Dynamics of Supervised Learning with Restricted Training Sets

A.C.C. Coolen
Dept of Mathematics
King's College London
Strand, London WC2R 2LS, UK
tcoolen@mth.kcl.ac.uk

D. Saad
Neural Computing Research Group
Aston University
Birmingham B4 7ET, UK
saadd@aston.ac.uk

Abstract

We study the dynamics of supervised learning in layered neural networks, in the regime where the size p of the training set is proportional to the number N of inputs. Here the local fields are no longer described by Gaussian distributions. We use dynamical replica theory to predict the evolution of macroscopic observables, including the relevant error measures, incorporating the old formalism in the limit $p/N \to \infty$.

1 INTRODUCTION

Much progress has been made in solving the dynamics of supervised learning in layered neural networks, using the strategy of statistical mechanics: by deriving closed laws for the evolution of suitably chosen macroscopic observables (order parameters) in the limit of an infinite system size [1, 2, 3, 4]. For a recent review and guide to references see e.g. [5]. The main successful procedure developed so far is built on the following cornerstones:

• *The task to be learned is defined by a 'teacher', which is itself a neural network.* This induces a natural set of order parameters (mutual weight vector overlaps between the teacher and the trained, 'student', network).

• *The number of network inputs is infinitely large.* This ensures that fluctuations in the order parameters will vanish, and enables usage of the central limit theorem.

• *The number of 'hidden' neurons is finite,* in both teacher and student, ensuring a finite number of order parameters and an insignificant cumulative impact of the fluctuations.

• *The size of the training set is much larger than the number of updates.* Each example presented is now different from the previous ones, so that the local fields will have Gaussian distributions, leading to closure of the dynamic equations.

In this paper we study the dynamics of learning in layered networks with *restricted* training sets, where the number p of examples scales linearly with the number N of inputs. Individual examples will now re-appear during the learning process as soon as the number of weight updates made is of the order of p. Correlations will develop between the weights

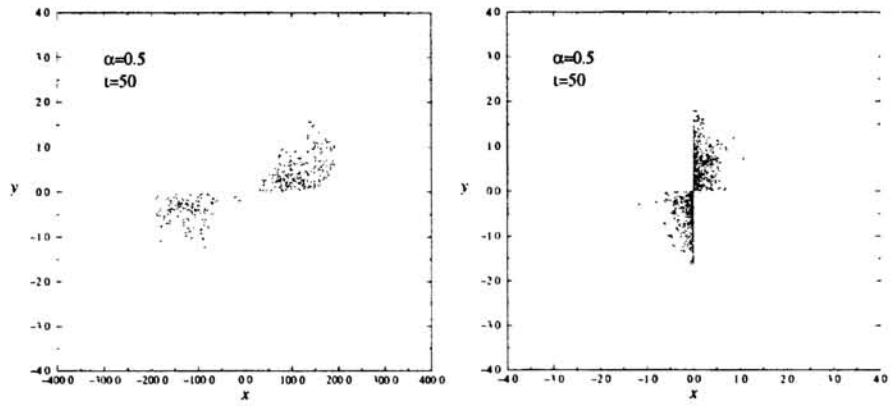

Figure 1: Student and teacher fields (x, y) (see text) observed during numerical simulations of on-line learning (learning rate $\eta = 1$) in a perceptron of size $N = 10,000$ at $t = 50$, using examples from a training set of size $p = \frac{1}{2}N$. Left: Hebbian learning. Right: AdaTron learning [5]. Both distributions are clearly non-Gaussian.

and the training set examples and the student's local fields (activations) will be described by non-Gaussian distributions (see e.g. Figure 1). This leads to a breakdown of the standard formalism: the field distributions are no longer characterized by a few moments, and the macroscopic laws must now be averaged over realizations of the training set. The first rigorous study of the dynamics of learning with restricted training sets in non-linear networks, via generating functionals [6], was carried out for networks with binary weights. Here we use dynamical replica theory (see e.g. [7]) to predict the evolution of macroscopic observables for finite α, incorporating the old formalism as a special case ($\alpha = p/N \to \infty$). For simplicity we restrict ourselves to single-layer systems and noise-free teachers.

2 FROM MICROSCOPIC TO MACROSCOPIC LAWS

A 'student' perceptron operates a rule which is parametrised by the weight vector $\boldsymbol{J} \in \Re^N$:

$$S : \{-1, 1\}^N \to \{-1, 1\} \qquad S(\boldsymbol{\xi}) = \text{sgn}\,[\boldsymbol{J} \cdot \boldsymbol{\xi}] \equiv \text{sgn}\,[x] \tag{1}$$

It tries to emulate a teacher perceptron which operates a similar rule, characterized by a (fixed) weight vector $\boldsymbol{B} \in \Re^N$. The student modifies its weight vector $\boldsymbol{J}$ iteratively, using examples of input vectors $\boldsymbol{\xi}$ which are drawn at random from a fixed (randomly composed) training set $\tilde{D} = \{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p\} \subset D = \{-1, 1\}^N$, of size $p = \alpha N$ with $\alpha > 0$, and the corresponding values of the teacher outputs $T(\boldsymbol{\xi}) = \text{sgn}[\boldsymbol{B} \cdot \boldsymbol{\xi}] \equiv \text{sgn}\,[y]$. Averages over the training set $\tilde{D}$ and over the full set D will be denoted as $\langle \Phi(\boldsymbol{\xi}) \rangle_{\tilde{D}}$ and $\langle \Phi(\boldsymbol{\xi}) \rangle_D$, respectively. We will analyze the following two classes of learning rules:

$$\begin{aligned} \text{on-line}: \quad & \boldsymbol{J}(m+1) = \boldsymbol{J}(m) + \tfrac{\eta}{N}\, \boldsymbol{\xi}(m)\, \mathcal{G}\,[\boldsymbol{J}(m) \cdot \boldsymbol{\xi}(m), \boldsymbol{B} \cdot \boldsymbol{\xi}(m)] \\ \text{batch}: \quad & \boldsymbol{J}(m+1) = \boldsymbol{J}(m) + \tfrac{\eta}{N}\, \langle \boldsymbol{\xi}\, \mathcal{G}\,[\boldsymbol{J}(m) \cdot \boldsymbol{\xi}, \boldsymbol{B} \cdot \boldsymbol{\xi}] \rangle_{\tilde{D}} \end{aligned} \tag{2}$$

In on-line learning one draws at each step m a question $\boldsymbol{\xi}(m)$ at random from the training set, the dynamics is a stochastic process; in batch learning one iterates a deterministic map. Our key dynamical observables are the training- and generalization errors, defined as

$$E_{\rm t}(\boldsymbol{J}) = \langle \theta[-(\boldsymbol{J} \cdot \boldsymbol{\xi})(\boldsymbol{B} \cdot \boldsymbol{\xi})] \rangle_{\tilde{D}} \qquad E_{\rm g}(\boldsymbol{J}) = \langle \theta[-(\boldsymbol{J} \cdot \boldsymbol{\xi})(\boldsymbol{B} \cdot \boldsymbol{\xi})] \rangle_D \tag{3}$$

Only if the training set $\tilde{D}$ is sufficiently large, and if there are no correlations between $\boldsymbol{J}$ and the training set examples, will these two errors be identical. We now turn to *macroscopic* observables $\boldsymbol{\Omega}[\boldsymbol{J}] = (\Omega_1[\boldsymbol{J}], \ldots, \Omega_k[\boldsymbol{J}])$. For $N \to \infty$ (with finite times $t = m/N$

and with finite k), and if our observables are of a so-called mean-field type, their associated macroscopic distribution $P_t(\Omega)$ is found to obey a Fokker-Planck type equation, with flow- and diffusion terms that depend on whether on-line or batch learning is used. We now choose a *specific* set of observables $\Omega[\boldsymbol{J}]$, taylored to the present problem:

$$Q[\boldsymbol{J}] = \boldsymbol{J}^2, \qquad R[\boldsymbol{J}] = \boldsymbol{J}\cdot\boldsymbol{B}, \qquad P[x,y;\boldsymbol{J}] = \langle \delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\,\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde{D}} \tag{4}$$

This choice is motivated as follows: (i) in order to incorporate the old formalism we need $Q[\boldsymbol{J}]$ and $R[\boldsymbol{J}]$, (ii) the training error involves field statistics calculated over the training set, as given by $P[x,y;\boldsymbol{J}]$, and (iii) for $\alpha < \infty$ one cannot expect closed equations for a finite number of order parameters, the present choice effectively represents an infinite number. We will assume the number of arguments (x,y) for which $P[x,y;\boldsymbol{J}]$ is evaluated to go to infinity *after* the limit $N\to\infty$ has been taken. This eliminates technical subtleties and allows us to show that in the Fokker-Planck equation all diffusion terms vanish as $N\to\infty$. The latter thereby reduces to a Liouville equation, describing deterministic evolution of our macroscopic observables. For on-line learning one arrives at

$$\frac{d}{dt}Q = 2\eta\int dxdy\; P[x,y]\,x\,\mathcal{G}[x;y] + \eta^2\int dxdy\; P[x,y]\,\mathcal{G}^2[x;y] \tag{5}$$

$$\frac{d}{dt}R = \eta\int dxdy\; P[x,y]\,y\,\mathcal{G}[x;y] \tag{6}$$

$$\begin{aligned}\frac{\partial}{\partial t}P[x,y] = {}& \frac{1}{\alpha}\left[\int dx' P[x',y]\delta[x-x'-\eta\mathcal{G}[x',y]] - P[x,y]\right] \\ & -\eta\frac{\partial}{\partial x}\int dx'dy'\;\mathcal{G}[x',y']\,\mathcal{A}[x,y;x',y'] \\ & +\frac{1}{2}\eta^2\int dx'dy'\;P[x',y']\,\mathcal{G}^2[x',y']\,\frac{\partial^2}{\partial x^2}P[x,y]\end{aligned} \tag{7}$$

Expansion of these equations in powers of η, and retaining only the terms linear in η, gives the corresponding equations describing batch learning. The complexity of the problem is fully concentrated in a Green's function $\mathcal{A}[x,y;x',y']$, which is defined as

$$\mathcal{A}[x,y;x',y'] = \lim_{N\to\infty}\langle\langle([1-\delta_{\boldsymbol{\xi}\boldsymbol{\xi}'}]\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\,\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}](\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\,\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\rangle_{\tilde{D}}\rangle_{\tilde{D}}\rangle_{\mathrm{QRP};t}$$

It involves a *sub-shell* average, in which $p_t(\boldsymbol{J})$ is the weight probability density at time t:

$$\langle K[\boldsymbol{J}]\rangle_{\mathrm{QRP};t} = \frac{\int d\boldsymbol{J}\;K[\boldsymbol{J}]\,p_t(\boldsymbol{J})\delta[Q-Q[\boldsymbol{J}]]\delta[R-R[\boldsymbol{J}]]\prod_{xy}\delta[P[x,y]-P[x,y;\boldsymbol{J}]]}{\int d\boldsymbol{J}\;p_t(\boldsymbol{J})\delta[Q-Q[\boldsymbol{J}]]\delta[R-R[\boldsymbol{J}]]\prod_{xy}\delta[P[x,y]-P[x,y;\boldsymbol{J}]]}$$

where the sub-shells are defined with respect to the order parameters. The solution of (5,6,7) can be used to generate the errors of (3):

$$E_{\mathrm{t}} = \int dxdy\; P[x,y]\theta[-xy] \qquad\qquad E_{\mathrm{g}} = \frac{1}{\pi}\arccos[R/\sqrt{Q}] \tag{8}$$

3 CLOSURE VIA DYNAMICAL REPLICA THEORY

So far our analysis is still exact. We now close the macroscopic laws (5,6,7) by making, for $N\to\infty$, the two key assumptions underlying dynamical replica theory [7]:

(i) Our macroscopic observables $\{Q,R,P\}$ obey *closed* dynamic equations.

(ii) These equations are self-averaging with respect to the realisation of $\tilde{D}$.

(i) implies that probability variations within the $\{Q,R,P\}$ subshells are either absent or irrelevant to the evolution of $\{Q,R,P\}$. We may thus make the simplest choice for $p_t(\boldsymbol{J})$:

$$p_t(\boldsymbol{J}) \;\to\; \overline{p}(\boldsymbol{J}) \sim \delta[Q-Q[\boldsymbol{J}]]\,\delta[R-R[\boldsymbol{J}]]\prod_{xy}\delta[P[x,y]-P[x,y;\boldsymbol{J}]] \tag{9}$$

$\overline{p}(\boldsymbol{J})$ depends on time implicitly, via the order parameters $\{Q, R, P\}$. The procedure (9) leads to exact laws if our observables $\{Q, R, P\}$ indeed obey closed equations for $N \to \infty$. It gives an approximation if they don't. (ii) allows us to average the macroscopic laws over all training sets; it is observed in numerical simulations, and can probably be proven using the formalism of [6]. Our assumptions result in the closure of (5,6,7), since now $\mathcal{A}[\ldots]$ is expressed fully in terms of $\{Q, R, P\}$. The final ingredient of dynamical replica theory is the realization that averaging fractions is simplified with the replica identity [8]

$$\left\langle \frac{\int d\boldsymbol{J}\, W[\boldsymbol{J}, z] G[\boldsymbol{J}, z]}{\int d\boldsymbol{J}\, W[\boldsymbol{J}, z]} \right\rangle_z = \lim_{n\to 0} \int d\boldsymbol{J}^1 \cdots d\boldsymbol{J}^n \, \langle G[\boldsymbol{J}^1, z] \prod_{\alpha=1}^{n} W[\boldsymbol{J}^\alpha, z] \rangle_z$$

What remains is to perform integrations. One finds that $P[x, y] = P[x|y]P[y]$ with $P[y] = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}y^2}$. Upon introducing the short-hands $Dy = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}y^2} dy$ and $\langle f(x, y)\rangle = \int Dy dx\, P[x|y] f(x, y)$ we can write the resulting macroscopic laws as follows:

$$\frac{d}{dt} Q = 2\eta V + \eta^2 Z \qquad \frac{d}{dt} R = \eta W \tag{10}$$

$$\frac{\partial}{\partial t} P[x|y] = \frac{1}{\alpha} \int dx' P[x'|y] \{\delta[x - x' - \eta G[x', y]] - \delta[x - x']\} + \frac{1}{2}\eta^2 Z \frac{\partial^2}{\partial x^2} P[x|y]$$
$$- \eta \frac{\partial}{\partial x} \{P[x|y] [U(x - Ry) + Wy + [V - RW - (Q - R^2)U]\Phi[x, y]]\} \tag{11}$$

with

$$U = \langle \Phi[x, y] \mathcal{G}[x, y] \rangle, \quad V = \langle x \mathcal{G}[x, y] \rangle, \quad W = \langle y \mathcal{G}[x, y] \rangle, \quad Z = \langle \mathcal{G}^2[x, y] \rangle$$

As before the batch equations follow upon expanding in η and retaining only the linear terms. Finding the function $\Phi[x, y]$ (in replica symmetric ansatz) requires solving a saddle-point problem for a scalar observable q and a function $M[x|y]$. Upon introducing

$$B = \frac{\sqrt{qQ - R^2}}{Q(1 - q)} \qquad \langle f[x, y, z] \rangle_\star = \frac{\int dx\, M[x|y] e^{Bxz} f[x, y, z]}{\int dx\, M[x|y] e^{Bxz}}$$

(with $\int dx\, M[x|y] = 1$ for all y) the saddle-point equations acquire the form

$$\text{for all } X, y: \quad P[X|y] = \int Dz\, \langle \delta[X - x] \rangle_\star$$

$$\langle (x - Ry)^2 \rangle + (qQ - R^2)[1 - \frac{1}{\alpha}] = [Q(1 + q) - 2R^2] \langle x \Phi[x, y] \rangle$$

The solution $M[x|y]$ of the functional saddle-point equation, given a value for q in the physical range $q \in [R^2/Q, 1]$, is unique [9]. The function $\Phi[x, y]$ is then given by

$$\Phi[X, y] = \left\{ \sqrt{qQ - R^2} P[X|y] \right\}^{-1} \int Dz\, z \langle \delta[X - x] \rangle_\star \tag{12}$$

4 THE LIMIT $\alpha \to \infty$

For consistency we show that our theory reduces to the simple (Q, R) formalism of infinite training sets in the limit $\alpha \to \infty$. Upon making the ansatz

$$P[x|y] = [2\pi(Q - R^2)]^{-\frac{1}{2}} \, e^{-\frac{1}{2}[x - Ry]^2/(Q - R^2)}$$

one finds that the saddle-point equations are simultaneously and uniquely solved by

$$M[x|y] = P[x|y], \qquad q = R^2/Q$$

and $\Phi[x, y]$ reduces to

$$\Phi[x, y] = (x - Ry)/(Q - R^2)$$

Insertion of our ansatz into equation (11), followed by rearranging of terms and usage of the above expression for $\Phi[x, y]$, shows that this equation is satisfied. Thus from our general theory we indeed recover for $\alpha \to \infty$ the standard theory for infinite training sets.

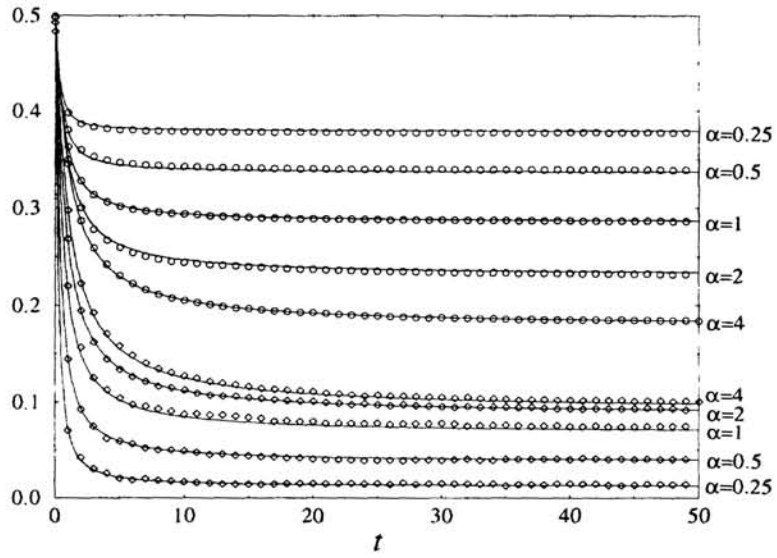

Figure 2: Simulation results for on-line Hebbian learning (system size $N = 10.000$) versus an approximate solution of the equations generated by dynamical replica theory (see main text), for $\alpha \in \{0.25, 0.5, 1.0, 2.0, 4.0\}$. Upper five curves: E_g as functions of time. Lower five curves: E_t as functions of time. Circles: simulation results for E_g; diamonds: simulation results for E_t. Solid lines: the corresponding theoretical predictions.

5 BENCHMARK TESTS: HEBBIAN LEARNING

Batch Hebbian Learning

For the Hebbian rule, where $\mathcal{G}[x, y] = \text{sgn}(y)$, one can calculate our order parameters exactly at any time, even for $\alpha < \infty$ [10], which provides an excellent benchmark for general theories such as ours. For batch execution all integrations in our present theory can be done and all equations solved explicitly, and our theory is found to predict the following:

$$R = R_0 + \eta t\sqrt{\frac{2}{\pi}} \qquad Q = Q_0 + 2\eta t R_0\sqrt{\frac{2}{\pi}} + \eta^2 t^2\left[\frac{2}{\pi}+\frac{1}{\alpha}\right] \qquad q = \frac{\alpha R^2 + \eta^2 t^2}{\alpha Q} \tag{13}$$

$$P[x|y] = \frac{e^{-\frac{1}{2}[x - Ry - (\eta t/\alpha)\,\text{sgn}(y)]^2/(Q-R^2)}}{\sqrt{2\pi(Q-R^2)}} \tag{14}$$

$$E_g = \frac{1}{\pi}\arccos\left[\frac{R}{\sqrt{Q}}\right] \qquad E_t = \frac{1}{2} - \frac{1}{2}\int Dy\ \text{erf}\left[\frac{|y|R+\eta t/\alpha}{\sqrt{2(Q-R^2)}}\right] \tag{15}$$

Comparison with the exact solution, calculated along the lines of [10] (where this was done for on-line Hebbian learning) shows that the above expressions are all rigorously exact.

On-Line Hebbian Learning

For on-line execution we cannot (yet) solve the functional saddle-point equation analytically. However, some explicit analytical predictions can still be extracted [9]:

$$R = R_0 + \eta t\sqrt{\frac{2}{\pi}} \qquad Q = Q_0 + 2\eta t R_0\sqrt{\frac{2}{\pi}} + \eta^2 t + \eta^2 t^2\left[\frac{2}{\pi}+\frac{1}{\alpha}\right] \tag{16}$$

$$\int dx\ x P[x|y] = Ry + (\eta t/\alpha)\,\text{sgn}(y) \tag{17}$$

$$P[x|y] \sim \left[\frac{\alpha}{2\pi\eta^2 t^2}\right]^{\frac{1}{2}} \exp\left[-\frac{\alpha(x - Ry - (\eta t/\alpha)\,\text{sgn}(y))^2}{2\eta^2 t^2}\right] \qquad (t \to \infty) \tag{18}$$

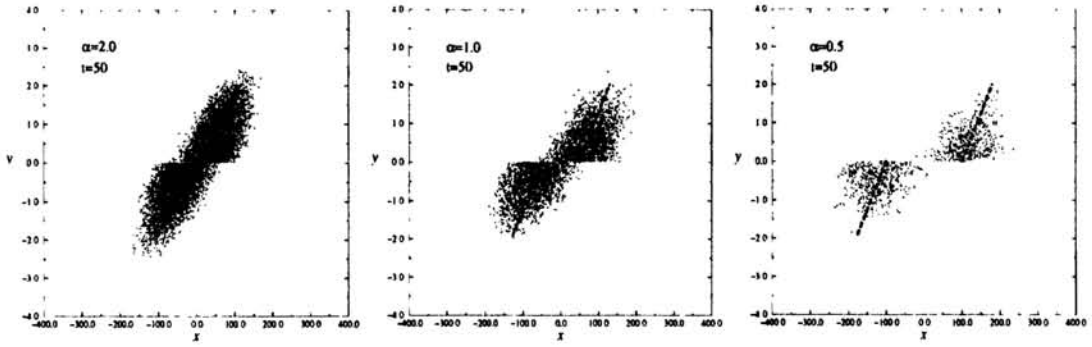

Figure 3: Simulation results for on-line Hebbian learning ($N = 10,000$) versus dynamical replica theory, for $\alpha \in \{2.0, 1.0, 0.5\}$. Dots: local fields $(x, y) = (J \cdot \xi, B \cdot \xi)$ (calculated for examples in the training set), at time $t = 50$. Dashed lines: conditional average of student field x as a function of y, as predicted by the theory, $\overline{x}(y) = Ry + (\eta t/\alpha)\ \mathrm{sgn}(y)$.

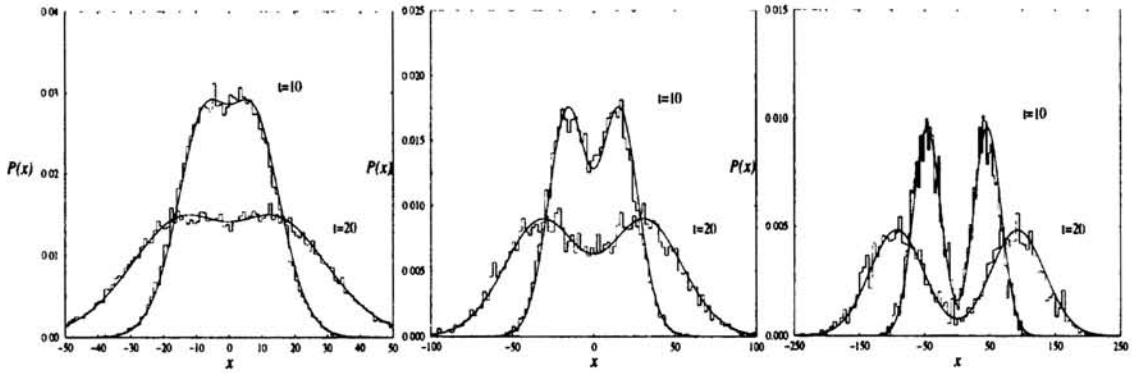

Figure 4: Simulations of Hebbian on-line learning with $N = 10,000$. Histograms: student field distributions measured at $t = 10$ and $t = 20$. Lines: theoretical predictions for student field distributions (using the approximate solution of the diffusion equation, see main text), for $\alpha = 4$ (left), $\alpha = 1$ (middle), $\alpha = 0.25$ (right).

Comparison with the exact result of [10] shows that the above expressions (16,17,18), and therefore also that of E_g at any time, are all rigorously exact.

At intermediate times it turns out that a good approximation of the solution of our dynamic equations for on-line Hebbian learning (exact for $t \ll \alpha$ and for $t \to \infty$) is given by

$$P[x|y] = \frac{e^{-\frac{1}{2}[x - Ry - (\eta t/\alpha)\ \mathrm{sgn}(y)]^2/(Q - R^2 + \eta^2 t/\alpha)}}{\sqrt{2\pi(Q - R^2 + \eta^2 t/\alpha)}} \tag{19}$$

$$E_g = \frac{1}{\pi} \arccos\left[\frac{R}{\sqrt{Q}}\right] \qquad E_t = \frac{1}{2} - \frac{1}{2}\int Dy\ \mathrm{erf}\left[\frac{|y|R + \eta t/\alpha}{\sqrt{2(Q - R^2 - \eta^2 t/\alpha)}}\right] \tag{20}$$

In Figure 2 we compare the approximate predictions (20) with the results obtained from numerical simulations ($N = 10,000$, $Q_0 = 1$, $R_0 = 0$, $\eta = 1$). All curves show excellent agreement between theory and experiment. We also compare the theoretical predictions for the distribution $P[x|y]$ with the results of numerical simulations. This is done in Figure 3 where we show the fields as observed at $t = 50$ in simulations (same parameters as in Figure 2) of on-line Hebbian learning, for three different values of α. In the same figure we draw (dashed lines) the theoretical prediction for the y-dependent average (17) of the conditional x-distribution $P[x|y]$. Finally we compare the student field distribution $P[x] =$

$\int Dy\ P[x|y]$ according to (19) with that observed in numerical simulations, see Figure 4. The agreement is again excellent (note: here the learning process has almost equilibrated).

6 DISCUSSION

In this paper we have shown how the formalism of dynamical replica theory [7] can be used successfully to build a general theory with which to predict the evolution of the relevant macroscopic performance measures, including the training- and generalisation errors, for supervised (on-line and batch) learning in layered neural networks with randomly composed but restricted training sets (i.e. for finite $\alpha = p/N$). Here the student fields are no longer described by Gaussian distributions, and the more familiar statistical mechanical formalism breaks down. For simplicity and transparency we have restricted ourselves to single-layer systems and realizable tasks. In our approach the joint distribution $P[x, y]$ for student and teacher fields is itself taken to be a dynamical order parameter, in addition to the conventional observables Q and R. From the order parameter set $\{Q, R, P\}$, in turn, we derive both the generalization error $E_{\rm g}$ and the training error $E_{\rm t}$. Following the prescriptions of dynamical replica theory one finds a diffusion equation for $P[x, y]$, which we have evaluated by making the replica-symmetric ansatz in the saddle-point equations. This equation has Gaussian solutions only for $\alpha \to \infty$; in the latter case we indeed recover correctly from our theory the more familiar formalism of infinite training sets, with closed equations for Q and R only. For finite α our theory is by construction exact if for $N \to \infty$ the dynamical order parameters $\{Q, R, P\}$ obey closed, deterministic equations, which are self-averaging (i.e. independent of the microscopic realization of the training set). If this is not the case, our theory is an approximation.

We have worked out our general equations explicitly for the special case of Hebbian learning, where the existence of an exact solution [10], derived from the microscopic equations (for finite α), allows us to perform a critical test of our theory. Our theory is found to be fully exact for batch Hebbian learning. For on-line Hebbian learning full exactness is difficult to determine, but exactness can be establised at least for (i) $t \to \infty$, (ii) the predictions for Q, R, $E_{\rm g}$ and $\overline{x}(y) = \int dx\ xP[x|y]$ at any time. A simple approximate solution of our equations already shows excellent agreement between theory and experiment. The present study clearly represents only a first step, and many extensions, applications and generalizations are currently under way. More specifically, we study alternative learning rules as well as the extension of this work to the case of noisy data and of soft committee machines.

References

[1] Kinzel W. and Rujan P. (1990), *Europhys. Lett.* **13**, 473

[2] Kinouchi O. and Caticha N. (1992), *J. Phys. A: Math. Gen.* **25**, 6243

[3] Biehl M. and Schwarze H. (1992), *Europhys. Lett.* **20**, 733
Biehl M. and Schwarze H. (1995), *J. Phys. A: Math. Gen.* **28**, 643

[4] Saad D. and Solla S. (1995), *Phys. Rev. Lett.* **74**, 4337

[5] Mace C.W.H. and Coolen A.C.C (1998), *Statistics and Computing* **8**, 55

[6] Horner H. (1992a), *Z. Phys. B* **86**, 291
Horner H. (1992b), *Z. Phys. B* **87**, 371

[7] Coolen A.C.C., Laughton S.N. and Sherrington D. (1996), *Phys. Rev. B* **53**, 8184

[8] Mézard M., Parisi G. and Virasoro M.A. (1987), *Spin-Glass Theory and Beyond* (Singapore: World Scientific)

[9] Coolen A.C.C. and Saad D. (1998), in preparation.

[10] Rae H.C., Sollich P. and Coolen A.C.C. (1998), these proceedings