EUROPHYSICS LETTERS 10 April 1996

Europhys. Lett., **34** (2), pp. 151-156 (1996)

Finite-size effects in on-line learning of multilayer neural networks

D. Barber[1], D. Saad[1] and P. Sollich[2]

[1] *Department of Computer Science and Applied Mathematics, Aston University Birmingham B4 7ET, UK*
[2] *Department of Physics, University of Edinburgh, Kings Buildings Edinburgh EH9 3JZ, UK*

(received 7 July 1995; accepted in final form 26 February 1996)

PACS. 87.10+e – General, theoretical, and mathematical biophysics (including logic of biosystems, quantum biology, and relevant aspects of thermodynamics, information theory, cybernetics, and bionics).
PACS. 02.50−r – Probability theory, stochastic processes, and statistics.

Abstract. – We complement recent advances in thermodynamic limit analyses of mean on-line gradient descent learning dynamics in multilayer networks by calculating fluctuations possessed by finite-dimensional systems. Fluctuations from the mean dynamics are largest at the onset of specialisation as student hidden unit weight vectors begin to imitate specific teacher vectors, increasing with the degree of symmetry of the initial conditions. In light of this, we include a term to stimulate asymmetry in the learning process, which typically also leads to a significant decrease in training time.

An attractive feature of neural networks is their ability to *learn* a parametrised rule from a set of input/output training examples, by which the parameters of the network are adapted to minimise an error measuring the misfit of the network mapping on the training examples. Different approaches to the learning process are typically evaluated by the expected error that the network will make on a randomly presented input example. In *on-line learning*, statistical mechanics plays a strong role in calculating this *generalisation error* (see [1], [2], [4] and references within) through self-averaging in the thermodynamic limit, for which an understanding of finite-size effects would benefit further advances. Connections to alternative finite-dimensional methods (see [3] and references within) will be pointed to in the course of our analysis.

In *on-line learning*, the *weights* parametrising the *student* network are successively updated according to the error incurred on a single example from a stream of input/output examples, $\{\boldsymbol{\xi}^\mu, \tau(\boldsymbol{\xi}^\mu)\}$, generated by a *teacher* network $\tau(\cdot)$. We assume that the *teacher* network the student attempts to learn is a *soft committee machine* [1], [4] of N inputs, and M hidden units, this being a one-hidden-layer network with weights connecting each hidden to output unit set to $+1$, and with each hidden unit n connected to all input units by $\mathbf{B}_n (n = 1, \ldots, M)$.

Explicitly, for the N-dimensional training input vector $\boldsymbol{\xi}^\mu$, the output of the teacher is given by

$$\zeta^\mu = \sum_{n=1}^{M} g(\mathbf{B}_n \cdot \boldsymbol{\xi}^\mu), \tag{1}$$

where $g(x)$ is the activation function of the hidden units, and we take $g(x) = \mathrm{erf}(x/\sqrt{2})$. The teacher generates a stream of training examples $(\boldsymbol{\xi}^\mu, \zeta^\mu)$, with input components drawn from a normal distribution of zero mean, unit variance. The *student* network that attempts to learn the teacher, by fitting the training examples, is also a soft committee machine, but with K hidden units. For input $\boldsymbol{\xi}^\mu$, the student output is

$$\sigma(\mathbf{J}, \boldsymbol{\xi}^\mu) = \sum_{i=1}^{K} g(\mathbf{J}_i \cdot \boldsymbol{\xi}^\mu), \tag{2}$$

where the student weights $\mathbf{J} = \{\mathbf{J}_i\}$ $(i = 1, \ldots, K)$ are sequentially modified to reduce the *error* that the student makes on an input $\boldsymbol{\xi}^\mu$,

$$\epsilon(\mathbf{J}, \boldsymbol{\xi}^\mu) = \frac{1}{2}\left(\sigma(\mathbf{J}, \boldsymbol{\xi}^\mu) - \zeta^\mu\right)^2 = \frac{1}{2}\left(\sum_{i=1}^{K} g(x_i^\mu) - \sum_{n=1}^{M} g(y_n^\mu)\right)^2, \tag{3}$$

where the *activations* are defined $x_i^\mu = \mathbf{J}_i \cdot \boldsymbol{\xi}^\mu$, and $y_n^\mu = \mathbf{B}_n \cdot \boldsymbol{\xi}^\mu$. Gradient descent on the error (3) results in an update of the student weight vectors,

$$\mathbf{J}^{\mu+1} = \mathbf{J}^\mu - \frac{\eta}{N}\delta_i^\mu \boldsymbol{\xi}^\mu, \tag{4}$$

where

$$\delta_i^\mu = g'(x_i^\mu)\left[\sum_{n=1}^{M} g(y_n^\mu) - \sum_{j=1}^{K} g(x_j^\mu)\right], \tag{5}$$

and g' is the derivative of the activation function g. The typical performance of the student on a randomly selected input example is given by the generalisation error, $\epsilon_\mathrm{g} = \langle \epsilon(\mathbf{J}, \boldsymbol{\xi}) \rangle$, where $\langle \ldots \rangle$ represents an average over the Gaussian input distribution. One finds that ϵ_g depends only on the *order parameters*, $R_{in} = \mathbf{J}_i \cdot \mathbf{B}_n$, $Q_{ij} = \mathbf{J}_i \cdot \mathbf{J}_j$, and $T_{nm} = \mathbf{B}_n \cdot \mathbf{B}_m (i, j = 1, \ldots, K; n, m = 1, \ldots, M)$ [4], for which, using (4), we derive (stochastic) *update equations*,

$$R_{in}^{\mu+1} - R_{in}^\mu = \frac{\eta}{N}\delta_i^\mu y_n^\mu, \tag{6}$$

$$Q_{ik}^{\mu+1} - Q_{ik}^\mu = \frac{\eta}{N}\left(\delta_i^\mu x_j^\mu + \delta_k^\mu x_i^\mu\right) + \frac{\eta^2}{N^2}\delta_i \delta_k \boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\mu. \tag{7}$$

We average over the input distribution to obtain deterministic equations for the mean values of the order parameters, which are self-averaging in the *thermodynamic limit*, $N \to \infty$.

The order parameter approach contrasts with approaches which analyse the dynamics of the individual weight components, based upon approximate Fokker-Planck equations (see [3] and references within). The advantage of the order parameter approach is that the system is modelled exactly in the thermodynamic limit, with only a small number of equations.

In this work we present a more realistic treatment by calculating the dynamic fluctuations induced by finite-dimensional random inputs [5].

In the thermodynamic limit, we treat $\mu/N = \alpha$ as a continuous variable and form differential equations for the *thermodynamic overlaps*, R^0_{in}, Q^0_{ik},

$$\frac{\mathrm{d}R^0_{in}}{\mathrm{d}\alpha} = \eta \left\langle \delta_i y_n \right\rangle, \qquad \frac{\mathrm{d}Q^0_{ik}}{\mathrm{d}\alpha} = \eta \left\langle \delta_i x_k + \delta_k x_i \right\rangle + \eta^2 \left\langle \delta_i \delta_k \right\rangle. \tag{8}$$

For given initial overlap conditions, (8) can be integrated to find the mean dynamical behaviour of a student learning a teacher with an arbitrary number of hidden units [4] (see fig. 1 *a*)). Typically, ϵ_{g} decays rapidly to a *symmetric phase* in which there is near perfect symmetry between the hidden units. Such phases exist in learnable scenarios until sufficient examples have been presented to determine which student hidden unit will mimic which teacher hidden unit. For perfectly symmetric initial conditions, such *specialisation* is impossible in a mean dynamics analysis. The more symmetric the initial conditions are, the longer the trapping in the symmetric phase (see fig. 2 *a*)). Large deviations from the mean dynamics can exist in the symmetric phase, as a small perturbation from symmetry can determine which student hidden unit will specialise on which teacher hidden unit [1].

We can rewrite (6), (7) in the general form

$$a^{\mu+1} - a^{\mu} = \frac{\eta}{N} \left(F_a + \eta G_a \right), \tag{9}$$

where $F_a + \eta G_a$ is the *update rule* for a general overlap parameter a. In order to investigate finite-size effects, we make the following "small-fluctuations" ansätze([1]) [6] for the deviations of the update rules F_a (the same form is made for G_a) and overlap parameters a from their thermodynamic values([2]),

$$F_a = F^0_a + \triangle F_a + \frac{1}{N} F^1_a, \qquad a = a^0 + \sqrt{\frac{\eta}{N}} \triangle a + \frac{\eta}{N} a^1, \tag{10}$$

where $\langle \triangle F_a \rangle = \langle \triangle a \rangle = 0$. Terms of the form $\triangle a$ represent *dynamic* corrections that arise due to the random examples. Terms like a^1 represent *static* corrections such that the mean of the overlap parameter a is given by $a^0 + \eta a^1/N$—the thermodynamic average plus a correction. In order to simplify the analysis, we assume a small learning rate, η, so that the thermodynamic overlaps are governed by

$$\frac{\mathrm{d}a^0}{\mathrm{d}\tilde{\alpha}} = F^0_a, \tag{11}$$

where F^0_a is the update rule F_a averaged over the input distribution, and the rescaled learning rate is given by $\tilde{\alpha} = \eta\alpha$. Substituting (10) in (9) and averaging over the input distribution, we derive a set of coupled differential equations([3]) for the (scaled) covariances $\langle \triangle a \triangle b \rangle$, and static corrections a^1,

$$\frac{\mathrm{d}\langle \triangle a \triangle b \rangle}{\mathrm{d}\tilde{\alpha}} = \sum_c \langle \triangle a \triangle c \rangle \frac{\partial F^0_a}{\partial c^0} + \sum_c \langle \triangle b \triangle c \rangle \frac{\partial F^0_b}{\partial c^0} + \langle \triangle F_a \triangle F_b \rangle, \tag{12}$$

$$\frac{1}{2} \frac{\mathrm{d}^2 a^0}{\mathrm{d}\tilde{\alpha}^2} + \frac{\mathrm{d}a^1}{\mathrm{d}\tilde{\alpha}} = \sum_b b^1 \frac{\partial F^0_a}{\partial b^0} + \frac{1}{2} \sum_{bc} \langle \triangle b \triangle c \rangle \frac{\partial^2 F^0_a}{\partial b^0 \partial c^0} + G^1_a. \tag{13}$$

([1]) The activations have variance $\mathcal{O}(1)$ which, iterated through (9), yield overlap variances of $\mathcal{O}\left(N^{-1}\right)$.

([2]) If the order parameter represented by c is Q_{11}, then $c^0 = Q^0_{11}$, and $\triangle c = \triangle Q_{11}$.

([3]) The small-fluctuations ansatz necessarily yields equations of the same form as presented in [3] for the weight component dynamics—here they are for the order parameter representation of the system.

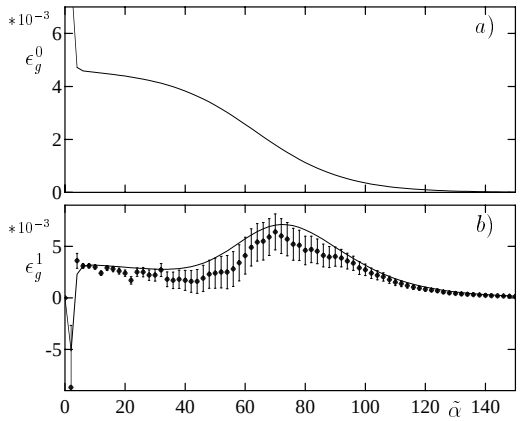

Fig. 1.

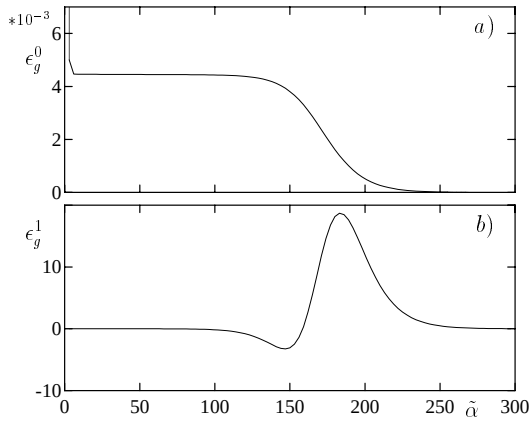

Fig. 2.

Fig. 1. – Two student hidden units, one teacher hidden unit. Non-zero initial parameters: $Q_{11} = 0.2, Q_{22} = R_{11} = 0.1$. *a*) Thermodynamic generalisation error, $\epsilon_{\rm g}^0$. *b*) $\mathcal{O}\left(N^{-1}\right)$ correction to the generalisation error, $\epsilon_{\rm g}^1$. Simulation results for $N = 10, \eta = 0.1$ and (half standard deviation) error bars are drawn.

Fig. 2. – Two student hidden units, one teacher hidden unit. Initially, $Q_{11} = 0.1$, with all other parameters set to zero. *a*) Thermodynamic generalisation error $\epsilon_{\rm g}^0$. *b*) $\mathcal{O}\left(N^{-1}\right)$ correction to the generalisation error, $\epsilon_{\rm g}^1$.

Summations are over all overlap parameters, $\{Q_{ij}, R_{in} | i, j = 1, \ldots, K, n = 1, \ldots, M\}$. The elements $\langle \triangle F_a \triangle F_b \rangle$ are found explicitly by calculating the covariance of the update rules F_a, and F_b. Initially, the fluctuations $\langle \triangle F_a \triangle F_b \rangle$ are set to zero, and eqs. (11)-(13) are then integrated to find the evolution of the covariances, $\mathrm{cov}(a, b) = (\eta/N) \langle \triangle a \triangle b \rangle$, and the corrections to the thermodynamic average values, $(\eta/N) a^1$. The average finite-size correction to the generalisation error is given by $\epsilon_{\rm g} = \epsilon_{\rm g}^0 + (\eta/N)\epsilon_{\rm g}^1$, where

$$\epsilon_{\rm g}^1 = \sum_a a^1 \frac{\partial e_{\rm g}^0}{\partial a} + \frac{1}{2} \sum_{ab} \langle \triangle a \triangle b \rangle \frac{\partial^2 \epsilon_{\rm g}^0}{\partial a^0 \partial b^0}. \tag{14}$$

These results enable the calculation of finite-size effects for an arbitrary teacher/student learning scenario. For demonstration, we calculate the finite-size effects for a student with two hidden units learning a teacher with one hidden unit. In this over-realisable case, one of the student hidden units eventually specialises to the single teacher hidden unit, while the other student hidden unit decays to zero. In fig. 1, we plot the thermodynamic limit generalisation error alongside the $\mathcal{O}\left(N^{-1}\right)$ correction. In fig. 1 *a*) there is no significant symmetric phase, and the finite-size corrections (fig. 1 *b*)) are small. For a finite-size correction of less than 10%, we would require an input dimension of around $N > 25\eta$. For the more symmetric initial conditions (fig. 2 *a*)) there is a very definite symmetric phase, for which a finite-size correction of less than 10% (fig. 2 *b*)) would require an input dimension of around $N > 50\,000\eta$. As the initial conditions approach perfect symmetry, the finite-size effects diverge, and the mean dynamical theory becomes inexact. Using the covariances, we can analyse the way in which the student breaks out of the symmetric phase by hidden-unit specialisation. For the isotropic teacher scenario $T_{nm} = \delta_{nm}$, and $M = K = 2$, learning proceeds such that one can approximate, $Q_{22} = Q_{11}, R_{22} = R_{11}$. By analysing the eigenvalues of the covariance matrix

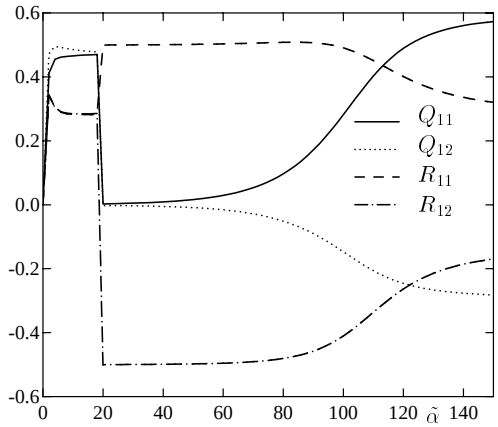

Fig. 3.

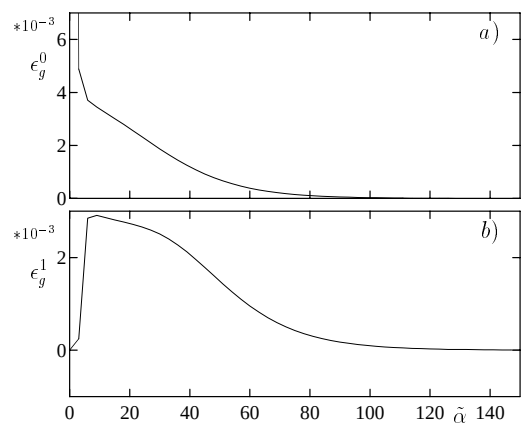

Fig. 4.

Fig. 3. – *a*) The normalised components of the principal eigenvector for an isotropic teacher. $M = K = 2$ ($Q_{22} = Q_{11}$, $R_{22} = R_{11}$). Non-zero initial parameters $Q_{11} = 0.2$, $Q_{22} = 0.1$, $R_{11} = 0.001$, $R_{22} = 0.001$.

Fig. 4. – Two student hidden units, one teacher hidden unit. The initial conditions are as in fig. 2. *a*) Thermodynamic generalisation error, $\epsilon_{\rm g}^0$. *b*) $\mathcal{O}\left(N^{-1}\right)$ correction to the generalisation error, $\epsilon_{\rm g}^1$.

$\langle \triangle a \triangle b \rangle$, we found a sharply defined principal direction, the components of which we show in fig. 3. Initially, all components of the principal direction are similarly correlated, which corresponds to the symmetric region. Then, around $\tilde{\alpha} = 20$, as the symmetry breaks, R_{11} and R_{21} become maximally anticorrelated, whilst there is minimal correlation between the Q_{11} and Q_{12} components, which corresponds with predictions from perturbation analysis [4]. The symmetry breaking is characterised by a specialisation process in which each student vector increases its overlap with one particular teacher weight, whilst decreasing its overlap with other teacher weights. After specialisation, there is a growth in the anticorrelation between the student length and its overlap with other students. The asymptotic values of these correlations are in agreement with the convergence fixed point, $R^2 = Q = 1$ [4].

In light of possible prolonged symmetric phases, we break the symmetry of the student hidden units by ordering the student lengths, $Q_{11} \geq Q_{22} \geq \ldots \geq Q_{KK}$. This constraint is enforced in a "soft" manner by including an extra term to (3),

$$\epsilon^\dagger = \frac{1}{2} \sum_{j=1}^{K-1} h\left(Q_{j+1j+1} - Q_{jj}\right), \tag{15}$$

where $h(x)$ approximates the step function, $h(x) = \left(1 + \mathrm{erf}\left(\beta x/\sqrt{2}\right)\right)/2$. This modification simply adds a Gaussian term in the student weight lengths to the weight update rule (cf.(4)). In fig. 4, we show the overlap parameters and their fluctuations for $\beta = 10$, $K = 2$, $M = 1$. This graph is to be compared to fig. 2 for which the initial conditions are the same. There is now no collapse to an initial symmetric phase from which the student will eventually specialise, and the initial convergence to the optimal values is much faster. The finite-size corrections are much reduced and are now largest around the initial value of $\tilde{\alpha}$ where the overlap parameters are very symmetric, becoming rapidly smaller due to the large driving force away from this near-symmetric region. For the case in which the teacher weights are equal, the constraint (15) will prevent the student from converging optimally, necessitating an adaptive soft constraint.

A naive scheme is to adapt the steepness, β, such that it is inversely proportional to the average of the gradients Q_{ii}, which decreases as the dynamics converge asymptotically.

In this work we have complemented the recent significant advances in the theory of on-line learning of multilayer networks by examining the conditions under which thermodynamic limit calculations are representative of real learning scenarios. Additionally, breaking the internal symmetries of the network reduces both finite-size effects and training time. We conjecture that such symmetry breaking is potentially of great benefit in the practical field of neural-network training. For extensions of thermodynamic limit analyses to the case of a limited number of adaptive hidden units to output unit weights, an understanding of finite-size effects will be of central importance.

***

This work was partially supported by the EU grant ERB CHRX-CT92-0063.

REFERENCES

[1] Biehl M. and Schwarze H., *J. Phys. A*, **28** (1995) 643.
[2] Copelli M. and Caticha N., *J. Phys. A*, **28** (1995) 1615.
[3] Heskes T., *J. Phys. A*, **27** (1994) 5145.
[4] Saad D. and Solla S., *Phys. Rev. E*, **52** (1995) 4225.
[5] Sollich P., *J. Phys. A*, **27** (1994) 7771.
[6] Van Kampen N., *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam) 1992.