Finite size effects in on-line learning of multi-layer neural networks.

David Barber[2], Peter Sollich[1] and David Saad[2]

[1] *Department of Physics, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom*

[2] *Neural Computing Research Group, Aston University, Birmingham B4 7ET, United Kingdom*

E-mail: D.Barber@aston.ac.uk

We complement the recent progress in thermodynamic limit analyses of mean on-line gradient descent learning dynamics in multi-layer networks by calculating the 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, and increase with the degree of symmetry of the initial conditions. Including a term to stimulate asymmetry in the learning process typically significantly decreases finite size effects and training time.

Recent advances in the theory of on-line learning have yielded insights into the training dynamics of multi-layer neural networks. In *on-line learning*, the *weights* parametrizing the *student* network are updated according to the error on a single example from a stream of examples, $\{\boldsymbol{\xi}^\mu, \tau(\boldsymbol{\xi}^\mu)\}$, generated by a *teacher* network $\tau(\cdot)$[1]. The analysis of the resulting weight dynamics has previously been treated by assuming an infinite input dimension (*thermodynamic limit*) such that a mean dynamics analysis is exact[2]. We present a more realistic treatment by calculating corrections to the mean dynamics induced by finite dimensional inputs[3].
We assume that the *teacher* network the student attempts to learn is a *soft committee machine*[1] 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 $B_n (n = 1..M)$. Explicitly, for the N dimensional training input vector $\boldsymbol{\xi}^\mu$, the output of the teacher is given by,

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

where $g(x)$ is the activation function of the hidden units, and we take $g(x) = erf(x/\sqrt{2})$. 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,

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

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

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

with the *activations* defined $x_i^\mu = J_i \cdot \boldsymbol{\xi}^\mu$, and $y_n^\mu = B_n \cdot \boldsymbol{\xi}^\mu$. Gradient descent on the error (3) results in an update of the student weight vectors,

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

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,

$$\epsilon_g = \langle \epsilon(J, \boldsymbol{\xi}) \rangle, \tag{6}$$

where $\langle .. \rangle$ represents an average over the gaussian input distribution. One finds that ϵ_g depends only on the *overlap parameters*, $R_{in} = J_i \cdot B_n$, $Q_{ij} = J_i \cdot J_j$, and $T_{nm} = B_n \cdot B_m (i, j = 1..K; n, m = 1..M)$[2], for which, using (4), we derive (stochastic) *update equations*,

$$R_{in}^{\mu+1} - R_{in}^\mu = \frac{\eta}{N}\delta_i^\mu y_n^\mu, \tag{7}$$

$$Q_{ik}^{\mu+1} - Q_{ik}^\mu = \frac{\eta}{N}\left(\delta_i^\mu x_j^\mu + \delta_k^\mu x_i^\mu\right) + \frac{\eta^2}{N^2}\delta_i \delta_k \boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\mu. \tag{8}$$

We average over the input distribution to obtain deterministic equations for the mean values of the overlap parameters, which are self-averaging in the thermodynamic limit. In this limit we treat $\mu/N = \alpha$ as a continuous variable and form differential equations for the *thermodynamic overlaps*, R_{in}^0, Q_{ik}^0,

$$\frac{dR_{in}^0}{d\alpha} = \eta \langle \delta_i y_n \rangle, \tag{9}$$

$$\frac{dQ_{ik}^0}{d\alpha} = \eta \left\langle \delta_i x_k + \delta_k x_i \right\rangle + \eta^2 \left\langle \delta_i \delta_k \right\rangle . \tag{10}$$

For given initial overlap conditions, (9,10) are integrated to find the mean dynamical behaviour of a student learning a teacher with an arbitrary numbers of hidden units[2] (see fig.(1a)). 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.(2a)). Large deviations from the mean dynamics can exist in this symmetric phase, as a small perturbation from symmetry can determine which student hidden unit will specialise on which teacher hidden unit[1]. We rewrite (7,8) in the general form

$$a^{\mu+1} - a^{\mu} = \frac{\eta}{N} \left(F_a + \eta G_a \right), \tag{11}$$

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 ansaetze 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,[1]

$$F_a = F_a^0 + \triangle F_a + \frac{1}{N} F_a^1, \qquad a = a^0 + \sqrt{\frac{\eta}{N}} \triangle a + \frac{\eta}{N} a^1, \tag{12}$$

where $\langle \triangle F_a \rangle = \langle \triangle a \rangle = 0$. The update rule ansatz is motivated by observing that the activations have variance $\mathcal{O}(1)$ which, iterated through (11) yield overlap variances of $\mathcal{O}(N^{-1})$. Terms of the form, $\triangle a$ represent *dynamic* corrections that arise due to the random examples, and 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,

$$\frac{da^0}{d\tilde{\alpha}} = F_a^0, \tag{13}$$

where F_a^0 is the update rule F_a averaged over the input distribution, and the rescaled learning rate is given by

$$\tilde{\alpha} = \eta \alpha. \tag{14}$$

Substituting (12) in (11) and averaging over the input distribution, we derive a set of coupled differential equations for the (scaled) covariances $\langle \triangle a \triangle b \rangle$, and static

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

corrections a^1,

$$\frac{d\langle \triangle a \triangle b\rangle}{d\tilde{\alpha}} = \sum_c \langle \triangle a \triangle c\rangle \frac{\partial F_a^0}{\partial c^0} + \sum_c \langle \triangle b \triangle c\rangle \frac{\partial F_b^0}{\partial c^0} + \langle \triangle F_a \triangle F_b\rangle \tag{15}$$

$$\frac{1}{2}\frac{d^2 a^0}{d\tilde{\alpha}^2} + \frac{da^1}{d\tilde{\alpha}} = \sum_b b^1 \frac{\partial F_a^0}{\partial b^0} + \frac{1}{2}\sum_{bc} \langle \triangle b \triangle c\rangle \frac{\partial^2 F_a^0}{\partial b^0 \partial c^0} + G_a^1. \tag{16}$$

Summations are over all overlap parameters, $\{Q_{ij}, R_{in} | i,j = 1..K, n = 1..M\}$. 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 equations (13,15) are then integrated to find the evolution of the covariances, $cov(a,b) = (\eta/N)\langle \triangle a \triangle b\rangle$, and the corrections to the thermodynamic average values, $(\eta/N)a^1$. The average finite size correction to the generalisation error is given by

$$\epsilon_g = \epsilon_g^0 + \frac{\eta}{N}\epsilon_g^1, \tag{17}$$

where,

$$\epsilon_g^1 = \sum_a a^1 \frac{\partial e_g^0}{\partial a} + \frac{1}{2}\sum_{ab} \langle \triangle a \triangle b\rangle \frac{\partial^2 \epsilon_g^0}{\partial a^0 \partial b^0}. \tag{18}$$

These results enable the calculation of finite size effects for an arbitrary 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 on 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 $\mathcal{O}(N^{-1})$ correction. In fig.(1a) there is no significant symmetric phase, and the finite size corrections (fig.(1b)) 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.(2a)) there is a very definite symmetric phase, for which a finite size correction of less than 10% (fig.(2b)) 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 specialising its hidden units. 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 $\langle \triangle a \triangle b\rangle$, we found that there is 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 anti-correlated, whilst there is minimal correlation between the Q_{11} and Q_{12} components. This corresponds

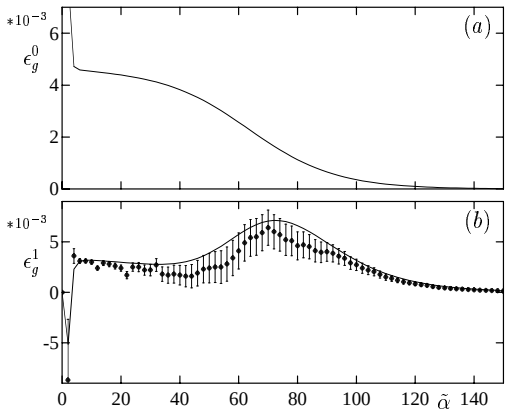

Fig. 1.

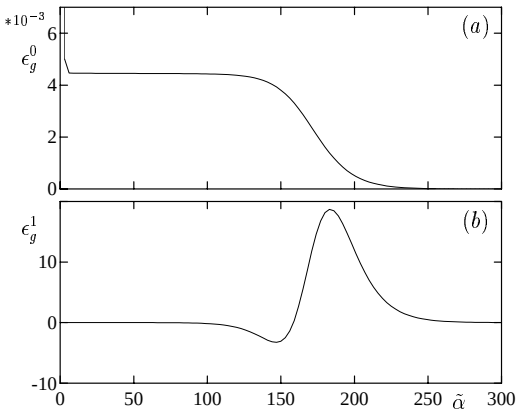

Fig. 2.

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, ϵ_g^0. (b) $\mathcal{O}(N^{-1})$ correction to the generalisation error , ϵ_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 ϵ_g^0. (b) $\mathcal{O}(N^{-1})$ correction to the generalisation error, ϵ_g^1.

well with predictions from perturbation analysis[2]. 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 the specialisation has occured, there is a growth in the anti-correlation between the student length and its overlap with other students. The asymptotic values of these correlations are in agreement with the convergence fixed point, $R^2 = Q = 1$.

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

$$\epsilon^{\dagger} = \frac{1}{2} \sum_{j=1}^{K-1} h\left(Q_{j+1j+1} - Q_{jj}\right), \tag{19}$$

where $h(x)$ approximates the step function,

$$h(x) = \frac{1}{2}\left(1 + erf\left(\frac{\beta}{\sqrt{2}}x\right)\right). \tag{20}$$

This straightforward modification involves the addition of a gaussian term in the student weight lengths to the weight update rule (4). In fig.(4), we show the over-

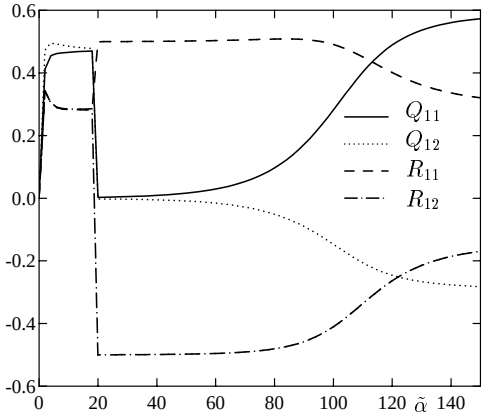

Fig. 3.

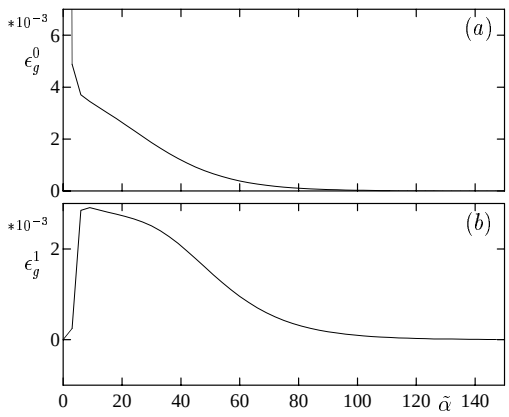

Fig. 4.

Fig. 3. - (a) The normalised components of the principal eigenvector for the 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, ϵ_g^0 . (b) $\mathcal{O}(N^{-1})$ correction to the generalisation error, ϵ_g^1.

lap parameters and their fluctuations for β=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 specialize. Also, the initial convergence to the optimal values is much faster. As there is no symmetric phase, the finite size corrections are much reduced and are largest around the initial value of $\tilde{\alpha}$ where the overlap parameters are most symmetric, decreasing rapidly due to the driving force away from this near-symmetric region. For the case in which the teacher weights are equal, the constraint (19) prevents the student from converging optimally. A naive scheme to prevent this 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. We conjecture that such symmetry breaking is potentially of great benefit in the practical field of neural network training.

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

References

[1] M. Biehl and H. Schwarze. *Journal of Physics A*, 28:643–656, 1995.
[2] D. Saad and S .Solla. *Physical Review Letters*, 74(21):4337-4340, 1995.
[3] P. Sollich. *Journal of Physics A*, 27:7771–7784, 1994.