Finite Size Effects in Neural Networks

Laura Viana[1], Arnulfo Castellanos[2], and A.C.C. Coolen[3]

[1] Centro de Ciencias de la Materia Condensada, UNAM
A. Postal 2681, 22800 Ensenada, B.C., México
laura@ccmc.unam.mx http:www.ccmc.unam.mx
[2] Dept. de Física, Universidad de Sonora
A. Postal 1626, Hermosillo 83000, Son., México
acastell@fisica.uson.mx
[3] Dept. of Mathematics, King's College, University of London
Strand, London WC2R 2LS, U.K.
tcoolen@mth.kcl.ac.uk

Abstract. In this paper we give an overview of a recently developed theory [1, 2] which allows for calculating finite size corrections to the dynamical equations describing the dynamics of separable Neural Networks, away from saturation. According to this theory, finite size effects are described by a linear-noise Fokker Planck equation for the fluctuations (corresponding to an Ornstein-Uhlenbeck process), whose solution is characterized by the first two moments. The theory is applied to a particular problem in which detailed balance does not hold.

PACS: 87.30, 05.20

1 Introduction

Most Statistical Mechanics theories for Infinite-range spin models of Neural Networks (NN) are valid in the thermodynamical limit. However, finite size effects have often been reported in the literature [3]. Secondly, the first surge of studies of NN, concentrated on the study of the equilibrium properties of these systems (for a review see [4]), whereas it is now generally accepted that due to the intrinsic storage properties of these systems, the study of the dynamics is essential to achieve a better understanding of the behaviour of NN. Here, we review a recently developed theory which considers both aspects, as it allows us to study the dynamics of finite separable NN, away from saturation, by making a correction of order (1/N) to the dynamical mean field equations. Besides, the theory can be used in systems either satisfying, or not satisfying, detailed balance.

In this paper we give an overview of the derivation of the theory, which has as a starting point the master equation for the microscopic probability distribution. This distribution is then rewritten in terms of the overlaps, which are macroscopic variables which -it is assumed- contain all relevant information about the state of the system; in this way all irrelevant information is eliminated from the theory. By doing this, a Kramers-Moyal equation is obtained which can be

expanded in powers of 1/N. If the two leading orders are kept, one obtains a linear-noise Fokker Planck equation, which describes the stochastic behaviour of the overlaps, at least on finite timescales (those not scaling with the system size N). The theory then provides us with a general solution to the Fokker-Planck equation, which is subsequantly applied to study a specific problem. We would like to point out that the problem considered has already been studied in [1]; however, in this pap! er it is solved with more generality and detail than before.

1.1 The Theory

This theory considers a spin model for a NN, composed by a large number N of interconnected neurons modeled as Ising spins $\sigma_i = \pm 1$, for $i = 1, \ldots, N$. The master equation for the microscopic probability distribution $p_t(\boldsymbol{\sigma})$ is given by

$$\frac{d}{dt} p_t(\boldsymbol{\sigma}) = \sum_i \{ w_i(F_i\boldsymbol{\sigma}) p_t(F_i\boldsymbol{\sigma}) - w_i(\boldsymbol{\sigma}) p_t(\boldsymbol{\sigma}) \} \ , \tag{1}$$

where F_i is an operator that flips the $i-$th spin, i.e. $F_i f(\sigma_1, \ldots, \sigma_N) = f(\sigma_1, \ldots, -\sigma_i, \ldots, \sigma_N)$, and $w_i(\sigma)$ is the probability per unit time of the $i - th$ spin being flipped at time t, and it is given by

$$w_i(\boldsymbol{\sigma}) = \frac{1}{2}[1 - \sigma_i \tanh(\beta h_i(\boldsymbol{\sigma}))] \ . \tag{2}$$

In this expression, the inverse of β $(= T^{-1})$ measures the noise level , and the local field $h_i(\boldsymbol{\sigma})$ is given by $h_i(\boldsymbol{\sigma}) = \sum_j J_{ij}\sigma_j$, where J_{ij} is the strength of the synaptic connection from neuron j to neuron i. These connections contain information about p randomly chosen (and fixed) binary patterns $\boldsymbol{\xi}^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu) \in \{-1, 1\}^N$, with $\mu = 1, \ldots, p$, by means of a learning rule given by

$$J_{ij} = [1 - \delta_{ij}] \frac{1}{N} \sum_{\mu\nu=1}^{p} \xi_i^\mu A_{\mu\nu} \xi_j^\nu \ . \tag{3}$$

It is possible to obtain a macroscopic description of the system, by introducing the pattern overlaps

$$\boldsymbol{m}(\boldsymbol{\sigma}) = (m_1(\boldsymbol{\sigma}), \ldots, m_p(\boldsymbol{\sigma})) \ , \qquad m_\mu(\boldsymbol{\sigma}) = \frac{1}{N} \sum_i^N \xi_i^\mu \sigma_i \ , \tag{4}$$

which measure the resemblance between the state of the system and each of the stored random patterns. The probability density for the macroscopic variables $\boldsymbol{m}$ is given by:

$$P_t(\boldsymbol{m}) \equiv \sum_{\boldsymbol{\sigma}} p_t(\boldsymbol{\sigma}) \delta \left[\boldsymbol{m} - \boldsymbol{m}(\boldsymbol{\sigma}) \right] \ . \tag{5}$$

By rewriting the microscopic master equation (1) in terms of this new variable, and doing some algebra, it is possible to arrive at a Kramers-Moyal expansion

for the probability density $P_t(\boldsymbol{m})$ of the macroscopic variables; this expression is then expanded in powers of (1/N) [for full details of the calculation see [1]]. In the thermodynamical limit $N \to \infty$, this expansion reduces to a Liouville equation:

$$\frac{d}{dt}P_t(\boldsymbol{m}) = \sum_{\mu=1}^{p} \frac{\partial}{\partial m_\mu} \left\{ P_t(\boldsymbol{m}) \left[m_\mu - \langle \xi_\mu \tanh \beta \left[\boldsymbol{\xi} \cdot \boldsymbol{A}\boldsymbol{m} \right] \rangle_{\boldsymbol{\xi}} \right] \right\} ,$$

with the deterministic solution

$$P_t(\boldsymbol{m}) = \delta \left[\boldsymbol{m} - \boldsymbol{m}^*(t) \right] , \qquad \frac{d}{dt}\boldsymbol{m}^*(t) = \langle \boldsymbol{\xi} \tanh \beta \left[\boldsymbol{\xi} \cdot \boldsymbol{A}\boldsymbol{m}^*(t) \right] \rangle_{\boldsymbol{\xi}} - \boldsymbol{m}^*(t) , \quad (6)$$

where we defined $\langle g[\boldsymbol{\xi}] \rangle_{\boldsymbol{\xi}} = \lim_{N\to\infty} \frac{1}{N} \sum_i g[\boldsymbol{\xi}_i]$, with $\boldsymbol{\xi}_i = (\xi_i^1, \ldots, \xi_i^p)$.

If instead of taking the thermodynamical limit, the two leading orders in the expansion are kept, a Fokker-Planck equation is obtained for the overlaps $\boldsymbol{m}(\boldsymbol{\sigma})$. Since we are interested in evaluating finite size effects, as corrections to the mean field equations, it is convenient to rewrite the overlap (4) as the sum of a deterministic term $\boldsymbol{m}^*(t)$ representing the state of the system, as predicted in the thermodynamical limit and given by (6), plus a fluctuating stochastic term $\boldsymbol{q}(\boldsymbol{\sigma})/\sqrt{N}$ resulting from the finite size effects

$$\boldsymbol{m}(\boldsymbol{\sigma}(t)) = \boldsymbol{m}^*(t) + \frac{1}{\sqrt{N}}\boldsymbol{q}(t) . \quad (7)$$

In terms of the new variable $\boldsymbol{q}(t)$, it is possible to write the Fokker-Planck equation in the form

$$\frac{d}{dt}\mathcal{P}_t(\boldsymbol{q}) = \sum_\mu \frac{\partial}{\partial q_\mu} \left\{ \mathcal{P}_t(\boldsymbol{q}) F_\mu[\boldsymbol{q};t] \right\} + \sum_{\mu\nu} \frac{\partial^2}{\partial q_\mu \partial q_\nu} \left\{ \mathcal{P}_t(\boldsymbol{q}) D_{\mu\nu}[\boldsymbol{q};t] \right\} , \quad (8)$$

in which the flow term is given by

$$F_\mu[\boldsymbol{q};t] = K_\mu(t) + \sum_\nu L_{\mu\nu}(t) q_\nu \quad (9)$$

with

$$K_\mu(t) = \lim_{N\to\infty} \sqrt{N} \left\{ \langle \xi_\mu \tanh\{\beta[\boldsymbol{\xi}\cdot\boldsymbol{A}\boldsymbol{m}^*(t)]\} \rangle_{\boldsymbol{\xi}} - \frac{1}{N} \sum_i \xi_i^\mu \tanh\{\beta \left[\boldsymbol{\xi}_i \cdot \boldsymbol{A}\boldsymbol{m}^*(t) \right]\} \right\} , \quad (10)$$

$$L_{\mu\nu}(t) = \delta_{\mu\nu} - \beta \sum_\lambda \langle \xi_\mu \xi_\lambda \left[1 - \tanh^2\{\beta[\boldsymbol{\xi}\cdot\boldsymbol{A}\boldsymbol{m}^*(t)]\} \right] \rangle_{\boldsymbol{\xi}} A_{\lambda\nu} . \quad (11)$$

As we can see, $K_\mu(t)$ describes a 'frozen' correction to the flow field, which depends explicitly on the microscopic realization of the pattern components $\{\xi_i^\mu\}$,

and vanishes in the thermodynamical limit. Secondly, the diffusion matrix $D_{\mu\nu}$ in (8) is found to be symmetric, independent of $\boldsymbol{q}(t)$, and given by

$$D_{\mu\nu}(t) = \delta_{\mu\nu} - e^{-t} \lim_{N\to\infty} \frac{1}{N} \sum_i \xi_i^\mu \xi_i^\nu \sigma_i(0) \tanh\{\beta[\boldsymbol{\xi}_i \cdot \boldsymbol{A}\boldsymbol{m}^*(t)]\}$$

$$- \int_0^t ds\ e^{s-t} \langle \xi_\mu \xi_\nu \tanh\{\beta[\boldsymbol{\xi}\cdot\boldsymbol{A}\boldsymbol{m}^*(s)]\} \tanh\{\beta[\boldsymbol{\xi}\cdot\boldsymbol{A}\boldsymbol{m}^*(t)]\}\rangle_{\boldsymbol{\xi}} \,. \qquad (12)$$

This (8) is a 'linear noise' Fokker-Planck equation, and it describes a so-called time dependent Ornstein-Uhlenbeck process (see e.g. [5]), whose solution is a Gaussian distribution.

$$\mathcal{P}_t(\boldsymbol{q}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \boldsymbol{\Xi}(t)}}\ \exp\left\{-\frac{1}{2}[\boldsymbol{q} - \langle \boldsymbol{q}\rangle_t]\cdot \boldsymbol{\Xi}^{-1}(t)[\boldsymbol{q} - \langle \boldsymbol{q}\rangle_t]\right\} , \qquad (13)$$

where $\boldsymbol{\Xi}(t)$ is defined as $\Xi_{\mu\nu}(t) = \langle q_\mu q_\nu\rangle - \langle q_\mu\rangle\langle q_\nu\rangle$, and $\langle\cdots\rangle$ denotes an average over an ensemble of initial states $\{\sigma_i(0)\}$ such that $\boldsymbol{m}(0) = \boldsymbol{m}^*(0) + \frac{1}{\sqrt{N}}\boldsymbol{q}(0)$, with $\boldsymbol{q}(0) \sim \mathcal{O}(1)$. Therefore, the complete solution is determined by the first two moments, which are given by the solution of

$$\frac{d}{dt}\langle \boldsymbol{q}\rangle = -\boldsymbol{L}(t)\langle \boldsymbol{q}\rangle - \boldsymbol{K}(t) \,, \qquad (14)$$

$$\frac{d}{dt}\boldsymbol{\Xi}(t) = -\boldsymbol{L}(t)\boldsymbol{\Xi}(t) - \boldsymbol{\Xi}(t)\boldsymbol{L}^\dagger(t) + 2\boldsymbol{D}(t) \,. \qquad (15)$$

2 Escape from a Basin of Attraction as a Final Size Effect

In this section we will use this theory to study a NN where detailed balance does not hold, so it is not possible to define an energy function whose minima are the attractors of the dynamics of the system. This kind of problem is characterized by a non symmetric interaction matrix $\boldsymbol{A}$ [c.f. (3)]. As we will see, the finite version of this system presents a qualitatively different behaviour to that corresponding to its infinite counterpart. The system to be considered has two patterns $\boldsymbol{\xi}^\mu$, $(\mu = 1, 2)$, stored according tho the learning rule (3) with an interaction matrix $\boldsymbol{A} = \{\{1,-1\},\{1,1\}\}$. These patterns are randomly drawn with equal probability $\xi_i^p = \pm 1$, so the overlaps between them are of order $N^{-1/2}$; therefore, in the thermodynamic limit the only states $\boldsymbol{m}(\boldsymbol{\sigma})$ which can exist are those enclosed by the rhombus formed by the dotted lines in Fig. (1-a). In the noiseless case, i.e. at $T = 0$, the infinite version has four basins of attraction, each corresponding to one of the quadrants in the space of states (m_1^*, m_2^*), so the separatrices of these regions are the lines $y = 0$, $x = 0$, $x \pm y = \pm 1$; there are four fixed points located over the separatices between two of the basins at $\pm(1, 0)$ and $\pm(0, 1)$. We will analyze the behaviour of this system assuming any arbitrary valid initial state within the first quadr! ant (satisfying $m_1^* + m_2^* \leq 1$), generalization of the results to other initial states is straightforward if the symmetries of the problem

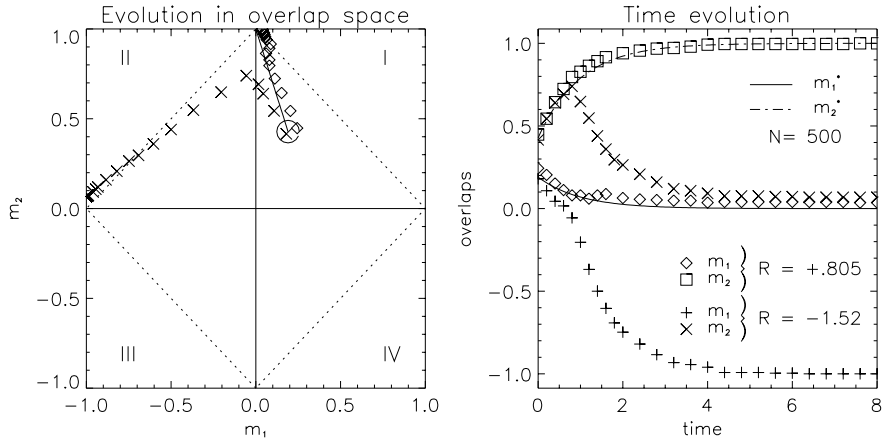

Fig. 1. a) Full line represents the evolution of an infinite system, with the centre of the circle showing the value of $\boldsymbol{m}^*(0)$, while markers represent the evolution of two particular finite systems of the same size ($N = 500$), with positive and negative R, respectively, and the same macroscopical initial state; dotted line encloses the region where the infinite system can exist. b) Solid (dot dash) line represents the time evolution of the overlaps $m_1^*(t)$ ($m_2^*(t)$) of the infinite system, while the symbols $\diamond$, and $+$ ($\square$ and $\times$) represent the actual evolution of the overlaps $m_1(t)$ ($m_2(t)$) for the same two systems

are considered. By solving (6) it is possible to demonstrate that the infinite system evolves in straight line, with asymptotically decreasing speed, towards the fixed point $(0, 1)$, with $\boldsymbol{m}^*(t)$ given by

$$m_1^*(t) \;=\; m_1^*(0)\; e^{-t}\,, \qquad\qquad m_2^*(t) \;=\; 1 + \{m_2^*(0) - 1\}\; e^{-t}\,, \tag{16}$$

as shown in Fig. (1-a). In order to study finite size effects, we need to consider an specific realization of the system, characterized by its value of R, the frozen correction (10) in the noiseless $T \to 0$ limit

$$R = \frac{1}{\sqrt{N}} \sum_i \xi_i^1 \xi_i^2\,, \tag{17}$$

and make an ansatz of initial conditions consistent with the value of $\boldsymbol{m}^*(0)$. We will consider the most general case which consists in assuming two condensed patterns at $t = 0$; this ansatz is represented by initial states given by the probability density [1]

$$p_0(\boldsymbol{\sigma}) = \prod_i \Big\{ \;|m_1^*(0)|\; \delta_{\sigma_i,\xi_i^1} \;+\; |m_2^*(0)|\; \delta_{\sigma_i,\xi_i^2} \;+$$

[1] Notice that we are considering a more general case than in [1]

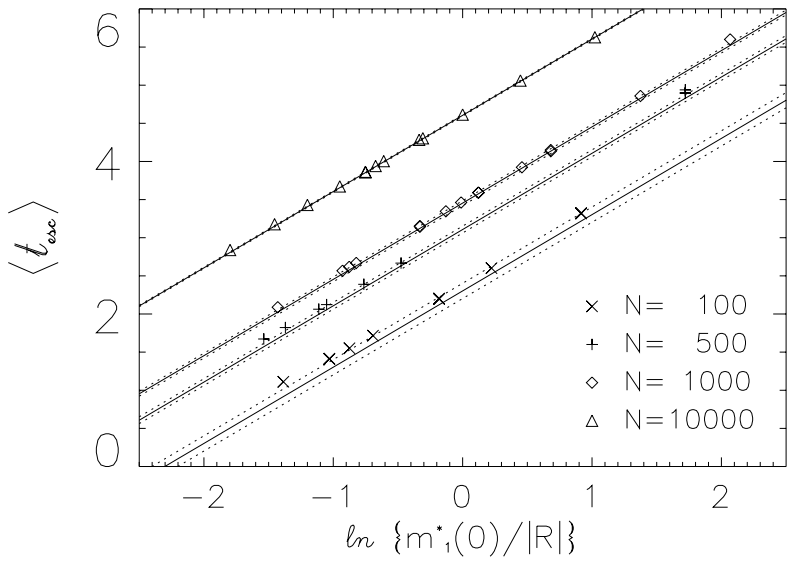

Fig. 2. In this figure solid lines show the theoretical predictions, while doted lines show the precision of the theory. Markers indicate the result of computer simulations in systems of different sizes.

$$+\frac{1}{2}\left(1-|m_1^*(0)|-|m_2^*(0)|\right)\ \left(\delta_{\sigma_i,1}+\delta_{\sigma_i,-1}\right)\Big\}\ . \tag{18}$$

If we take an average of $\boldsymbol{m}(0)$ over this ansatz, and keep in mind the definition of $\boldsymbol{q}(t)$, given by (7), it is easy to demonstrate that the initial conditions are given by

$$\langle \boldsymbol{q}_1(0)\rangle = m_2^*(0)R\,, \qquad \langle \boldsymbol{q}_2(0)\rangle = m_1^*(0)R\,,$$
$$\boldsymbol{\Xi}(0) = \left\{1-(m_1^*(0))^2-(m_2^*(0))^2\right\}\delta_{\mu\nu} - 2m_1^*(0)m_2^*(0)(1-\delta_{\mu\nu})\,. \tag{19}$$

On the other hand, the relevant quantities (10-12) in the Fokker-Planck equation (8) are given by

$$\boldsymbol{K}(t) = -\begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad \boldsymbol{L}(t) = \boldsymbol{I}, \qquad \boldsymbol{D}(t) = e^{-t}\begin{pmatrix} 1-m_2^*(0) & -m_1^*(0) \\ -m_1^*(0) & 1-m_2^*(0) \end{pmatrix},$$

where R is given by (17). By solving (14-15) with the initial conditions (19), we find that the finite system evolves following a random path with average fluctuations given by [c.f. eq(7)]

$$\langle q_1(t)\rangle \ = \ R\ m_2^*(t)\,, \qquad \langle q_2(t)\rangle \ = \ R\ m_1^*(t)\,, \tag{20}$$

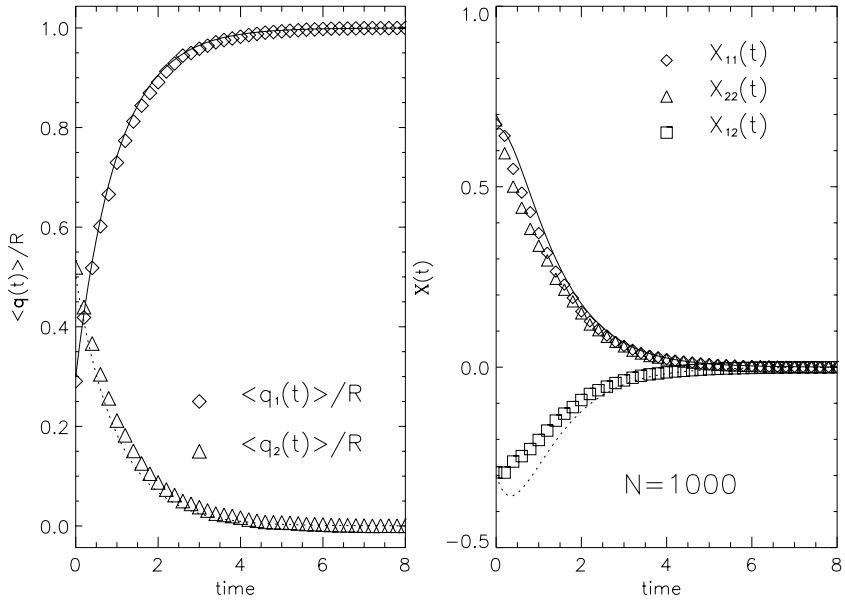

Fig. 3. Computer simulations (markers) versus theory (lines) for a system with $N = 1000$ and $R = +2.15$. a) average of finite size fluctuations $\langle q_1 \rangle / R$, $\langle q_2 \rangle / R$, b) Central correlations $\boldsymbol{\Xi}(t)$

.

$$\boldsymbol{\Xi}(t) = \boldsymbol{\Xi}(0) e^{-2t} + 2\boldsymbol{D}(0) e^{-t} \left(1 - e^{-t}\right) \ , \tag{21}$$

Figure (1) ilustrates the different behaviour of two systems of the same size ($N = 500$) and the same value of $\boldsymbol{m}^*(t)$, but with frozen corrections of different sign. As we can see, finite size effects displace the position of the fixed point on a direction which depends on the sign of R. This allows a system with $R < 0$ to jump out from the first quadrant, and evolve towards the fixed point of the second quadrant. The escape from this region will happen at $t = t_{esc}$ with t_{esc} defined by $m_1(t_{esc}) = m_1^*(t_{esc}) + q_1(t_{esc})/\sqrt{N} = 0$, which will have different values for different initial states $\{\sigma_i(0)\}$. Although we know nothing for a given particular process, this theory allows us to evaluate the average escape time, by using instead $\langle q_1(t_{esc}) \rangle$ in the above expression, together with (16,20), to obtain:

$$\langle t_{esc} \rangle = \frac{1}{2} \log N + \log \left\{ \frac{m_1^*(0)}{|R|} \right\} \ . \tag{22}$$

Figure (2) shows the theoretical predictions of this theory (22) for the average escape time, and the actual measured average escape time obtained from computing simulations performed in systems of different sizes.

Finally, we present on Fig. (3) a comparison between the theoretical predictions for the first two moments of the fluctuations, and computer simulations

performed over an ensemble of $n = 1600$ different initial conditions chosen according to (18), for one system with $N = 1000$, and $R = 2.15$. As we can see, the simulations are in reasonable agreement with this theory, since its theoretical precision is of $\mathcal{O}(N^{-1/2}) \sim 3.16\%$, while we have an uncertainty over the computer simulations of $\mathcal{O}(n^{-1/2}) = 2.5\%$.

3 Discussion

Finite size effects in NN can be very important, and as we have shown, some finite systems can even behave very differently from their infinite counterparts, so it is important to have a theory which allows for their evaluation. The present theory can be successfully applied to study such effects to leading order in the system size ($N^{-1/2}$), in a wide variety of systems, away from saturation, and the time dependent probability density can be explicitly calculated. These systems are allowed to have either symmetric or non symmetric interactions, may present non zero noise levels ($T > 0$), and even store biased patterns $\langle \boldsymbol{\xi}^{\mu} \rangle \neq 0$. Here, for mathematical simplicity, we chose to consider a system with only two stored patterns, but the results can be equally applied to systems with a higher number of them.

Acknowledgments

The authors wish to thank Carlos González S. and Citlali Martínez for their help in the formatting of the text. This project was partially supported by grants DGAPA IN100895, DGAPA 123098, from the National University of Mexico.

References

1. Castellanos, A., Coolen, A.C.C., Viana, L.: Finite Size effects in separable recurrent Neural Networks, J. Phys. A: Math. Gen.**31** (1998) 6615–6634
2. Castellanos, A., Ph.D. Thesis, CICESE-UNAM, México (1998).
3. Kohring G.A.: J. Phys. A: Math. Gen. **23** (1990) 2237
4. Coolen, A.A.C. and Sherrington D.: Mathematical Approaches to Neural Networks, ed. J.G. Taylor (Amsterdam, North Holland) p 293
5. Gardiner C W 1990 *Handbook of Stochastic Methods* (Berlin: Springer)