arXiv:1105.2790v2 [cond-mat.dis-nn] 16 May 2011

On the equivalence of Hopfield Networks and Boltzmann Machines

Adriano Barra[*], Alberto Bernacchia [†], Enrica Santucci [‡] and Pierluigi Contucci [§]

March 2011

Abstract

A specific type of the neural networks, the Restricted Boltzmann Machines (RBM), are implemented for classification and feature detection. They are characterized by separate layers of visible and hidden units, which are able to learn efficiently a generative model of the observed data. We study a "hybrid" version of RBM's, in which hidden units are analog and visible units are binary, and we show that the evolution and the thermodynamics of visible units are equivalent to those of a Hopfield network, in which the N visible units are the neurons and the P hidden units are the learned patterns.
We apply the method of stochastic stability to derive the thermodynamics of the machine, by considering a formal extension of this technique to the case of multiple sets of stored patterns, which may act as a benchmark for the study of correlated sets.
Our results imply that simulating the dynamics of a Hopfield network, requiring the update of N neurons and the storage of $N(N-1)/2$ synapses, can be accomplished by a hybrid Boltzmann Machine, requiring the update of $N+P$ neurons but only the storage of NP synapses. In addition, the well known glass transition of the Hopfield network has a counterpart in the Boltzmann Machine: It corresponds to an optimum criterion for selecting the relative sizes of the hidden and visible layers, resolving the trade-off between flexibility and generality of the model. The low storage phase of the Hopfield model corresponds to few hidden units and hence a very constrained RBM, while the spin-glass phase (too many hidden units) corresponds to an overly unconstrained RBM prone to overfitting of the observed data.

1 Introduction

A common problem in Machine Learning is to design a device able to copy, or reproduce a system, in the case in which a satisfying model of the system is not available and its underlying principles are not known [15]: In this case, learning to reproduce the system can be achieved only by the observation of a large number of samples. Here, "reproducing" is intended in a statistical sense, namely learning the arbitrarily complex distribution of all possible states of the system and reproducing those states with the correct likelihood [11]. For example, despite several attempts [23], no accurate model exists of the visual world (the set of all possible visual

[*]Dipartimento di Fisica, Sapienza Università di Roma.
[†]Department of Neurobiology, Yale University.
[‡]Dipartimento di Matematica, Università degli Studi dell'Aquila.
[§]Dipartimento di Matematica, Alma Mater Studiorum Università di Bologna.

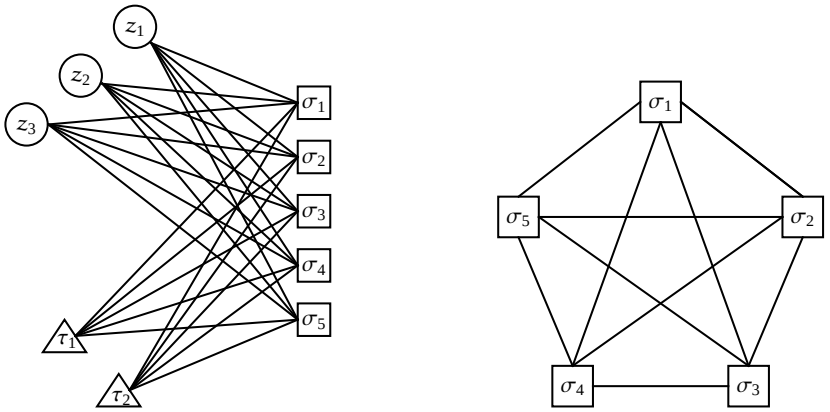

Figure 1: Left panel: Schematic representation of a Hybrid Boltzmann Machine (HBM) where the hidden units are analog (z, τ variables) and the visible units are digital (σ variables). The two sets of hidden units, z and τ, represent two feature sets that are both connected to the layer of visible units σ. The layers of hidden and visible units are reciprocally connected, but there are no intra-layer connections, thus forming a bipartite topology. Right panel: Schematic representation of the equivalent Hopfield neural network built up only by the visible units, with an internal fully connected structure.

stimuli). However, the ability to predict which stimuli will be encountered, and which are more or less likely to happen, is fundamental for the survival of living organisms. It is likely that the brain develops an internal model of the visual world, and is able to estimate the probability and respond to the occurrence of different events [6],[7].

Ising-type neural networks have been widely used as generative models of simple systems [16],[3]. Those models are generative in the sense that, after learning the synaptic weights from a set of observations (neural activations), the network is able to generate a sequence of states whose probabilities match those of the observations. Ising models are characterized by a quadratic energy function and a Boltzmann distribution of states, and a very popular example is the Hopfield model [2][19]. Boltzmann Machines (BM) have been designed to capture the complex statistics of arbitrary systems by dividing the neurons in two subsets, visible and hidden units: marginalizing the Boltzmann distribution over the hidden units allows the BM to reproduce, through the visible units, arbitrarily complex distribution of states, by learning the appropriate synaptic weights [17]. Efficient learning algorithms, used in state of the art feature detectors and classifiers [9], have been designed for a specific type of BM, the Restricted Boltzmann Machines (RBM), characterized by a bipartite topology in which hidden and visible units are coupled, but there is no interaction within either set of visible or hidden units [18] (see Fig.1 left).

All units of RBM's are binary, while the analog equivalent of RBM, the Restricted Diffusion Networks, have been described in [20][9], in which all units are analog. Here we study the case of a "hybrid" Boltzmann Machine (HBM), in which the hidden units are analog and the visible units are binary (the case of standard RBMs with all binary variables can be derived straightforwardly from our approach). We show that the HBM, when marginalized over the hidden units, is exactly equivalent to a Hopfield network, where the N visible units are the neurons and the P hidden units are the learned patterns. While the Hopfield network is limited in the space of

possible probability distributions it can generate, it has been widely studied for its associative and retrieval properties. The exact mapping introduces a new way to simulate Hopfield networks, and allows a novel interpretation of the spin glass transition, which in the HBM translates into an optimal criterion for selecting the relative size of the hidden and visible layers (see Conclusions). In order to derive the exact thermodynamics of the system, we use the method of stochastic stability previously described in [1], [4]: While the natural technique of investigation of the Hopfield model is the replica trick, the HBM can be studied by properly adapting the stochastic stability to this case. We analyze the model with two non-interacting sets of hidden units (uncorrelated patterns), and solve the thermodynamic at the replica symmetry level. We then extend the theory to cope with two sets of interacting hidden layers and show how their interaction acts as a further noise source for the retrieval.

2 Description of the model

We define a "hybrid" Boltzmann Machine (HBM, see Fig. 1 left) as a network in which the activity of units in the visible layer is discrete, $\sigma_i = \pm 1,\ i \in (1, ..., N)$ (digital layer), and the activity in the hidden layer is continuous (analog layer). For the sake of generality, we assume that the layer of hidden units is further divided in two sets, both described by continuous variables, $z_\mu, \tau_\nu \in \Re,\ \mu \in (1, .., P),\ \nu \in (1, ..., K)$. The layers of hidden and visible units are reciprocally connected, but there are no intra-layer connections, thus forming a bipartite topology. We will consider the case of interacting hidden units (connections between z and τ) in Section 3.5. The synaptic connections between the units in the two layers are fixed and symmetric, and are defined by the synaptic matrix ξ_i^μ, η_i^ν. The input to unit i in the digital layer is the sum of the activities in the analog layers weighted by the synaptic matrix, i.e. $\sum_\mu \xi_i^\mu z_\mu + \sum_\nu \eta_i^\nu \tau_\nu$, while the input to units $\mu,\ \nu$ in the analog layers is the sum of the activities in the digital layer only, again weighted by the (symmetric) synaptic matrix, i.e. $\sum_i \xi_i^\mu \sigma_i$, $\sum_i \eta_i^\nu \sigma_i$. The dynamics of the activity of the units is different in the two layers; in the analog layers it changes continuously in time, while in the digital counterpart it changes in discrete steps.

The activity in the analog layer follows the stochastic differential equations

$$T\frac{dz_\mu}{dt} = -z_\mu(t) + \sum_i \xi_i^\mu \sigma_i + \sqrt{\frac{2T}{\beta}}\,\zeta_\mu(t) \tag{1}$$

$$T\frac{d\tau_\nu}{dt} = -\tau_\nu(t) + \sum_i \eta_i^\nu \sigma_i + \sqrt{\frac{2T}{\beta}}\,\chi_\nu(t) \tag{2}$$

where ζ, χ are white gaussian noise with zero mean and covariance $\langle \zeta_\mu(t)\zeta_\nu(t')\rangle = \delta_{\mu\nu}\ \delta(t-t')$, $\langle \chi_\mu(t)\chi_\nu(t')\rangle = \delta_{\mu\nu}\ \delta(t-t')$, $\langle \zeta_\mu(t)\chi_\nu(t')\rangle = 0$. The parameter T quantifies the timescale of the dynamics, and the parameter β determines the strength of the fluctuations. The process described by the above equations is akin to an Ornstein-Uhlembeck diffusion process [25], where the first term in the right hand sides describes an energy leakage, the second term the input signals and the third term a noise source. Note that the activity of units in the digital layer, σ_i, also depends on time, and we assume here that the timescale of diffusion T is much faster than the rate at which the digital units are updated (see below). Under this assumption, the process corresponds to a diffusion in the quadratic potential

$$E(z,\tau) = \frac{\beta}{2}\Big(\sum_{\mu} z_{\mu}^2 + \sum_{\nu} \tau_{\nu}^2\Big) - \beta\left(\sum_{i,\mu} \sigma_i \xi_i^{\mu} z_{\mu} + \sum_{i,\nu} \sigma_i \eta_i^{\nu} \tau_{\nu}\right) \tag{3}$$

and the equilibrium distribution for the activity of the analog units is proportional to $e^{-E(z,\tau)}$.

The activity in the digital layer follows a standard Glauber dynamics for Ising-type systems. At a specified sequence of time intervals (much larger than T), the activity of a unit in the digital layer is updated according to its input, where the probability of the unit's activity is equal to a rectified value of the input (logit transfer function), i.e.

$$Pr(\sigma_i(t) = \pm 1) = \frac{1}{1 + \exp[\mp 2\beta(\sum_{\mu} \xi_i^{\mu} z_{\mu}(t) + \sum_{\nu} \eta_i^{\nu} \tau_{\nu}(t)]}. \tag{4}$$

When updating the digital units according to the above expression of the probability, the analog variables are "frozen", namely, the update of the digital units is instantaneous. Therefore, the dynamics of analog and digital units occurs separately, and no concurrent update is implemented. The Glauber dynamics defined by the above equation implies that the equilibrium distribution for the digital units is proportional to the distribution $\exp[\beta(\sum_{i,\mu} \sigma_i \xi_i^{\mu} z_{\mu} + \sum_{i,\nu} \sigma_i \eta_i^{\nu} \tau_{\nu})]$. Since the dynamics of the digital and analog units occurs separately, the whole system including both layers admits the joint equilibrium distribution

$$P(\{z_{\mu}\}, \{\tau_{\nu}\}, \{\sigma_i\}) \propto \exp\Big(-\frac{\beta}{2}(\sum_{\mu} z_{\mu}^2 + \sum_{\nu} \tau_{\nu}^2) + \beta(\sum_{i,\mu} \sigma_i \xi_i^{\mu} z_{\mu} + \sum_{i,\nu} \sigma_i \eta_i^{\nu} \tau_{\nu})\Big), \tag{5}$$

and consequently an Hamiltonian representation as

$$H(\sigma, z, \tau; \xi, \eta) = \frac{1}{2}(\sum_{\mu} z_{\mu}^2 + \sum_{\nu} \tau_{\nu}^2) - \sum_{i,\mu} \sigma_i \xi_i^{\mu} z_{\mu} - \sum_{i,\nu} \sigma_i \eta_i^{\nu} \tau_{\nu}. \tag{6}$$

where the joint equilibrium distribution is proportional to $\exp(-\beta H)$
As a consequence, we can apply from now on methods stemmed from statistical mechanics in the canonical ensemble.

3 The statistical mechanics approach

3.1 Thermodynamical equivalence of HBM and Hopfield networks

We are interested in the thermodynamical properties (i.e. phase diagram, critical capacity, blackouts) of the system defined through eq. (6): Within the standard canonical ensemble, we define the partition function $Z_{N,P,K}$ as

$$Z_{N,P,K}(\beta; \xi, \eta) = \sum_{\sigma} \int \prod_{\mu=1}^{P} dz_{\mu} \int \prod_{\nu=1}^{K} d\tau_{\nu} \exp\left(-\beta H(\sigma, z, \tau; \xi, \eta)\right), \tag{7}$$

The equivalence of HBM and the Hopfield network is evident by marginalizing over the analog variables, i.e. performing the Gaussian integration over the one-body terms in z, τ:

$$Z_{N,P,K}(\beta; \xi, \eta) \propto \sum_{\sigma} \exp(-\beta H(\sigma; \xi, \eta)), \tag{8}$$

where $H(\sigma;\xi,\eta)$ is the following reduced Hamiltonian:

$$H(\sigma;\xi,\eta) = -\frac{1}{2}\sum_{i,j}^{N}\Big(\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu + \sum_{\nu=1}^{K}\eta_i^\nu\eta_j^\nu\Big)\sigma_i\sigma_j, \tag{9}$$

This is the Hamiltonian of a Hopfield neural network whose kernel is the sum of the contribution of two subsets of analog (hidden) variables.
This result connects the two Hamiltonians of the Hopfield network and the Boltzmann machine and states that the thermodynamics obtained by the first cost function is the same as the one obtained by the second one, implicitly offering a connection among free energy minimization for the retrieval in the former case and log-likelihood estimation for learning in the latter [2][9] provided that the noise level of the former is the square of the noise level of the latter.

3.2 Definitions

In order to study the thermodynamics of the system, in this section we make a list of useful definitions. We introduce the Boltzmann states ω as

$$\omega_\beta(O) = (Z_{N,P,K}(\beta;\xi,\eta))^{-1}\sum_{\sigma}\int\prod_{\mu=1}^{P}dz_\mu\int\prod_{\nu=1}^{K}d\tau_\nu\; O(\sigma)\exp\left(-\beta H(\sigma,z,\tau;\xi,\eta)\right). \tag{10}$$

$\Omega_s = \omega_1\times\omega_2\times...\times\omega_s$ is introduced as the s-replicated state (as we will average over different replicas of the system to account for the weights ξ,η). We define the operator $\mathbb{E}$ as

$$\mathbb{E}[F(\xi,\eta)] = \int d\mu(\xi)\int d\mu(\eta)F(\xi,\eta), \tag{11}$$

where μ is the standard Gaussian measure. We define the main quantity of interest, the quenched free energy $A_{N,P,K}(\beta)$ as

$$A_{N,P,K}(\beta) = \frac{1}{N}\mathbb{E}\log Z_{N,P,K}(\beta;\xi,\eta). \tag{12}$$

In Hopfield networks, the regime of high storage corresponds to the case in which the number of stored patterns is linearly increasing with the number of neurons [2]. In HBM, this corresponds to the case in which the sizes of the hidden and visible layers are comparable. Their relative size is defined by introducing the two control parameters $\alpha,\gamma\in\mathbb{R}^+$ as

$$\lim_{N\to\infty}\frac{P}{N} = \alpha \quad \lim_{N\to\infty}\frac{K}{N} = \gamma, \tag{13}$$

so that the total amount of load is $\alpha+\gamma$. We further introduce the order parameters q,p,r, the overlaps, as

$$q_{ab} = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^a\sigma_i^b,\;\; p_{ab} = \frac{1}{P}\sum_{\mu=1}^{P}z_\mu^a z_\mu^b,\;\; r_{ab} = \frac{1}{K}\sum_{\nu=1}^{K}\tau_\nu^a\tau_\nu^b. \tag{14}$$

Our goal is to minimize the free energy with respect to the order parameters. This corresponds to energy minimization and entropy maximization and allows, by looking at its changes (phase transitions), to obtain a phase diagram highlighting where the HBM (in the α,β,γ volume) may work akin to a cooperative associative network.

3.3 The triple stochastic stability

Once we demonstrated that studying the thermodynamics of the HBM is the same as of studying the thermodynamic of the Hopfield network, we focus on the former and extend the method of the stochastic stability [1] (on the same line as recently achieved by Guerra and coworkers [4] in the neural scenario), whose scheme is quite intuitive:
Because those systems are mean field models, the overall stimulus felt by an element of a given layer is the sum (synaptically weighted) of all the entries by the other layers: For instance the field acting on the variable σ_i is $\phi_i = \sum_\mu \xi_i^\mu z_\mu + \sum_\nu \eta_i^\nu \tau_\nu$.
However, if we are able to substitute the real stimuli (which even in principle are extremely hard to be evaluated) with analytically tractable fictitious ones -denoted by a tilde in the following- whose probability distributions are shared by the formers, thermodynamical observable would be unaffected but the calculations much easier: For instance, still focusing on σ_i we replace ϕ_i with $\tilde{\eta}_i$, a standard $\mathcal{N}[0,1]$ variable.
To accomplish this substitution smoothly, we introduce an extended free energy, which depends further on the interpolating parameter $t \in [0,1]$, such that for $t = 1$ the original free energy of the model is recovered, while for $t = 0$ a straightforward free energy is achieved (obtained from a system without two-body terms), as an immediate check trough eq. (15) may confirm:

$$\tilde{A}_{N,P,K}(\beta,t) = \frac{1}{N}\mathbb{E}\log\sum_\sigma \int \prod_{\mu=1}^{P} d\mu(z_\mu) \int \prod_{\nu=1}^{K} d\mu(\tau_\nu) \exp\sqrt{t}\Big(- \beta H_N(\sigma,z,\tau;\xi,\eta)\Big)$$
$$\exp\sqrt{1-t}\Big(a\sum_{i=1}^{N}\tilde{\eta}_i\sigma_i + b\sum_{\mu=1}^{P}\tilde{\eta}_\mu z_\mu + c\sum_{\nu=1}^{K}\tilde{\eta}_\nu\tau_\nu\Big)\exp\Big[(1-t)\Big(\frac{h}{2}\sum_{\mu=1}^{P}z_\mu^2 + \frac{\epsilon}{2}\sum_{\nu=1}^{K}\tau_\nu^2\Big)\Big]. \qquad (15)$$

The two expressions $\tilde{A}(\beta, t=0)$ and $\tilde{A}(\beta, t=1)$ are not equivalent, but we can express the difference between them in terms of a sum-rule:

$$\tilde{A}_{N,P,K}(\beta,t=1) = \tilde{A}_{N,P,K}(\beta,t=0) + \int_0^1 dt' \Big(\frac{d\tilde{A}_{N,P,K}(\beta,t)}{dt}\Big)_{t=t'} \qquad (16)$$

In order to tackle $\tilde{A}(\beta,t=1)$, the interpolating free energy at $t=0$ must be evaluated (which is a sum of known one body problems), and the the integral over the support $[0,1]$ of its t-derivative. So far no particular advantage is evident in this sum rule (as, despite the simple calculation of $\tilde{A}(\beta,t=0)$, the integration of $\partial_t\tilde{A}(\beta,t)$ is in generally prohibitive), however, we will show that, by properly choosing the scalars $a,b,c,h,\epsilon \in \mathbb{R}^+$ we can obtain sum rules as $A(t=1) = \mathbb{A} + \mathbb{F}$, where $\mathbb{A}$ stands for terms of order one ("average" of the observable) and $\mathfrak{F}$ for terms of order two ("fluctuations" of the observable from its average), such that, neglecting $\mathbb{F}$ -and easily keeping the theory fully solvable-, we can retain a meaningful, replica symmetric, scenario[1].
Let us now focus on the derivative. As the interpolating parameter t appears seven times in the exponential, its derivative contains seven different terms, that we label using capital letters: $\dot{A} = A + B + C + D + E + F + G$[2]. Their derivation is long but straightforward, here we report

[1] The replica symmetric solution is the standard level of approximation used for the calculus of collective properties of neural networks even in standard approaches, i.e. the replica trick [2, 11, 16].

[2] For the sake of clearness we note that A in this sequence does not represent the free energy but only a term in the derivation.

directly the results:

$$A = -\frac{\alpha\beta}{2}\langle q_{12}p_{12}\rangle + \frac{\beta}{2N}\mathbb{E}\sum_{\mu}\omega(z_\mu^2), \tag{17}$$

$$B = -\frac{\gamma\beta}{2}\langle q_{12}r_{12}\rangle + \frac{\beta}{2N}\mathbb{E}\sum_{\nu}\omega(\tau_\nu^2), \tag{18}$$

$$C = -\frac{a^2}{2}(1-\langle q_{12}\rangle), \tag{19}$$

$$D = \frac{\alpha b^2}{2}\langle p_{12}\rangle - \frac{b^2}{2N}\mathbb{E}\sum_{\mu}\omega(z_\mu^2), \tag{20}$$

$$E = \frac{\gamma c^2}{2}\langle r_{12}\rangle - \frac{c^2}{2N}\mathbb{E}\sum_{\nu}\omega(\tau_\nu^2), \tag{21}$$

$$F = \frac{h}{2N}\sum_{\mu}\mathbb{E}\sum_{\mu}\omega(z_\mu^2), \tag{22}$$

$$G = \frac{\epsilon}{2N}\sum_{\mu}\mathbb{E}\sum_{\nu}\omega(\tau_\nu^2). \tag{23}$$

Pasting the various terms together we get

$$\begin{aligned}\dot{\tilde{A}} =& \mathbb{E}\sum_{\mu}\omega(z_\mu^2)\Big(\frac{\beta}{2N}-\frac{b^2}{2N}-\frac{h}{2N}\Big)+\mathbb{E}\sum_{\nu}\omega(\tau_\nu^2)\Big(\frac{\beta}{2N}-\frac{c^2}{2N}-\frac{\epsilon}{2N}\Big)-\frac{\alpha\beta}{2}\langle q_{12}p_{12}\rangle \\ &-\frac{\gamma\beta}{2}\langle q_{12}r_{12}\rangle-\frac{a^2}{2}(1-\langle q_{12}\rangle)+\frac{\alpha b^2}{2}\langle p_{12}\rangle+\frac{\gamma c^2}{2}\langle r_{12}\rangle.\end{aligned} \tag{24}$$

Now we are left with the freedom of choosing the most convenient parameters; we see that with the particular choice

$$a=\sqrt{\beta(\alpha\bar{p}+\gamma\bar{r})}, \qquad b=\sqrt{\beta\bar{q}}, \qquad c=\sqrt{\beta\bar{q}}, \qquad h=\epsilon=\beta(1-\bar{q}),$$

we can express the whole derivative as the source of the overlap fluctuations:

$$\dot{\tilde{A}} = S(\alpha,\beta,\gamma)+\frac{\beta}{2}(\bar{q}-1)(\alpha\bar{p}+\gamma\bar{r})-\frac{\beta(\alpha+\gamma)}{2}, \tag{25}$$

$$S(\alpha,\beta,\gamma) = -\frac{\beta}{2}\langle(q_{12}-\bar{q})[\alpha(p_{12}-\bar{p})+\gamma(r_{12}-\bar{r})]\rangle. \tag{26}$$

3.4 Replica-symmetric sum rule

Neglecting in the future the source term $S(\alpha,\beta,\gamma)$ implies managing a theory where overlaps among different replicas are delta distributed, namely a replica symmetric approximation.
We can turn our attention now to the average-source term, the $t=0$ evaluation, whose calculation works by direct evaluation and allows the following representation

$$\begin{aligned}\tilde{A}(t=0) =& \log 2+\int d\mu(\eta)\log\cosh(\sqrt{\beta(\alpha\bar{p}+\gamma\bar{r})}\eta)+\frac{\alpha+\gamma}{2}\log(1-\beta(1-\bar{q}))^{-1}+ \\ &\frac{\beta(\alpha+\gamma)}{2}\frac{\bar{q}}{1-\beta(1-\bar{q})}.\end{aligned} \tag{27}$$

At the end we have our replica symmetric free energy written as a trivial sum rule: $\tilde{A}_{RS}(\beta, t = 1) = \tilde{A}(\beta, t = 0) + \beta(\bar{q} - 1)(\alpha\bar{p} + \gamma\bar{r})/2 - \beta(\alpha + \gamma)/2$.

Now, once the general strategy has been outlined, in order to introduce further a Mattis magnetization term to check retrival (to mirror standard AGS theory [2]), we introduce a Lagrange multiplier inside the free energy, still within the outlined spirit of the interpolation[3]; Using m_1 for the real magnetization (symbolically coupled to the first pattern $\mu = 1$) and M for its replica-symmetric approximation, the sum rule extends to

$$\begin{aligned}
&\frac{1}{N}\mathbb{E}(\log Z_{N,P,K}(\beta, \xi, \eta)) + \frac{\beta}{2}\int_0^1 dt\langle(q_{12} - \bar{q})[\alpha(p_{12} - \bar{p}) + \gamma(r_{12} - \bar{r})]\rangle \\
&= \frac{\beta}{2}\int_0^1 dt\langle(m_1 - M)^2\rangle + \bar{A}(\bar{p}, \bar{q}, \bar{r}, M; \alpha, \beta, \gamma)
\end{aligned} \tag{28}$$

by which the final replica symmetric free energy can be written as

$$\begin{aligned}
\bar{A}(\bar{p}, \bar{q}, \bar{r}, M; \alpha, \beta, \gamma) &= \log 2 + \int d\mu(\eta) \log\cosh\Big(\eta\sqrt{\beta(\alpha\bar{p} + \gamma\bar{r}) + \beta^2 M^2}\Big) \\
&+ \frac{\alpha + \gamma}{2}\log\Big(\frac{1}{1 - \beta(1 - \bar{q})}\Big) + \frac{(\alpha + \gamma)\beta}{2}\frac{\bar{q}}{1 - \beta(1 - \bar{q})} \\
&- \frac{\beta}{2}(\alpha\bar{p} + \gamma\bar{r})(1 - \bar{q}) - \frac{(\alpha + \gamma)\beta}{2} - \frac{\beta}{2}M^2.
\end{aligned} \tag{29}$$

Now we have to extremize the free energy (29) with respect to the replica symmetric order parameters $\bar{q}, \bar{p}, \bar{r}, M$, namely we impose that

$$\partial_{\bar{q}} A^{RS}(\beta; \alpha, \gamma) = 0, \qquad \partial_{\bar{p}} A^{RS}(\beta; \alpha, \gamma) = 0, \qquad \partial_{\bar{r}} A^{RS}(\beta; \alpha, \gamma) = 0, \qquad \partial_M A^{RS}(\beta; \alpha, \gamma) = 0.$$

This gives the following system of integrodifferential equations to be simultaneously satisfied

$$\partial_{\bar{q}} A^{RS} = \frac{\beta}{2}\Big(\alpha\bar{p} + \gamma\bar{r} - \frac{(\alpha + \gamma)\bar{q}\beta}{(1 - \beta(1 - \bar{q}))^2}\Big) = 0, \tag{30}$$

$$\partial_{\bar{p}} A^{RS} = \frac{\alpha\beta}{2}\Big(\bar{q} - \int d\mu(\eta)\tanh^2\Big(\eta\sqrt{\beta(\alpha\bar{p} + \gamma\bar{r}) + \beta^2 M^2}\Big) = 0, \tag{31}$$

$$\partial_{\bar{r}} A^{RS} = \frac{\gamma\beta}{2}\Big(\bar{q} - \int d\mu(\eta)\tanh^2\Big(\eta\sqrt{\beta(\alpha\bar{p} + \gamma\bar{r}) + \beta^2 M^2}\Big) = 0, \tag{32}$$

$$\partial_M A^{RS} = \int d\mu(\eta)\tanh\Big(\eta\sqrt{\beta(\alpha\bar{p} + \gamma\bar{r}) + \beta^2 M^2}\Big), \tag{33}$$

by which we can generalize the well known AGS equation for the overlap as

$$\alpha\bar{p} + \gamma\bar{r} = \bar{q}(\alpha + \gamma)\beta/(1 - \beta(1 - \bar{q}))^2, \tag{34}$$

$$\bar{q} = \int d\mu(\eta)\tanh^2\Big(\beta[\frac{\sqrt{(\alpha + \gamma)\bar{q}}\eta}{1 - \beta(1 - \bar{q})} + M]\Big), \tag{35}$$

and write down the critical line as

$$\beta_c = \frac{1}{1 + \sqrt{\alpha + \gamma}}. \tag{36}$$

[3]It can be alternatively though of as a functional generator [11], given the exponential structure of the Maxwell-Boltzmann weight.

This result is not surprising because, as far as the z, τ variables of the hidden layer are uncorrelated, the Hebbian synaptic matrix is a linear sum of their contribution (see eq. 9), and, as a consequence, equivalent to a single Hopfield model which level of storage $P + K$.
However a far from trivial consequence can be inferred for the corresponding Boltzmann machine counterpart as the phase diagram of the Hopfield model is well known (see for instance [2][11]): In particular, at low level of noise, the relative sizes of the analog layers with respect to the digital one should not exceed the 14%, otherwise the equivalent Hopfield model is pushed in a spin glass phase where learning and retrieval is no longer possible. This offers a concrete prescription for the explicit construction of a Boltzmann Machines of the specific type considered here, and it suggests a possible way to study this problem in more general cases.

3.5 Bounds on interacting features

As far as the two hidden layers of the HBM do not interact, the corresponding Hebbian synaptic matrix works at its optimal capacity, appearing as an unweighed linear sum of the patterns (see eq. 10). When the hidden units in the two separate sets are made to interact, performances on the cognitive capabilities may change: We use a mean field approximation to show that, at least for small interaction strength (where calculations are handily), they act reciprocally as further noise sources for the retrieval of the σ.
To this task let us introduce the fully interacting cost function of the HBM as H_I, where I stands for "interacting features":

$$H_I(\sigma, z, \tau; \xi, \eta) = \frac{-1}{\sqrt{N}} \sum_{i\mu} \xi_i^\mu \sigma_i z_\mu + \frac{-1}{\sqrt{N}} \sum_{ik} \xi_i^k \sigma_i \tau_k + \frac{-\epsilon}{\sqrt{N}} \sum_{\mu k} \xi_{\mu k} z_\mu \tau_k, \tag{37}$$

where the last term accounts for interacting features, whose strength is ruled by ϵ which is assumed to be small.
The complete control of the model in this case is far from trivial as the investigation of spin glasses with gaussian spins is still under debate [5], however, within some assumptions (namely retaining only the leading order in ϵ), bounds can be obtained.
Mirroring Section 3.1, we integrate the partition function Z_I coupled to the cost function (37) over the τ variables, to get

$$Z_I = \sum_\sigma \exp\Big(\frac{\beta}{2N} \sum_{ij}^{N} (\sum_k^{\alpha N} \xi_i^k \xi_j^k) \sigma_i \sigma_j\Big) \int d\mu(z_\mu) \exp\Big(\sum_\mu^{\gamma N} z_\mu \Phi_\mu + \epsilon^2 \sum_{\mu\nu}^{\gamma N} z_\mu \Psi_{\mu\nu} z_\nu\Big), \tag{38}$$

with

$$\Phi_\mu = \frac{\sqrt{\beta}}{\sqrt{N}} \sum_i \xi_i^\mu \sigma_i + \epsilon \frac{\beta}{N} \sum_i \sigma_i (\sum_k \xi_i^k \xi_\mu^k), \tag{39}$$

$$\Psi_{\mu\nu} = \frac{\beta}{2N} (\sum_k \xi_\mu^k \xi_\nu^k). \tag{40}$$

It is straightforward to obtain the corresponding counterpart when integrating over the z variables first.
Now, within a pure mean field approach, $\sum_\nu \Psi_{\mu\nu} z_\nu \sim -(\alpha\beta^2/2) z_\mu$, and we can bound the

expression above with a partition function as

$$Z_I \sim \sum_{\sigma} \exp\Big(\frac{\beta}{2N}\sum_{ij}^{N}(\sum_{k}^{\alpha N}\xi_i^k\xi_j^k + \sum_{\mu}^{\gamma N}\xi_i^\mu\xi_j^\mu\frac{1}{\sqrt{1+\epsilon\alpha\beta^2}})\sigma_i\sigma_j\Big). \tag{41}$$

Performing the other integration first we get to the equivalent form of

$$Z_I = \sum_{\sigma} \exp\Big(\frac{\beta}{2N}\sum_{ij}^{N}(\sum_{\mu}^{\gamma N}\xi_i^\mu\xi_j^\mu + \frac{1}{\sqrt{1+\epsilon\gamma\beta^2}}\sum_{k}^{\alpha N}\xi_i^k\xi_j^k)\sigma_i\sigma_j\Big), \tag{42}$$

by which, again retaining only the leading ϵ-terms, the overall equivalent Hamiltonian in the Hopfield counterpart reads off as

$$H(\sigma;\xi,\eta) = \frac{\beta}{4N}\sum_{ij}^{N}\Big(\sum_{\mu}^{\alpha N}\xi_i^\mu\xi_j^\mu[1+\frac{1}{\sqrt{1+\epsilon\beta^2\gamma}}] + \sum_{k}^{\gamma N}\xi_i^k\xi_j^k[1+\frac{1}{\sqrt{1+\epsilon\beta^2\alpha}}]\Big). \tag{43}$$

As we can see, at least for small interaction coupling among the hidden units, the effect on their retrieval (performed by the external unit), is an extra-noise which affects the amplitude of the signal afferent to the output layer: In particular, the larger the size of the unit which is perturbing the retrieval on the other one, the smaller is the signal available for the retrieval itself.

4 Conclusions

First, by demonstrating the exact mapping between the Hopfield network and the HBM, we pave the way to a novel procedure for simulating large Hopfield networks: In particular, Hopfield networks require updating N neurons and storing $N(N-1)/2$ synapses, while HBM require updating $N+P$ neurons and storing only NP synapses, where P is the number of stored patterns. In addition, the well known spin glass transition of the Hopfield model has a counterpart in the HBM. In Boltzmann Machines, the ratio among the sizes of the hidden and visible layers is arbitrary and needs to be optimized in order to obtain the best possible generative model of the observed data. If the number of hidden units is too small, then the generative model is overconstrained and is not able to learn, while if it is too big then the model "overlearns" (overfits) the observed data and is not able to generalize [9]. Interestingly, these two extrema correspond in the Hopfield model to, respectively, the low storage phase, in which only a few patterns can be represented, and the spin glass phase, in which there is an exponentially increasing number of stable states. Therefore, the corresponding phase transition in the HBM can be understood as the optimal trade-off between flexibility and generality, thus effectively representing a statistical regularization procedure [8].

Furthermore we showed that, as far as we sum stimuli of independent nature, they linearly fill the memory of the equivalent Hopfield network, namely, probabilistic independence among the features is mapped into linearity into the cost functions, while, if input features are made to interact, within a first approximation, we show that they affect their reciprocal retrieval acting as noise sources.

Although the replica trick has represented a breakthrough for studying the thermodynamics of the Hopfield model, we argue that the "natural" mathematical backbone required for solving

the thermodynamic of the Boltzmann machine is the (suitably extendend) stochastic stability, whose implementation on bipartite spin-glasses is much more handily.
This further contributes on connecting scientific communities quite far apart, such as the mathematical physicists of spin glasses (see i.e. [10]) and the computer scientists of artificial intelligence (see i.e. [21]).

Acknowledgments

The strategy outlined in this research article belongs to the study supported by the Italian Ministry for Education and Research (FIRB grant number $RBFR08EKEV$) and by Sapienza Universita' di Roma.
AB is partially funded by GNFM (Gruppo Nazionale per la Fisica Matematica) which is also acknowledged
AB is grateful to Elena Agliari and Francesco Guerra for useful discussions.

References

[1] M. Aizenman, P. Contucci, *On the stability of the quenched state in mean field spin glass models*, J. Stat. Phys. **92**, 765-783 (1998).

[2] D.J. Amit, *Modeling brain function: The world of attractor neural network*, Cambridge Univerisity Press, (1992).

[3] A. Barra, *The mean field Ising model trough interpolating techniques*, J. Stat. Phys. **132**, 12-32, (2008).

[4] A. Barra, G. Genovese, F. Guerra, *The replica symmetric behavior of the analogical neural network*, J. Stat. Phys. **140**, 784-796, (2010).

[5] A. Barra, G. Genovese, F. Guerra, D. Tantari, *The fully solvable Gaussian spin glass*, to appear.

[6] A. Bernacchia, H. Seo, D. Lee, X.-J. Wang, *A reservoir of time constants for memory traces in cortical neurons*, Nature Neuroscience, **14**, 366-372, (2011).

[7] A. Bernacchia, D.J. Amit, *Impact of spatiotemporally correlated images on the structure of memory*, Proc. Natl. Acad. Sci. USA, **104**, 3544-3549, (2007).

[8] A. Bernacchia, S. Pigolotti, *Self-consistent method for density estimation*, J. Roy. Stat. Soc. B Met., **73**, 407-422, (2011).

[9] Y. Bengio, *Learning Deep Architectures for Artificial Intelligence*, Machine Learning **2**, 1, 127, (2009).

[10] A. Bovier, P. Picco, *Mathematical Aspects of Spin Glasses and Neural Networks*, Birkheauser Editor, (1998) and references therein.

[11] A.C.C. Coolen, R. Kuehn, P. Sollich, *Theory of Neural Information Processing Systems*, Oxford University Press, (2005).

[12] I. Gallo, P. Contucci, *Bipartite mean field spin systems. Existence and solution* , Math. Phys. E. J. **14**, 463, (2008).

[13] F. Guerra, F. L. Toninelli, *The Thermodynamic Limit in Mean Field Spin Glass Models*, Comm. Math. Phys. **230**, 71-79, (2002).

[14] D.O. Hebb, *Organization of Behaviour*, Wiley, New York, (1949).

[15] V. Honavar, L. Uhr, *Artificial Intelligence and Neural Networks: Steps Toward Principled Integration*, Elsevier, Boston, Academic Press, (1994).

[16] J. Hertz, A. Krogh, R. Palmer, *Introduction to the theory of neural computation*, Santa Fe Institute Studies in the Sciences of Complexity, (1991).

[17] G.E. Hinton, *Learning multiple layers of representation*, Trends in Cognitive Science **11**, 10, 428-434, (2007).

[18] G.E. Hinton, R. R. Salakhutdinov, *Reducing the Dimensionality of Data with Neural Networks*, Science **313**, 5786, 504-507, (2006).

[19] J.J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, P.N.A.S. **79**, 2554-2558, (1982).

[20] T. K. Marks, J. R. Movellan, *Diffusion Networks, Products of Experts, and Factor Analysis*, Proc. Third Int. Conf. Independent Component Anal. Signal Separation, (2001).

[21] M. Mezard, A. Montanari, *Information,Physics and Computation*, Cambridge Press, (2007).

[22] M. Mezard, G. Parisi, M.A. Virasoro, *Spin glass theory and beyond*, World Scientific, Singapore, (1987).

[23] X. Pitkow, *Exact feature probabilities in images with occlusion*, J. of Vision **10**, 14,

[24] L. Pastur, M. Scherbina, B. Tirozzi, *The replica symmetric solution of the Hopfield model without replica trick*, J. Stat. Phys. **74**, 1161-1183, (1994).

[25] H.C. Tuckwell, *Introduction to theoretical neurobiology*, Vol.2, Cambridge University Press (1988).