Noname manuscript No.
(will be inserted by the editor)

J. P. Neirotti

Learning in ultrametric committee machines

the date of receipt and acceptance should be inserted later

Abstract The problem of learning by examples in ultrametric committee machines (UCMs) is studied within the framework of statistical mechanics. Using the replica formalism we calculate the average generalization error in UCMs with L hidden layers and for a large enough number of units. In most of the regimes studied we find that the generalization error, as a function of the number of examples presented, develops a discontinuous drop at a critical value of the load parameter. We also find that when $L > 1$ a number of teacher networks with the same number of hidden layers and different overlaps induce learning processes with the same critical points.

Keywords Multilayer-Networks, Learning-by-Examples

1 Introduction

From a theoretical perspective, neural networks are the archetype of disordered systems whose study has motivated researchers to develop new statistical mechanics techniques. Being idealizations used to model some aspects of the brain's behavior, they represent a new approach to the problem of computation, based on a paralleled processing of information. As a consequence, neural networks research is multidisciplinary; models inspired in biologic observations have been used to better understand emergent phenomena [1], pattern recognition and task reproduction [2], associative memory capacity [3] and neural developing [4]; several aspect of the learning process have been investigated by using recurrent [5] and spiking networks [6]; applications using neural networks have been recently developed for credit assignment [7] and Bayesian inference [8]. These research has also played a complementary role to studies *in vivo* [9–11]. The work described here is motivated by the need to better understand the learning by examples process in artificial networks to help the development of more efficient automatic systems.

One of the most well-studied and better understood feed-forward networks is the perceptron [12–14] which, because of its simplicity, has very limited computational capabilities. Ultrametric Committee Machines (UCMs), as presented in [15], represent one step forward in architectural complexity and, as a consequence, they are potentially better suited for real world applications. UCMs are fully connected committee machines with K units in the first of its L hidden layers, a tree like structure linking hidden-to-hidden layers and non-zero synaptic overlaps at the hidden-to-input level only. These overlaps, i.e. the inner products between synaptic vectors belonging to different units, form an ordered set $\left\{\tilde{\zeta}_j,\, j = 1, 2, \ldots, L\right\}$, where $\tilde{\zeta}_j \gg \tilde{\zeta}_i$ for all $j > i$. The overlap's subindex indicates the number of layers we have to go forward to find the common root to both units thus indicating the ultrametric distance between them (see figure 1). Although more sophisticated than the perceptron, some of the UCM's computational properties can be analytically obtained. Indeed, we show in this article that

NCRG, Aston University, Birmingham B4 7ET, UK.
E-mail: j.p.neirotti@aston.ac.uk

by the application of the replica trick, it is possible to study the learning by examples process when, considering students and teachers of the same architecture, the number of examples presented to the student is proportional to the number of units in one of the L hidden layers of the teacher. This scaling emerges naturally from the expression of the free energy of the system (see below).

Most of the studies found in the literature, in the area of learning process in networks, consider only the case where the teacher's synaptic overlaps are set to zero. Recently, [15], we found that there is a clear relationship between the magnitude of these overlaps and how difficult is to reproduce the teacher's classification. In this respect our results give support to the definition of network complexity presented in [17]. Understanding the link between task difficulty and network complexity is fundamental for the development of tools for practical applications.

A more complete description of UCM's is presented as follows. UCMs are fully connected committee machines with L hidden layers organized in such a way that the number of units in the L-th layer (last hidden) is K_L, each one of them linked to K_{L-1} units in the $(L-1)$-th layer through unit weights. This tree-like structure is repeated until reaching the first hidden layer, thus the total number of units populating the ℓ-th hidden layer is $K_L K_{L-1} \dots K_\ell$. The $K_L \dots K_1 = K$ units in the first hidden layer are connected to the inputs through synaptic vectors $\mathbf{w}_{\mathbf{k}_1} \in \mathbb{R}^N$ whose overlap matrix $[\boldsymbol{\Omega}]_{\mathbf{k}_1,\mathbf{k}'_1} \equiv \mathbf{w}^{\mathsf{T}}_{\mathbf{k}_1}\mathbf{w}_{\mathbf{k}'_1}/N$ satisfies the following relationship:

$$[\boldsymbol{\Omega}]_{\mathbf{k}_1,\mathbf{k}'_1} = \delta_{\mathbf{k}_1,\mathbf{k}'_1}\left(1-\tilde{\zeta}_1\right) + \dots + \delta_{k_L,k'_L}\left(\tilde{\zeta}_{L-1}-\tilde{\zeta}_L\right) + \tilde{\zeta}_L, \tag{1}$$

where $\mathbf{w}^{\mathsf{T}}$ is the transpose of the vector $\mathbf{w}$, the indexes $\mathbf{k}_\ell \equiv [k_L, k_{L-1}, \dots, k_\ell]$ run over all hidden units of the ℓ-th layer, $\delta_{\mathbf{k}_\ell \mathbf{k}'_\ell} \equiv \prod_{m=\ell}^{L} \delta_{k_m k'_m}$ and $\delta_{ij} = 1$ if and only if $i = j$ and 0 otherwise and the overlaps $\tilde{\zeta}_j$ satisfy the relationship:

$$\tilde{\zeta}_\ell = \frac{\zeta_\ell}{\prod_{j=1}^{\ell} K_j}, \tag{2}$$

where the ζ_j are independent on the size of the system. The matrix of overlaps $\boldsymbol{\Omega}$ is ultrametric [18], hence UCMs. All the units that compose the network are binary, and process their inputs according to the following rules: the output unit $\sigma_{\mathbb{W}}(\mathbf{S}) \equiv \operatorname{sgn}\left(\sum_{k_L=1}^{K_L} \sigma_{k_L}(\mathbf{S})/\sqrt{K_L}\right)$, the ℓ-th hidden layer unit $\sigma_{\mathbf{k}_\ell}(\mathbf{S}) \equiv \operatorname{sgn}\left(\sum_{k_{\ell-1}=1}^{K_{\ell-1}} \sigma_{[\mathbf{k}_{\ell-1}]}(\mathbf{S})/\sqrt{K_{\ell-1}}\right)$ and the first hidden layer unit $\sigma_{\mathbf{k}_1}(\mathbf{S}) \equiv \operatorname{sgn}\left(\mathbf{w}^{\mathsf{T}}_{\mathbf{k}_1}\mathbf{S}/\sqrt{N}\right)$, where $\mathbb{W} = \left\{\mathbf{w}_{\mathbf{k}_1} \in \mathbb{R}^N \text{ and } \mathbf{w}^{\mathsf{T}}_{\mathbf{k}_1}\mathbf{w}_{\mathbf{k}_1} = N\right\}$ is the set of synaptic vectors associated to the first hidden layer's units and $\mathbf{S} \in \{1,-1\}^N$. In figure 1 we present an UCM with $L = 3$ hidden layers and K_3, K_2K_3 and $K_1K_2K_3 = K$ units in each layer.

We present as follows an investigation on the learning by examples process in UCMs, which is based upon [19–21] and generalizes the results found in [22]. In section 2 we present the problem in the language of statistical mechanics, in section 3 we present our results for the general case and a study on the particular cases where $L = 1$ and 2. Section 4 synthesizes our conclusions and final considerations.

2 Replica symmetric analysis

Given a set of examples $\mathcal{S}_P = \left\{\left(\boldsymbol{\xi}_\mu, \sigma_{\mathbb{W}^0}(\boldsymbol{\xi}_\mu)\right)\right\}_{\mu=1}^{P}$, where the patterns $\boldsymbol{\xi}_\mu \in \{1,-1\}^N$ have been classified by an UCM teacher $\mathbb{W}^0$ with labels $\sigma_{\mathbb{W}^0}(\boldsymbol{\xi}_\mu) \in \{1,-1\}$, we can define the Hamiltonian of the student $\mathbb{W}$ by $H_N(\mathbb{W};\mathcal{S}_P) \equiv \sum_{\mu=1}^{P} \Theta(-\sigma_{\mathbb{W}^0}(\boldsymbol{\xi}_\mu)\,\sigma_{\mathbb{W}}(\boldsymbol{\xi}_\mu))$, where $\sigma_{\mathbb{W}}(\boldsymbol{\xi}_\mu) \in \{1,-1\}$ is the classification given by the student to the μ-th pattern and $\Theta(x > 0) = 1$ and 0 otherwise. We are interested on computing the system's statistical properties at zero temperature in the thermodynamic limit $(N \uparrow \infty)$, and in the large network regime $(1 \ll K < \infty)$, through the calculation of the partition function:

$$Z_N(\mathcal{S}_P) = \int \mathrm{d}\rho(\mathbb{W}) \prod_{\mu=1}^{P} \Theta(\sigma_{\mathbb{W}^0}(\boldsymbol{\xi}_\mu)\,\sigma_{\mathbb{W}}(\boldsymbol{\xi}_\mu)), \tag{3}$$

where $\mathrm{d}\rho(\mathbb{W})$ is a measure of the synaptic vectors compatible with the UCM character of the student. $Z_N(\mathcal{S}_P)$ can be interpreted as the fraction of vectors in $\mathbb{R}^{NK}$ satisfying the constraints imposed by the

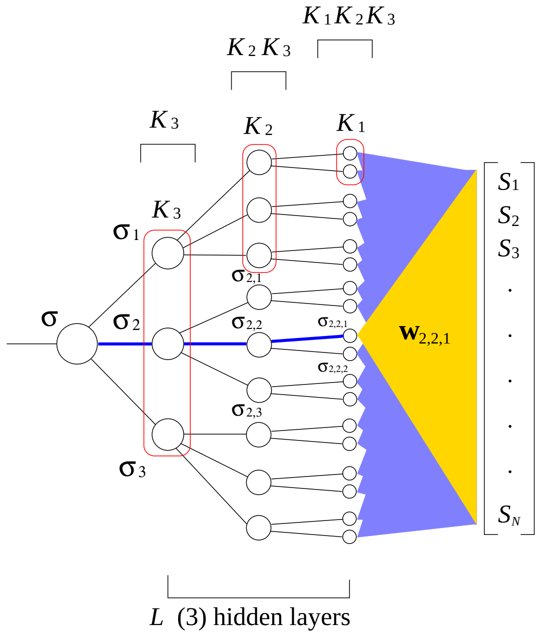

Fig. 1 Typical feed-forward network architecture studied in this article. This committee has $L = 3$ hidden layers with synaptic overlaps in the hidden-to-input layer. All the synaptic weights linking hidden-to-hidden and hidden-to-output units are set to one. Observe that the highlighted synaptic vector $\mathbf{w}_{2,2,1}$ corresponds to the unit with index $\mathbf{k}_1 = [2, 2, 1]$, i.e. the unit linked to the output through the path 2, 2, 1. The network is composed by K_3, K_2K_3 and $K_1K_2K_3$ units in each layer and the boxes illustrate the meaning of the numbers K_1, K_2 and K_3.

Θ function. The zero temperature free energy of the system is defined as $f_N(\mathcal{S}_P) \equiv -\frac{1}{NK} \ln Z_N(\mathcal{S}_P)$, which is of order one and self averaging in the $N \uparrow \infty$ limit. Thus, we have that the statistical properties of the system are conveyed by the quenched average of the free energy, that can be computed by the replica trick:

$$f = -\lim_{N\to\infty} \frac{1}{NK} \lim_{n\to 0} \frac{\left\langle \langle Z_N^n(\mathcal{S}_P) \rangle_{\mathbb{W}^0} \right\rangle_{\boldsymbol{\xi}} - 1}{n}. \tag{4}$$

Assuming that the number of patterns P presented to the student scales with the size of the synaptic vectors as $P = \alpha N$, where α is the load parameter, the quenched average of the replicated system can be expressed as:

$$\left\langle \langle Z_N^n(\mathcal{S}_P) \rangle_{\mathbb{W}^0} \right\rangle_{\boldsymbol{\xi}} = \int \prod_{a=0}^{n} \mathrm{d}\rho(\mathbb{W}^a) \, \exp\left(\alpha N G_{E,N}^{(n)}(\{\mathbb{W}^a\})\right),$$

where the pattern distribution $\mathcal{P}(\boldsymbol{\xi}_\mu) = \prod_{j=1}^{N} \mathcal{P}(\xi_{j,\mu}) = 2^{-N}$ is independent of μ and

$$G_{E,N}^{(n)}(\{\mathbb{W}^a\}) \equiv -\ln \left\langle \prod_{a=1}^{n} \Theta(\sigma_{\mathbb{W}^0}(\boldsymbol{\xi})\, \sigma_{\mathbb{W}^a}(\boldsymbol{\xi})) \right\rangle_{\boldsymbol{\xi}}.$$

By means of delta functions we can introduce the parameters $N q^{a,b}_{\mathbf{k}_1,\mathbf{m}_1} \equiv \mathbf{w}^{a\mathsf{T}}_{\mathbf{k}_1} \mathbf{w}^{b}_{\mathbf{m}_1}$ for all $0 < a < b \leq n$, $N t^{a}_{\mathbf{k}_1,\mathbf{m}_1} \equiv \mathbf{w}^{a\mathsf{T}}_{\mathbf{k}_1} \mathbf{w}^{a}_{\mathbf{m}_1}$ for all $\mathbf{k}_1 \neq \mathbf{m}_1$ and $N r^{a}_{\mathbf{k}_1,\mathbf{m}_1} \equiv \mathbf{w}^{a\mathsf{T}}_{\mathbf{k}_1} \mathbf{w}^{0}_{\mathbf{m}_1}$, and considering that the student and teacher committees satisfy $N = \mathbf{w}^{a\mathsf{T}}_{\mathbf{k}_1} \mathbf{w}^{a}_{\mathbf{k}_1}$ for all $0 < a \leq n$ and $N[\boldsymbol{\Omega}]_{\mathbf{k}_1,\mathbf{m}_1} = \mathbf{w}^{0\mathsf{T}}_{\mathbf{k}_1} \mathbf{w}^{0}_{\mathbf{m}_1}$, where $\boldsymbol{\Omega}$ is as defined in (1), we can express the zero temperature, replicated partition function as:

$$\left\langle \langle Z_N^n(\mathcal{S}_P) \rangle_{\mathbb{W}^0} \right\rangle_{\boldsymbol{\xi}} = \int \prod_{\mathbf{k}_1 \mathbf{m}_1} \left(\prod_{a<b} \frac{\mathrm{d}q^{a,b}_{\mathbf{k}_1,\mathbf{m}_1} \, \mathrm{d}\hat{q}^{a,b}_{\mathbf{k}_1,\mathbf{m}_1}}{2\pi/N} \prod_{a} \frac{\mathrm{d}r^{a}_{\mathbf{k}_1,\mathbf{m}_1} \, \mathrm{d}\hat{r}^{a}_{\mathbf{k}_1,\mathbf{m}_1}}{2\pi/N} \right)$$

$$\int \prod_{a} \left(\prod_{\mathbf{k}_1 < \mathbf{m}_1} \frac{\mathrm{d}t^{a}_{\mathbf{k}_1,\mathbf{m}_1} \, \mathrm{d}\hat{t}^{a}_{\mathbf{k}_1,\mathbf{m}_1}}{2\pi/N} \prod_{\mathbf{k}_1} \frac{\mathrm{d}\hat{\kappa}^{a}_{\mathbf{k}_1}}{4\pi} \right) \exp\left[-\alpha N G_{E,N}^{(n)}(\mathscr{P}) - \frac{1}{2} N G_{S,N}^{(n)}(\mathscr{P}; \hat{\mathscr{P}}) \right],$$

where $\mathscr{P} \equiv \{t^a_{\mathbf{k}_1,\mathbf{m}_1}\} \cup \{r^a_{\mathbf{k}_1,\mathbf{m}_1}\} \cup \{q^{a,b}_{\mathbf{k}_1,\mathbf{m}_1}\}$ and $\hat{\mathscr{P}} \equiv \{\hat{\kappa}^a_{\mathbf{k}_1}\} \cup \{\hat{r}^a_{\mathbf{k}_1,\mathbf{m}_1}\} \cup \{\hat{t}^a_{\mathbf{k}_1,\mathbf{m}_1}\} \cup \{\hat{q}^{a,b}_{\mathbf{k}_1,\mathbf{m}_1}\}$ and:

$$
\begin{aligned}
NG^{(n)}_{S,N}(\mathscr{P};\hat{\mathscr{P}}) \equiv -2\ln\Bigg\langle \exp\Bigg[& \sum_{(a,b)}\sum_{(\mathbf{k}_1,\mathbf{m}_1)} \hat{q}^{a,b}_{\mathbf{k}_1,\mathbf{m}_1}\left(Nq^{a,b}_{\mathbf{k}_1,\mathbf{m}_1} - \mathbf{w}^{a\mathsf{T}}_{\mathbf{k}_1}\mathbf{w}^b_{\mathbf{m}_1}\right) + \\
& + \sum_a \sum_{(\mathbf{k}_1,\mathbf{m}_1)} \hat{t}^a_{\mathbf{k}_1,\mathbf{m}_1}\left(Nt^a_{\mathbf{k}_1,\mathbf{m}_1} - \mathbf{w}^{a\mathsf{T}}_{\mathbf{k}_1}\mathbf{w}^a_{\mathbf{m}_1}\right) + \\
& + \sum_a \sum_{\mathbf{k}_1}\left(\sum_{\mathbf{m}_1}\hat{r}^a_{\mathbf{k}_1,\mathbf{m}_1}\left(Nr^a_{\mathbf{k}_1,\mathbf{m}_1} - \mathbf{w}^{0\mathsf{T}}_{\mathbf{k}_1}\mathbf{w}^a_{\mathbf{m}_1}\right) + \frac{1}{2}\hat{\kappa}^a_{\mathbf{k}_1}\left(N - \mathbf{w}^{a\mathsf{T}}_{\mathbf{k}_1}\mathbf{w}^a_{\mathbf{k}_1}\right)\right)\Bigg]\Bigg\rangle_{\mathbb{W}^{n+1}}.
\end{aligned}
$$

The solution of the saddle point equation on the variables in $\hat{\mathscr{P}}$ can be expressed in terms of the variables in $\mathscr{P}$ and, by considering the replica symmetric ansatz and the UCM character of the student, we can express these parameters as:

$$
\begin{aligned}
t^a_{\mathbf{k}_1,\mathbf{m}_1} &\equiv \delta_{\mathbf{k}_1\mathbf{m}_1}\left(1-\tilde{t}_1\right) + \delta_{\mathbf{k}_2\mathbf{m}_2}\left(\tilde{t}_1-\tilde{t}_2\right) + \ldots + \delta_{k_L m_L}\left(\tilde{t}_{L-1}-\tilde{t}_L\right) + \tilde{t}_L \\
q^{a,b}_{\mathbf{k}_1,\mathbf{m}_1} &\equiv \delta_{\mathbf{k}_1\mathbf{m}_1}\left(1-\tilde{q}_1\right) + \delta_{\mathbf{k}_2\mathbf{m}_2}\left(\tilde{q}_1-\tilde{q}_2\right) + \ldots + \delta_{k_L m_L}\left(\tilde{q}_{L-1}-\tilde{q}_L\right) + \tilde{q}_L \\
r^a_{\mathbf{k}_1,\mathbf{m}_1} &= \delta_{\mathbf{k}_1\mathbf{m}_1}\left(1-\tilde{r}_1\right) + \delta_{\mathbf{k}_2\mathbf{m}_2}\left(\tilde{r}_1-\tilde{r}_2\right) + \ldots + \delta_{k_L m_L}\left(\tilde{r}_{L-1}-\tilde{r}_L\right) + \tilde{r}_L,
\end{aligned}
$$

where $\tilde{q}_j$, $\tilde{r}_j$ and $\tilde{t}_j$ obey the scaling law (2). We find that:

$$
G^{(n)}_{E,N}(\mathscr{P}) \simeq -2n\int \mathcal{D}z\,\mathcal{H}\left(\sqrt{\frac{\mathscr{R}_L}{1-\mathscr{R}_L}}\,z\right)\ln\mathcal{H}\left(\sqrt{\frac{\mathscr{Q}_L}{1-\mathscr{Q}_L}}\,z\right) + O(n^2, N^{-1}),
$$

where:

$$
\begin{aligned}
\mathscr{Q}_\ell &\equiv \frac{\gamma_{\ell-1}}{\gamma_\ell}\left[\frac{2}{\pi}\arcsin\left(\mathscr{Q}_{\ell-1}\right) + \left(\frac{2}{\pi}\right)^\ell \frac{q_\ell}{\gamma_{\ell-1}}\right] \\
\mathscr{R}_\ell &\equiv \frac{\rho_{\ell-1}}{\rho_\ell}\left[\frac{2}{\pi}\arcsin\left(\sqrt{\frac{\gamma_{\ell-2}}{\gamma_{\ell-1}}\frac{\rho_{\ell-1}}{\rho_{\ell-2}}}\mathscr{R}_{\ell-1}\right) + \left(\frac{2}{\pi}\right)^\ell \frac{r_\ell}{\sqrt{\gamma_{\ell-1}\rho_{\ell-1}}}\right],
\end{aligned}
$$

with $\mathscr{Q}_0 \equiv q_0$, $\mathscr{R}_0 \equiv r_0$, $\gamma_\ell \equiv 1 + \sum_{j=1}^{\ell}\left(\frac{2}{\pi}\right)^j t_j$ and $\rho_\ell \equiv 1 + \sum_{j=1}^{\ell}\left(\frac{2}{\pi}\right)^j \zeta_j$, for all $\ell = 0,\ldots,L$. We have also use the definitions $\mathcal{D}z \equiv \mathrm{d}z\,\mathrm{e}^{-z^2/2}/\sqrt{2\pi}$ and $\mathcal{H}(z) \equiv \int_z^\infty \mathcal{D}x$.

Assuming that the number of units per layer in the committees respect the relations $K_1 \ldots K_{\ell-1}K_\ell^2 > K$ for all $\ell = 1,\ldots,L$, which simply indicates that the closer to the input the more densely populated the layer, the dominant contribution to the configurational term can be expressed as:

$$
G^{(n)}_{S,N}(\mathscr{P}) \simeq nK\left[g_0(\mathscr{P}) + \frac{1}{K_1}g_1(\mathscr{P}) + \frac{1}{K_1K_2}g_2(\mathscr{P}) + \ldots + \frac{1}{K}g_L(\mathscr{P})\right],
$$

where

$$
g_0(\mathscr{P}) \equiv \ln(1-q_0) + \frac{q_0 - r_0^2}{1-q_0}
$$

$$
\begin{aligned}
g_\ell(\mathscr{P}) \equiv{}& \ln\left(\frac{1-q_0+\sum_{j=1}^{\ell}(t_j-q_j)}{1-q_0+\sum_{j=1}^{\ell-1}(t_j-q_j)}\right) + \frac{\sum_{j=0}^{\ell} q_j - \dfrac{\left(\sum_{j=0}^{\ell} r_j\right)^2}{1+\sum_{j=1}^{\ell}\zeta_j}}{1-q_0+\sum_{j=1}^{\ell}(t_j-q_j)} - \frac{\sum_{j=0}^{\ell-1} q_j - \dfrac{\left(\sum_{j=0}^{\ell-1} r_j\right)^2}{1+\sum_{j=1}^{\ell-1}\zeta_j}}{1-q_0+\sum_{j=1}^{\ell-1}(t_j-q_j)} + \\
&+ \left[\frac{\sum_{j=0}^{\ell-1} q_j - \dfrac{\left(\sum_{j=0}^{\ell-1} r_j\right)^2}{1+\sum_{j=1}^{\ell-1}\zeta_j}}{1-q_0+\sum_{j=1}^{\ell-1}(t_j-q_j)} - 1\right]\frac{t_\ell - q_\ell}{1-q_0+\sum_{j=1}^{\ell-1}(t_j-q_j)} - \frac{q_\ell + \dfrac{\zeta_\ell}{1+\sum_{j=1}^{\ell-1}\zeta_j}\dfrac{\left(\sum_{j=0}^{\ell-1} r_j\right)^2}{1+\sum_{j=1}^{\ell-1}\zeta_j} - 2r_\ell\dfrac{\sum_{j=0}^{\ell-1} r_j}{1+\sum_{j=1}^{\ell-1}\zeta_j}}{1-q_0+\sum_{j=1}^{\ell-1}(t_j-q_j)}.
\end{aligned}
$$

According to (4) and assuming that we can interchange the limits in N and n, we have that:

$$f(\alpha, \mathscr{P}) = -2\frac{\alpha}{K}\int \mathcal{D}z\,\mathcal{H}\left(\sqrt{\frac{\mathscr{R}_L}{1-\mathscr{R}_L}}\,z\right)\ln\mathcal{H}\left(\sqrt{\frac{\mathscr{Q}_L}{1-\mathscr{Q}_L}}\,z\right) - \\ -\frac{1}{2}\left[g_0(\mathscr{P}) + \frac{1}{K_1}g_1(\mathscr{P}) + \frac{1}{K_1K_2}g_2(\mathscr{P}) + \ldots + \frac{1}{K}g_L(\mathscr{P})\right], \tag{5}$$

is the free energy of the system.

The generalization error is defined by

$$\varepsilon_G(\mathscr{P}) \equiv \langle\!\langle \Theta\left(-\sigma_{\mathbb{W}^0}(\mathbf{S})\,\sigma_{\mathbb{W}}(\mathbf{S})\right)\rangle_{\mathbb{W}^0}\rangle_{\mathbf{S}} = \frac{1}{\pi}\arccos\left(\sqrt{\frac{\gamma_{L-1}}{\gamma_L}\frac{\rho_L}{\rho_{L-1}}}\mathscr{R}_L\right),$$

which is computed in a similar way as the average sensitivity [15] and should be evaluated in the parameters obtained from the optimization of (5).

3 Results

It is clear from the structure of (5) that there are different regimes corresponding to values of the load parameter α proportional to the number of units in a given layer, i.e. $\alpha \sim O(1)$, $O(K_L)$, $O(K_LK_{L-1})$, $\ldots, O(K)$ respectively. If $\alpha \sim O(1)$ the optimization of the first L leading terms in the free energy produces $q_\ell = r_\ell = 0$ for $0 \le \ell < L$ and $t_\ell = 0$ for $0 < \ell < L$. The optimal values of the remaining parameters are obtained by solving the set of equations:

$$0 = \frac{\partial}{\partial\eta}\left[-2\alpha\int \mathcal{D}z\,\mathcal{H}\left(\sqrt{\frac{\mathscr{R}_L}{1-\mathscr{R}_L}}\,z\right)\ln\mathcal{H}\left(\sqrt{\frac{\mathscr{Q}_L}{1-\mathscr{Q}_L}}\,z\right) - \frac{1}{2}g_L(q_L, r_L, t_L)\right], \tag{6}$$

where $\eta = q_L$, r_L and t_L. In the particular case where $\zeta_\ell = 0$ for all $0 < \ell < L$ we can propose $t_L = \zeta_L$ and $r_L = q_L$ leaving only one equation (in q_L) to be solved. This particularly symmetric case is characterized by the equation $\mathscr{R}_L = \mathscr{Q}_L$ and, although descriptively simpler, presents the same qualitative behavior as the case with all non-zero overlaps. By defining the integral

$$\mathcal{I}(x) \equiv \frac{1}{2\pi}\frac{1}{\sqrt{1-x^2}}\int \mathcal{D}z\,\mathcal{H}^{-1}\left(\sqrt{\frac{x}{1+x}}\,z\right), \tag{7}$$

the saddle point equation in q_L is:

$$0 = \alpha\left(\frac{2}{\pi}\right)^L\frac{\mathcal{I}(\mathscr{Q}_L)}{1+\left(\frac{2}{\pi}\right)^L\zeta_L} - \frac{1}{2}\frac{q_L}{(1+\zeta_L)(1+\zeta_L-q_L)}, \tag{8}$$

where $\mathscr{Q}_L = \left(\frac{2}{\pi}\right)^L\left[1+\left(\frac{2}{\pi}\right)^L\zeta_L\right]^{-1}q_L$. For large values of α the generalization error asymptotically approaches the value:

$$\varepsilon_G(\alpha) \simeq \frac{1}{\pi}\arccos(\mathscr{Q}_L^{(\infty)}) + \left[\int \mathcal{D}z\,\mathcal{H}^{-1}\left(\sqrt{\frac{\mathscr{Q}_L^{(\infty)}}{1+\mathscr{Q}_L^{(\infty)}}}\,z\right)\right]^{-1}\frac{1}{\alpha} + O(\alpha^{-2}), \tag{9}$$

where

$$\mathscr{Q}_L^{(\infty)} \equiv \left(\frac{2}{\pi}\right)^L\frac{1+\zeta_L}{1+\left(\frac{2}{\pi}\right)^L\zeta_L}, \tag{10}$$

which implies that total generalization can not be achieved for finite values of ζ_L.

Let us consider the case where the load parameter is proportional to the number of units in the ℓ-th hidden layer, i.e. $\alpha = \hat{\alpha}K_LK_{L-1}\ldots K_{\ell+1}$ with $0 < \ell < L$ and $\hat{\alpha}$ independent on $K_1, \ldots, K_L$. In this regime we found that the optimization of the first ℓ terms of the free energy is achieved by $q_j = r_j = 0$

for all $0 \le j < \ell$ and $t_j = 0$ for all $0 < j < \ell$. The optimization of the term of $O(K_1^{-1} \dots K_\ell^{-1})$ is achieved through

$$0 = \frac{\partial}{\partial \eta} \left[-2\hat{\alpha} \int \mathcal{D}z \, \mathcal{H}\left(\sqrt{\frac{\mathscr{R}_L}{1-\mathscr{R}_L}}\, z \right) \ln \mathcal{H}\left(\sqrt{\frac{\mathscr{Q}_L}{1-\mathscr{Q}_L}}\, z \right) - \frac{1}{2} g_\ell(\mathscr{P}) \right], \tag{11}$$

where $\eta = q_\ell$, r_ℓ, t_ℓ and $t_{\ell-1}$, in the cases where $\ell > 0$. The optimization of the remaining terms (order $K_1^{-1} \dots K_{\ell+1}^{-1}$ and higher) produces the relationships

$$0 = 1 + \sum_{j=\ell}^{m} (t_j - q_j) \quad \text{and} \quad 0 = \left(1 + \sum_{j=\ell}^{m} t_j \right) \left(1 + \sum_{j=1}^{m} \zeta_j \right) - \left(\sum_{j=\ell}^{m} r_j \right)^2,$$

for all $\ell < m \le L$. This system of equations gets a simpler form for a teacher with overlaps $\zeta_j = 0$ for all $0 < j < \ell$. In this case we can chose $t_j = \zeta_j$ for all $0 < j \le L$, $r_j = q_j$ for all $0 \le j \le L$ and $q_j = \zeta_j$ for all $\ell + 2 \le j \le L$ and the set of equations gets reduced to the relationship $q_{\ell+1} = 1 + \zeta_\ell + \zeta_{\ell+1} - q_\ell$ and the equation:

$$0 = \hat{\alpha} \left(\frac{2}{\pi} \right)^L \prod_{j=\ell+1}^{L-1} \frac{1}{\sqrt{1-\mathscr{Q}_j^2}} \left[\frac{1}{\sqrt{1-\mathscr{Q}_\ell^2}} - 1 \right] \frac{\mathcal{I}(\mathscr{Q}_L)}{1 + \sum_{j=\ell}^{L} \left(\frac{2}{\pi}\right)^j \zeta_j} - \frac{1}{2} \frac{q_\ell}{(1+\zeta_\ell)(1+\zeta_\ell - q_\ell)}. \tag{12}$$

The learning processes induced by these teachers are characterized by the equation $\mathscr{Q}_L = \mathscr{R}_L$. Although apparently simpler, these processes present a behavior qualitative similar to the processes induced by teachers with a full set of non-zero overlaps.

Observe that (12) always admits the solution $q_\ell = 0$, which is the global minimum of the free energy for small values of $\hat{\alpha}$. In this phase there is no specialization of the units and all the overlaps associated to the ℓ-th layer are zero. For values of the load parameter $\hat{\alpha} > \hat{\alpha}_s$ the free energy develops a second minimum at $q_\ell^\star > 0$. This minimum becomes global for $\hat{\alpha} > \hat{\alpha}_c > \hat{\alpha}_s$. For large values of $\hat{\alpha}$ the asymptotic expression for the generalization error matches (9) with:

$$\mathscr{Q}_\ell^{(\infty)} \equiv \left(\frac{2}{\pi} \right)^\ell \frac{1+\zeta_\ell}{1 + \left(\frac{2}{\pi}\right)^\ell \zeta_\ell} \tag{13}$$

$$\mathscr{Q}_m^{(\infty)} \equiv \frac{1 + \sum_{j=\ell}^{m-1} \left(\frac{2}{\pi}\right)^j \zeta_j}{1 + \sum_{j=\ell}^{m} \left(\frac{2}{\pi}\right)^j \zeta_j} \left[\frac{2}{\pi} \arcsin(\mathscr{Q}_{m-1}^{(\infty)}) + \left(\frac{2}{\pi} \right)^m \frac{\zeta_m}{1 + \sum_{j=\ell}^{m-1} \left(\frac{2}{\pi}\right)^j \zeta_j} \right], \tag{14}$$

where $\ell < m \le L$. In this way the generalization error asymptotically converges to a non-zero value if and only if $\ell > 0$ and $\zeta_j < \infty$ for all $j = 1, \dots, L$.

To illustrate our results let us consider a network with one hidden layer ($L = 1$). If $\alpha \sim O(1)$ we have that the learning process presents only a perceptron-like phase where its asymptotic generalization error is given by (9) and (10) and no transition is observed. If ζ_1 is zero we recover the result obtained in [22]. Large values of the overlap ζ_1 indicate that the hidden units of the teacher work almost like K identical perceptrons. Only in this case $\mathscr{Q}_1^{(\infty)}$ becomes close to one and the generalization error approaches zero asymptotically.

When $\alpha = \hat{\alpha} K$ the free energy has the form:

$$f(q; \hat{\alpha}) = -2\hat{\alpha} \int \mathcal{D}z \, \mathcal{H}\left(\sqrt{\frac{\mathscr{Q}_1}{1-\mathscr{Q}_1}}\, z \right) \ln \mathcal{H}\left(\sqrt{\frac{\mathscr{Q}_1}{1-\mathscr{Q}_1}}\, z \right) - \frac{1}{2} \left[\ln(1-q) + q \right],$$

where $\mathscr{Q}_1 = \frac{2}{\pi}(1 + \frac{2}{\pi}\zeta_1)^{-1}[\arcsin(q) + 1 + \zeta_1 - q]$, that is precisely the expression found in [22] when $\zeta_1 = 0$. The free energy has a minimum at $q = 0$ for all values of $\hat{\alpha}$ and develops a second minimum $0 < q^\star < 1$ at $\hat{\alpha} > \hat{\alpha}_s$. This minimum becomes global at $\hat{\alpha} > \hat{\alpha}_c > \hat{\alpha}_s$. $\hat{\alpha}_c$ can be obtained by solving the equations (12) simultaneously with $f_0(0, \hat{\alpha}_c) = f_0(q_c; \hat{\alpha}_c)$ where $q_c = q^\star(\hat{\alpha}_c)$. In the region where the minimum $q = 0$ is the dominant (small $\hat{\alpha}$), the generalization error is a constant. For $\hat{\alpha} > \hat{\alpha}_c$ the generalization error decays asymptotically to zero. The larger the parameter ζ_1 the larger $\hat{\alpha}_c$ and the

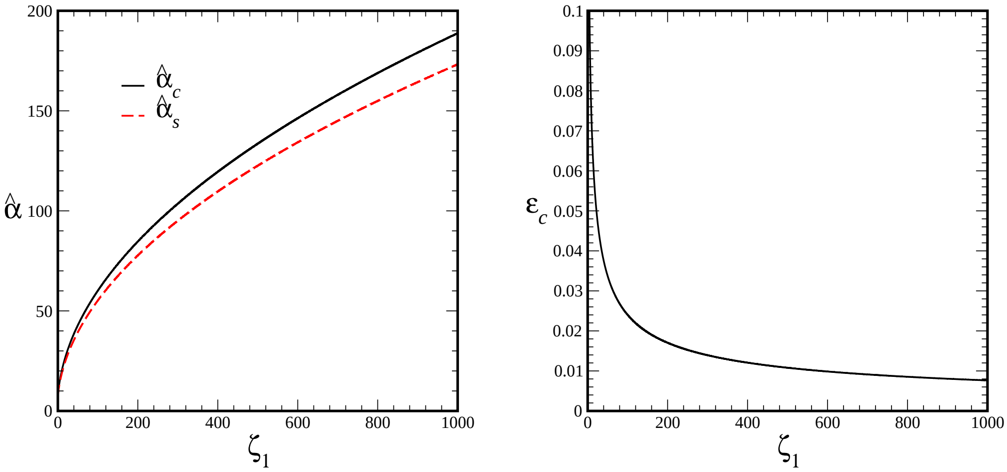

Fig. 2 Critical parameters $\hat{\alpha}_s$, $\hat{\alpha}_c$ and ε_c as functions of the overlap ζ_1. In the left panel we observe the load parameters as a function of the teacher's overlap. The asymptotic behavior of both quantities is proportional to $\sqrt{\zeta_1}$. In the left panel we observe the asymptotic decay to zero of the generalization error at the transition $\hat{\alpha}_c$.

lower the generalization error at the critical point $\varepsilon_c \equiv \varepsilon_G(\hat{\alpha}_c)$. Thus, for large values of the overlap ζ_1 we recover the perceptron like behavior where full generalization is achieved at very low values of $\hat{\alpha}$. For large values of the overlap ζ_1 we have that the critical parameter obeys the following equation:

$$0 = 2\lambda_c \sqrt{\frac{1+q_c}{1-q_c}}\, \frac{1+\sqrt{1-q_c^2}}{q_c} \left(\lambda_c - \sqrt{\frac{\pi}{2}-1}\right) - \ln(1-q_c) - q_c, \tag{15}$$

where $\lambda_c \equiv \frac{\pi}{2} - 1 + q_c - \arcsin(q_c)$. The numerical solution of (15) is $q_c \simeq 0.931$ which implies that $\hat{\alpha}_c \simeq 5.94\sqrt{\zeta_1}$ and the value of the generalization error at the criticality is $\varepsilon_c \simeq 0.248/\sqrt{\zeta_1}$. Graphs of the critical parameters $\hat{\alpha}_s$, $\hat{\alpha}_c$ and ε_c as functions of ζ_1, are presented in figure 2. The regression of these curves $\hat{\alpha}_c$ vs. ζ_1 and ε_c vs. ζ_1 confirm within a 1% the results presented above.

Thus, for $L = 1$ we have that the larger the overlap ζ_1 the larger the volume of information must be presented to the network to enter the learning (decreasing generalization error) phase. Although the generalization error in the *data acquisition* phase ($q = 0$) gets smaller, this tradeoff relationship suggests that, if the overlap is large enough, maybe a perceptron that saturates to a non-zero generalization error and a $O(N)$ training set is more economical than a $L = 1$ UCM with a large overlap and a $O(KN)$ training set.

For $L > 1$ we have a new and interesting result regarding multilayer feed-forward networks. By proceeding in similar way as in the example before we can study the behavior of the critical parameters as functions of the teacher overlaps $\zeta_1, \ldots, \zeta_L$. As we do so, we can find different values of the teacher's overlaps that induce equivalent learning processes, i.e. produce the same critical parameters. This result, for $L = 2$ is presented in figure 3.

To illustrate the behavior of the student network with respect to the large overlap limit, we consider the case of a UCM learning from a teacher with $L = 2$. We assume the symmetric regime, thus if $\alpha \sim O(1)$ we suppose that $\zeta_1 = 0$. The saddle point equation (8) can be written as:

$$0 = \alpha\, \mathscr{Q}_2^{(\infty)} \mathcal{I}(\mathscr{Q}_2) - \frac{1}{2} \frac{\mathscr{Q}_2}{\mathscr{Q}_2^{(\infty)} - \mathscr{Q}_2}, \tag{16}$$

where $\mathscr{Q}_2^{(\infty)}$ is as presented in (13). Equation (16) is precisely the saddle point equation of the perceptron when $\mathscr{Q}_2^{(\infty)} \uparrow 1$, which is the limit value reached when $\zeta_2 \uparrow \infty$. Observe that total generalization is achievable at this limit for large values of α. In a very similar way, when $\alpha = \hat{\alpha} K_2$, we can write the saddle point equation as

$$0 = \hat{\alpha}\, \frac{2}{\pi} \left[\frac{1}{\sqrt{1-\mathscr{Q}_1^2}} - 1 \right] \mathscr{Q}_1^{(\infty)} \mathcal{I}(\mathscr{Q}_2^{(\mathrm{eff})}) - \frac{1}{2} \frac{\mathscr{Q}_1}{\mathscr{Q}_1^{(\infty)} - \mathscr{Q}_1} + O(\varphi), \tag{17}$$

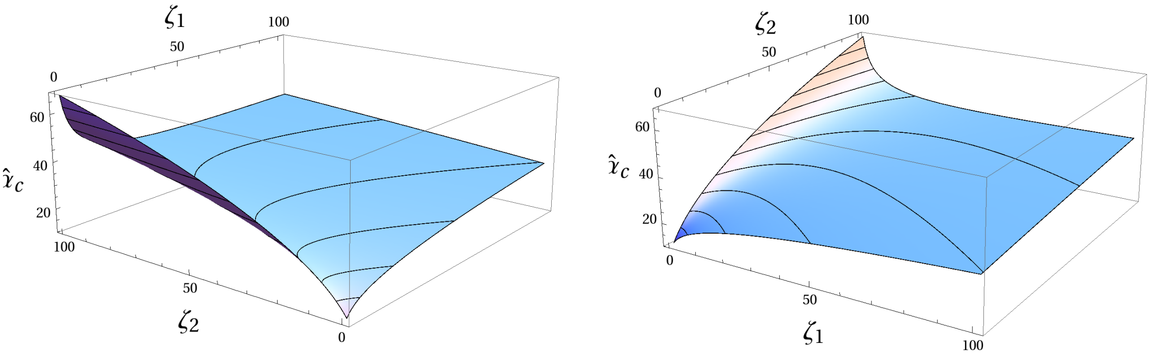

Fig. 3 Critical value of $\hat{\alpha}_c$ against ζ_1 and ζ_2. The lines are drawn at constant $\hat{\alpha}$ and thus represent sets of points whose coordinates describe different teachers with identical $\hat{\alpha}_c$.

where $\varphi \equiv (\frac{2}{\pi})^2(1+\frac{2}{\pi}\zeta_1)^{-1}\zeta_2$ and $\mathscr{Q}_2^{(\mathrm{eff})} \equiv \frac{2}{\pi}\left[\arcsin(\mathscr{Q}_1)+\mathscr{Q}_1^{(\infty)}-\mathscr{Q}_1\right]$. Equation (17) is equivalent to the saddle point equation of a system with $L=1$ in the limit $\mathscr{Q}_1^{(\infty)} \uparrow 1$ and $\varphi \downarrow 0$. Such a behavior is achieved for large values of the overlap ζ_1. Again, for the large overlap limit and for large values of $\hat{\alpha}$, total generalization is asymptotically obtained.

4 Conclusions

UCMs are feed-forward, binary neural networks that can be considered as the perceptron's next level of architectural complexity. They have only recently been study for the first time and, although they theoretically present more capabilities than the perceptron, they have not been used in real world applications yet.

Probably the most appealing feature UCMs have is the direct relationship between network complexity (as a function of the UCM's overlaps) and task difficulty. If a task difficulty can be assessed by measuring its sensitivity (as presented in [15] and [17]), then a suitable UCM may be constructed to cope with it. It is natural to continue this research by exploring the learning process in UCMs.

In this article we presented a study of the learning-by-examples process in UCMs. Our results were obtained by the application of statistical mechanics techniques, more precisely, by the application of the replica trick, with the imposition of replica symmetric ansatz. Our analysis is based on the study of regimes characterized by the number of examples presented to the student. The regimes of interest are those where the load parameter α is proportional to the number of units in the ℓ-th hidden layer of the teacher. To simplify this analysis we considered teachers with the first $\ell-1$ overlaps equal to zero. Although apparently less complex, such systems present a qualitatively identical critical behavior to systems with a full set of non-vanishing teacher overlaps.

Our first result, equation (8), is the saddle point equation correspondent to the regime where α is of order 1. This equation admits only one solution, which is the only minimum of the free energy, for all values of α. Thus, this regime is characterized by a lack of transitions and a decay of the generalization error to a non-vanishing value.

The first regime that admits a phase transition occurs when the load parameter is proportional to the number of hidden units in the more external (closest to the output) hidden layer. This first order transition is from a symmetric phase with zero inter-replica overlap and constant value of the generalization error (data acquisition phase), to a specialized phase, with a non-zero inter-replica overlap and decaying (and non-negligible) generalization error (generalization phase). This behavior is repeated for regimes where α is proportional to the number of units in a given hidden layer, but the first. If $\alpha \sim O(K)$ then after the transition the generalization error asymptotically vanishes when $\hat{\alpha} \uparrow \infty$. In all the cases where the transition is observed, the drop in the generalization error is discontinuous. We also observed that teachers with large overlaps effectively appear to their students as less architecturally complex UCMs. This result appeals to the consideration of the tradeoff between architectural complexity and network performance. It is probably due to this tradeoff that more economical (and simpler) networks could effectively perform equally good as a more complex (and training demanding) network.

Finally, we found that if $L > 1$ we can find many teachers with different synaptic overlaps and identical critical parameters. This indicates that, although representing different Boolean functions and implementing different classification tasks, these teachers induce the same learning process in student networks. This motivates a classification of teachers in equivalence classes that may simplify the study of Boolean functions so implemented.

Acknowledgements The author would like to acknowledge the friendly criticisms from Dr Roberto C. Alamino and Dr Laura Rebollo-Neira.

References

1. Variano E., AMcCoy J. H. and Lipson H.: Networks, Dynamics, and Modularity. Phys. Rev. Lett. **92**, 188701 (2004).
2. Huerta R. and Rabinovich M.: Reproducible Sequence Generation In Random Neural Ensembles. Phys. Rev. Lett. **93**, 238104 (2004).
3. Yoshioka M.: Learning of Spatiotemporal Patterns in Ising-Spin Neural Networks: Analysis of Storage Capacity by Path Integral Methods. Phys. Rev. Lett. **102**, 158102 (2009).
4. Lin I. -H., Wu R. -K. and Chen C. -M.: Synchronization in a noise-driven developing neural network. Phys. Rev. E **84**, 051923 (2011).
5. Saito A., Taiji M. and Ikegami T.: Inaccessibility in Online Learning of Recurrent Neural Networks. Phys. Rev. Lett. **93**, 168101 (2004).
6. Fiete I. R. and Seung H. S.: Gradient Learning in Spiking Neural Networks by Dynamic Perturbation of Conductances. Phys. Rev. Lett. **97**, 048104 (2006).
7. Saito H., Katahira K., Okanoya K. and Okada M.: Statistical mechanics of structural and temporal credit assignment effects on learning in neural networks. Phys. Rev. E **83**, 051125 (2011).
8. Braunstein A., Ramezanpour A., Zecchina R. and Zhang P.: Inference and learning in sparse systems with multiple states. Phys. Rev. E **83**, 056114 (2011).
9. Koralek A. C., Jin X., Long II J. D., Costa R. M. and Carmena J. M.: Corticostriatal plasticity is necessary for learning intentional neuroprosthetic skills. Nature **483**, 331 (2012).
10. Bardin J.: Neuroscience: Making connections. Nature **483**, 394 (2012).
11. Mesgarani N. and Chang E. F.: Selective cortical representation of attended speaker in multi-talker speech perception. Nature **485**, 233 (2012).
12. Seung H. S., Sompolinsky H. and Tishby N.: Statistical mechanics of learning by examples. Phys. Rev. A **45**, 6056-6091 (1992).
13. Neirotti J. P.: Can a student learn optimally from two different teachers? J. Phys. A **43**, 015101 (2010).
14. Neirotti J. P.: Parallel strategy for optimal learning in perceptrons. J. Phys. A **43**, 125101 (2010).
15. Neirotti J. P. and Franco L.: Computational capabilities of multilayer committee machines. J. Phys. A **43**, 445103 (2010).
16. Neirotti J. P. and Caticha N.: Dynamics of the evolution of learning algorithms by selection. Phys. Rev. E **67**, 041912 (2003).
17. Franco L. and Anthony M.: The influence of oppositely classified examples on the generalization complexity of Boolean functions. IEEE Trans. Neural Netw. **17**, 578-590 (2006).
18. Rammal R., Toulouse G. and Virasoro M. A.: Ultrametricity for physicists. Rev. Mod. Phys. **58**, 765-788 (1986).
19. Engel A., Khler H. M., Tschepke F., Vollmayr H. and Zippelius A.: Storage capacity and learning algorithms for two-layer neural networks. Phys. Rev. A **45**, 7590-7609 (1992).
20. Monasson R. and O'Kane D.: Domain of solutions and replica symmetry breaking in multilayer neural networks. Europhys. Lett. **27**, 85-90 (1994).
21. Monasson R. and Zecchina R.: Weight Space Structure and Internal Representations: A Direct Approach to Learning and Generalization in Multilayer Neural Networks. Phys. Rev. Lett. **75**, 2432-2435 (1995).
22. Schwarz H.: Learning rule in a multilayer neural network. J. Phys. A **26**, 5781-5794 (1993).
23. Schwarze H. and Hertz J.: Learning from examples in fully connected committee machines. J. Phys. A **26**, 4919-4936 (1993).