Princeton (1970).

[86] F Rosenblatt, *Principles of neurodynamics*, Spartan, New York (1962).

[87] D Ruelle, *Thermodynamic formalism*, Addison-Wesley, Reading (1978).

[88] D E Rumelhart, G E Hinton, R J Williams, Learning representations by back-propagating errors, *Nature*, **323**, 533–??? (1986).

[89] H Sakaguchi, Learning rules for an oscillator network, *Phys. Lett. A*, **174**, 289–292 (1993).

[90] E Scacciatelli, B Tirrozi, Fluctuation of the free energy in the Hopfield model, *J. Stat. Phys.*, **67**, 981–1008 (1992).

[91] H Scharze, J Hertz, Learning from examples in fully connected committee machines, *J. Phys. A: Math. Gen.*, **26**, 4919–4936 (1993).

[92] M Schlüter, E Wagner, ???, *Phys. Rev.*, **E49**, 1690–???? (1994).

[93] M Shcherbina, B Tirozzi, The free energy for for a class of Hopfield models, *J. Stat. Phys.*, **72**, 113–125 (1993).

[94] B Simon, The statistical mechanics of lattice gases, Princeton University press, Princeton (1993).

[95] S Solla, *Learning and generalisation in layered neural networks*, in Redes neuronales, teoria y aplicaciones, Escuela de Verano 88 en Fisica estadistica y sistemas cooperativos (1988).

[96] S Solla, A theory of supervised learning, in *Neural networks: from biology to high energy physics*, Proc. Elba Int. Physics Centre, O Benhar, C Bosio, P del Giudice, E Tabet eds. ETS Editrice, Pisa (1991).

[97] M Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, preprint Université Paris VI (1995).

[98] M Talagrand, *A new look at independence*, preprint Université Paris VI (1995).

[99] M Talagrand, *Résultats rigoureux pour le modèle de Hopfield*, preprint Université de Paris VI (1995).

[100] W Tarkowski, M Lewenstein, Storage of sets of correlated data in neural network memories, *J. Phys. A: Math. Gen.*, **26**, 2453–2469 (1993).

[101] N Tishby, E Levin, S Solla, Consistent inference of probabilities in layered networks: predictions and generalisation, *IEEE Neural Net.*, **2**, 403–410 (1989).

[102] F Vermet, F, Convergence de la variance de l'énergie libre du modèle de Hopfield, *C. R. Acad. Sci. Paris*, **315**, 1001–1004 (1992).

[103] F Vermet, *Étude asymptotique d'un réseau neuronal : le modèle de mémoire associative de Hopfield*, PhD Thesis, Université de Rennes I, presented on 28 January (1994).

[104] F Vermet, *Asymptotic study of a neural network*, preprint Université de Rennes I (1994).