On-Line Learning with Restricted Training Sets: Exact Solution as Benchmark for General Theories

H.C. Rae
hamish.rae@kcl.ac.uk

P. Sollich
psollich@mth.kcl.ac.uk

A.C.C. Coolen
tcoolen@mth.kcl.ac.uk

Department of Mathematics
King's College London
The Strand
London WC2R 2LS, UK

Abstract

We solve the dynamics of on-line Hebbian learning in perceptrons exactly, for the regime where the size of the training set scales linearly with the number of inputs. We consider both noiseless and noisy teachers. Our calculation cannot be extended to non-Hebbian rules, but the solution provides a nice benchmark to test more general and advanced theories for solving the dynamics of learning with restricted training sets.

1 Introduction

Considerable progress has been made in understanding the dynamics of supervised learning in layered neural networks through the application of the methods of statistical mechanics. A recent review of work in this field is contained in [1]. For the most part, such theories have concentrated on systems where the training set is much larger than the number of updates. In such circumstances the probability that a question will be repeated during the training process is negligible and it is possible to assume for large networks, via the central limit theorem, that the local field distribution is Gaussian. In this paper we consider *restricted training sets*; we suppose that the size of the training set scales linearly with N, the number of inputs. The probability that a question will reappear during the training process is no longer negligible, the assumption that the local fields have Gaussian distributions is not tenable, and it is clear that correlations will develop between the weights and the questions in the training set as training progresses. In fact, the non-Gaussian character of the local fields should be a *prediction* of any satisfactory theory of learning with restricted training sets, as this is clearly demanded by numerical simulations. Several authors [2, 3, 4, 5, 6, 7] have discussed learning with restricted training sets but a general theory is difficult. A simple model of learning with restricted training sets which can be solved *exactly* is therefore particularly attractive and provides a yardstick against which more difficult and sophisticated general theories can, in due course, be tested and compared. We show how this can be accomplished for on-line Hebbian learning in perceptrons with restricted training sets and we obtain exact solutions for the generalisation error and the training error for a class of noisy teachers and students with arbitrary weight decay. Our theory is in excellent agreement with numerical simulations and our prediction of the probability density of the student field is a striking confirmation of them, making it clear that we are indeed dealing with local fields which are non-Gaussian.

2 Definitions

We study on-line learning in a student perceptron S, which tries to perform a task defined by a teacher perceptron characterised by a fixed weight vector $\boldsymbol{B}^* \in \Re^N$. We assume, however, that the teacher is noisy and that the *actual* teacher output T and the corresponding student response S are given by

$$T : \{-1,1\}^N \to \{-1,1\} \qquad T(\boldsymbol{\xi}) = \operatorname{sgn}[\boldsymbol{B}\cdot\boldsymbol{\xi}],$$
$$S : \{-1,1\}^N \to \{-1,1\} \qquad S(\boldsymbol{\xi}) = \operatorname{sgn}[\boldsymbol{J}\cdot\boldsymbol{\xi}],$$

where the vector $\boldsymbol{B}$ is drawn *independently* of $\boldsymbol{\xi}$ with probability $p(\boldsymbol{B})$ which may depend explicitly on the correct teacher vector $\boldsymbol{B}^*$. Of particular interest are the following two choices, described in literature as output noise and Gaussian input noise, respectively:

$$p(\boldsymbol{B}) = \lambda\ \delta(\boldsymbol{B}+\boldsymbol{B}^*) + (1-\lambda)\ \delta(\boldsymbol{B}-\boldsymbol{B}^*) \tag{1}$$

where $\lambda \geq 0$ represents the probability that the teacher output is incorrect, and

$$p(\boldsymbol{B}) = \left[\frac{N}{2\pi\Sigma^2}\right]^{\frac{N}{2}} e^{-\frac{N}{2}(\boldsymbol{B}-\boldsymbol{B}^*)^2/\Sigma^2}. \tag{2}$$

The variance Σ^2/N has been chosen so as to achieve appropriate scaling for $N \to \infty$.

Our learning rule will be the on-line Hebbian rule, i.e.

$$\boldsymbol{J}(\ell{+}1) = \big(1-\frac{\gamma}{N}\big)\boldsymbol{J}(\ell) + \frac{\eta}{N}\ \boldsymbol{\xi}(\ell)\ \ \operatorname{sgn}[\boldsymbol{B}(\ell)\cdot\boldsymbol{\xi}(\ell)] \tag{3}$$

where the non-negative parameters γ and η are the decay rate and the learning rate, respectively. At each iteration step ℓ an input vector $\boldsymbol{\xi}(\ell)$ is picked at random from a training set consisting of $p = \alpha N$ randomly drawn vectors $\boldsymbol{\xi}^\mu \in \{-1,1\}^N$, $\mu = 1,\dots p$. This set remains unchanged during the learning dynamics. At the same time the teacher selects at random, and independently of $\boldsymbol{\xi}(\ell)$, the vector $\boldsymbol{B}(\ell)$, according to the probability distribution $p(\boldsymbol{B})$. Iterating equation (3) gives

$$\boldsymbol{J}(m) = \left(1-\frac{\gamma}{N}\right)^m \boldsymbol{J}_0 \ + \ \frac{\eta}{N}\sum_{\ell=0}^{m-1}\left(1-\frac{\gamma}{N}\right)^{m-\ell-1} \boldsymbol{\xi}(\ell)\ \operatorname{sgn}[\boldsymbol{B}(\ell)\cdot\boldsymbol{\xi}(\ell)] \tag{4}$$

We assume that the (noisy) teacher output is *consistent* in the sense that if a question $\boldsymbol{\xi}$ reappears at some stage during the training process the teacher makes the same choice of $\boldsymbol{B}$ in both cases, i.e. if $\boldsymbol{\xi}(\ell) = \boldsymbol{\xi}(\ell')$ then also $\boldsymbol{B}(\ell) = \boldsymbol{B}(\ell')$. This consistency allows us to define a generalised training set $\tilde{D}$ by including with the p

questions the corresponding teacher vectors:

$$\tilde{D} = \{(\boldsymbol{\xi}^1, \boldsymbol{B}^1), \ldots, (\boldsymbol{\xi}^p, \boldsymbol{B}^p)\}$$

There are two sources of randomness in this problem. First of all there is the random realisation of the 'path' $\Omega = \{(\boldsymbol{\xi}(0), \boldsymbol{B}(0)), (\boldsymbol{\xi}(1), \boldsymbol{B}(1)), \ldots, (\boldsymbol{\xi}(\ell), \boldsymbol{B}(\ell)), \ldots\}$. This is simply the randomness of the stochastic process that gives the evolution of the vector $\boldsymbol{J}$. Averages over this process will be denoted as $\langle \ldots \rangle$. Secondly there is the randomness in the composition of the training set. We will write averages over all training sets as $\langle \ldots \rangle_{\text{sets}}$. We note that

$$\langle f[\boldsymbol{\xi}(\ell), \boldsymbol{B}(\ell)] \rangle = \frac{1}{p} \sum_{\mu=1}^{p} f(\boldsymbol{\xi}^\mu, \boldsymbol{B}^\mu) \qquad \text{(for all } \ell\text{)}$$

and that averages over all possible realisations of the training set are given by

$$\langle f[(\boldsymbol{\xi}^1, \boldsymbol{B}^1), (\boldsymbol{\xi}^2, \boldsymbol{B}^2), \ldots, (\boldsymbol{\xi}^p, \boldsymbol{B}^p)] \rangle_{\text{sets}}$$

$$= \sum_{\boldsymbol{\xi}^1} \sum_{\boldsymbol{\xi}^2} \cdots \sum_{\boldsymbol{\xi}^p} \frac{1}{2^{Np}} \int \left[\prod_{\mu=1}^{p} p(\boldsymbol{B}^\mu)\, d\boldsymbol{B}^\mu \right] f[(\boldsymbol{\xi}^1, \boldsymbol{B}^1), (\boldsymbol{\xi}^2, \boldsymbol{B}^2), \ldots, (\boldsymbol{\xi}^p, \boldsymbol{B}^p)]$$

where $\boldsymbol{\xi}^\mu \in \{-1, 1\}^N$. We normalise $\boldsymbol{B}^*$ so that $[\boldsymbol{B}^*]^2 = 1$ and choose the time unit $t = m/N$. We finally assume that $\boldsymbol{J}_0$ and $\boldsymbol{B}^*$ are statistically independent of the training vectors $\boldsymbol{\xi}^\mu$, and that they obey $J_i(0), B_i^* = \mathcal{O}(N^{-\frac{1}{2}})$ for all i.

3 Explicit Microscopic Expressions

At the m-th stage of the learning process the two simple scalar observables $Q[\boldsymbol{J}] = \boldsymbol{J}^2$ and $R[\boldsymbol{J}] = \boldsymbol{B}^* \cdot \boldsymbol{J}$, and the joint distribution of fields $x = \boldsymbol{J} \cdot \boldsymbol{\xi}$, $y = \boldsymbol{B}^* \cdot \boldsymbol{\xi}$, $z = \boldsymbol{B} \cdot \boldsymbol{\xi}$ (calculated over the questions in the training set $\tilde{D}$), are given by

$$Q[\boldsymbol{J}(m)] = \boldsymbol{J}^2(m) \qquad R[\boldsymbol{J}(m)] = \boldsymbol{B}^* \cdot \boldsymbol{J}(m) \tag{5}$$

$$P[x, y, z; \boldsymbol{J}(m)] = \frac{1}{p} \sum_{\mu=1}^{p} \delta[x - \boldsymbol{J}(m) \cdot \boldsymbol{\xi}^\mu]\, \delta[y - \boldsymbol{B}^* \cdot \boldsymbol{\xi}^\mu]\, \delta[z - \boldsymbol{B}^\mu \cdot \boldsymbol{\xi}^\mu] \tag{6}$$

For infinitely large systems one can prove that the fluctuations in mean-field observables such as $\{Q, R, P\}$, due to the randomness in the dynamics, will vanish [6]. Furthermore one assumes, with convincing support from numerical simulations, that for $N \to \infty$ the evolution of such observables, observed for different random realisations of the training set, will be reproducible (i.e. the sample-to-sample fluctuations will also vanish, which is called 'self-averaging'). Both properties are central ingredients of all current theories. We are thus led to the introduction of the averages of the observables in (5,6), with respect to the dynamical randomness and with respect to the randomness in the training set (to be carried out in precisely this order):

$$Q(t) = \lim_{N \to \infty} \langle\, \langle Q[\boldsymbol{J}(tN)] \rangle\, \rangle_{\text{sets}} \qquad R(t) = \lim_{N \to \infty} \langle\, \langle R[\boldsymbol{J}(tN)] \rangle\, \rangle_{\text{sets}} \tag{7}$$

$$P_t(x, y, z) = \lim_{N \to \infty} \langle\, \langle P[x, y, z; \boldsymbol{J}(tN)] \rangle\, \rangle_{\text{sets}} \tag{8}$$

A fundamental ingredient of our calculations will be the average $\langle \xi_i \ \text{sgn}(\boldsymbol{B} \cdot \boldsymbol{\xi}) \rangle_{(\boldsymbol{\xi}, \boldsymbol{B})}$, calculated over all realisations of $(\boldsymbol{\xi}, \boldsymbol{B})$. We find, for a wide class of $p(\boldsymbol{B})$, that

$$\langle \xi_i \ \text{sgn}(\boldsymbol{B} \cdot \boldsymbol{\xi}) \rangle_{(\boldsymbol{\xi}, \boldsymbol{B})} = \rho B_i^* + \mathcal{O}(N^{-3/2}) \tag{9}$$

where, for example,

$$\rho = \sqrt{\frac{2}{\pi}}\,(1-2\lambda) \qquad\qquad \text{(output noise)} \tag{10}$$

$$\rho = \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{1+\Sigma^2}} \qquad\qquad \text{(Gaussian input noise)} \tag{11}$$

4 Averages of Simple Scalar Observables

Calculation of $Q(t)$ and $R(t)$ using (4, 5, 7, 9) to execute the path average and the average over sets is relatively straightforward, albeit tedious. We find that

$$\begin{aligned} Q(t) = e^{-2\gamma t}Q_0 + 2\eta\rho R_0\frac{e^{-\gamma t}(1-e^{-\gamma t})}{\gamma} + \frac{\eta^2}{2\gamma}(1-e^{-2\gamma t}) \\ + \eta^2\frac{(1-e^{-\gamma t})^2}{\gamma^2}(\frac{1}{\alpha}+\rho^2) \end{aligned} \tag{12}$$

and that

$$R(t) = e^{-\gamma t}R_0 + \eta\rho\gamma^{-1}(1-e^{-\gamma t}) \tag{13}$$

where ρ is given by equations (10, 11) in the examples of output noise and Gaussian input noise, respectively. We note that the generalisation error is given by

$$E_g = \frac{1}{\pi}\arccos\left[R(t)/\sqrt{Q(t)}\right] \tag{14}$$

All models of the teacher noise which have the same ρ will thus have the same generalisation error at any time. This is true, in particular, of output noise and Gaussian input noise when their respective parameters λ and Σ are related by

$$1-2\lambda = \frac{1}{\sqrt{1+\Sigma^2}}. \tag{15}$$

With each type of teacher noise for which (9) holds, one can thus associate an effective output noise parameter λ. Note, however, that this effective teacher error probability λ will in general not be identical to the *true* teacher error probability associated with a given $p(\boldsymbol{B})$, as can immediately be seen by calculating the latter for the Gaussian input noise (2).

5 Average of the Joint Field Distribution

The calculation of the average of the joint field distribution starting from equation (8) is more difficult. Writing $\sigma = (1-\gamma/N)$, and expressing the δ functions in terms of complex exponentials, we find that

$$\begin{aligned} P_t(x,y,z) = \int\frac{d\hat{x}d\hat{y}d\hat{z}}{8\pi^3}\,e^{i(x\hat{x}+y\hat{y}+z\hat{z})}\lim_{N\to\infty}\Big\langle e^{-i[\hat{x}e^{-\gamma t}\boldsymbol{J}_0\cdot\boldsymbol{\xi}^1+\hat{y}\boldsymbol{B}^*\cdot\boldsymbol{\xi}^1+\hat{z}\boldsymbol{B}^1\cdot\boldsymbol{\xi}^1]} \\ \times\prod_{\ell=0}^{tN}\left[\frac{1}{p}\sum_{\nu=1}^{p}e^{-[i\eta\hat{x}N^{-1}\sigma^{tN-\ell}(\boldsymbol{\xi}^1\cdot\boldsymbol{\xi}^\nu)\ \mathrm{sgn}(\boldsymbol{B}^\nu\cdot\boldsymbol{\xi}^\nu)]}\right]\Big\rangle_{\mathrm{sets}} \end{aligned} \tag{16}$$

In this expression we replace $\boldsymbol{\xi}^1$ by $\boldsymbol{\xi}$ and $\boldsymbol{B}^1$ by $\boldsymbol{B}$, and abbreviate $S = \prod_{\ell=0}^{tN}[\cdots]$. Upon writing the latter product in terms of the auxiliary variables $v_\nu = (\boldsymbol{\xi}^1\cdot\boldsymbol{\xi}^\nu)/\sqrt{N}$ and $\omega_\nu = \boldsymbol{B}^\nu\cdot\boldsymbol{\xi}^\nu$, we find that for large N

$$\log S \ \sim\ \chi(\hat{x}\ \mathrm{sgn}[\boldsymbol{B}\cdot\boldsymbol{\xi}],t) - \frac{i\eta\hat{x}u_1}{\gamma}(1-e^{-\gamma t}) - \frac{\eta^2\hat{x}^2u_2}{4\gamma}(1-e^{-2\gamma t}) \tag{17}$$

where $u_1,\ u_2$ are the random variables given by

$$u_1 = \frac{1}{\alpha\sqrt{N}} \sum_{\nu>1} v_\nu \; \mathrm{sgn}(\omega_\nu), \qquad\qquad u_2 = \frac{1}{p} \sum_{\nu>1} {v_\nu}^2.$$

and with

$$\chi(w,t) = \frac{1}{\alpha} \int_0^t ds \; [e^{-[i\eta w e^{\gamma(s-t)}]} - 1] \tag{18}$$

A study of the statistics of u_1 and u_2 shows that $\lim_{N\to\infty} u_2 = 1$, and that

$$u_1 = \rho \boldsymbol{B}^* \cdot \boldsymbol{\xi} + \alpha^{-1/2} u \qquad\qquad (N \to \infty),$$

where u is a Gaussian random variable with mean equal to zero and variance unity. On the basis of these results and equations (16, 17) we find that

$$P_t(x,y,z) = \int \frac{d\hat{x} d\hat{y} d\hat{z}}{8\pi^3} \; e^{i(x\hat{x}+y\hat{y}+z\hat{z}) - \frac{1}{2}\hat{x}^2[Q-R^2-e^{-2\gamma t}(Q_0-R_0^2)] + \chi(\hat{x}\;\mathrm{sgn}[z],t) - i\hat{x}\hat{y}(R-R_0 e^{-\gamma t})}$$

$$\times \lim_{N\to\infty} \langle e^{-i[\hat{x} e^{-\gamma t} \boldsymbol{J}_0 \cdot \boldsymbol{\xi} + \hat{y} \boldsymbol{B}^* \cdot \boldsymbol{\xi} + \hat{z} \boldsymbol{B} \cdot \boldsymbol{\xi}]} \rangle_{(\boldsymbol{\xi}, \boldsymbol{B})} \tag{19}$$

where Q and R are given by the expressions (12,13) (note: $Q-R^2$ is independent of ρ, i.e. of the distribution $p(\boldsymbol{B})$). Let $x_0 = \boldsymbol{J}_0 \cdot \boldsymbol{\xi}$, $y = \boldsymbol{B}^* \cdot \boldsymbol{\xi}$, $z = \boldsymbol{B} \cdot \boldsymbol{\xi}$. We assume that, *given* y, z is independent of x_0. This condition, which reflects in some sense the property that the teacher noise preserves the perceptron structure, is certainly satisfied for the models which we are considering and is probably true of all reasonable noise models. The joint probability density then has the form $p(x_0, y, z) = p(x_0|y)p(y,z)$. Equation (19) then leads to the following expression for the conditional probability of x, given y and z:

$$P_t(x|y,z) = \int \frac{d\hat{x}}{2\pi} \; e^{i\hat{x}[x-Ry] - \frac{1}{2}\hat{x}^2[Q-R^2] + \chi(\hat{x}\;\mathrm{sgn}[z],t)} \tag{20}$$

We observe that this probability distribution is the same for all models with the same ρ and that the dependence on z is through $\tau = \mathrm{sgn}[z]$, a directly observable quantity. The training error and the student field probability density are given by

$$E_{\mathrm{tr}} = \int dx dy \sum_{\tau=\pm 1} \theta(-x\tau) P_t(x|y,\tau) P(\tau|y) P(y) \tag{21}$$

$$P_t(x) = \int dy \sum_{\tau=\pm 1} P_t(x|y,\tau) P(\tau|y) P(y) \tag{22}$$

in which $P(y) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}y^2}$. We note that the dependence of E_{tr} and $P_t(x)$ on the specific noise model arises solely through $P(\tau|y)$ which we find is given by

$$P(\tau|y) = \lambda\theta(-\tau y) + (1-\lambda)\theta(\tau y) \qquad\qquad P(\tau|y) = \frac{1}{2}(1 + \tau\,\mathrm{erf}[\mathrm{y}/\sqrt{2}\Sigma])$$

in the output noise and Gaussian input noise models, respectively. In order to simplify the numerical computation of the remaining integrals one can further reduce the number of integrations analytically. Details will be reported elsewhere.

6 Comparison with Numerical Simulations

It will be clear that there is a large number of parameters that one could vary in order to generate different simulation experiments with which to test our theory. Here we have to restrict ourselves to presenting a number of representative results. Figure 1 shows, for the output noise model, how the probability density $P_t(x)$ of

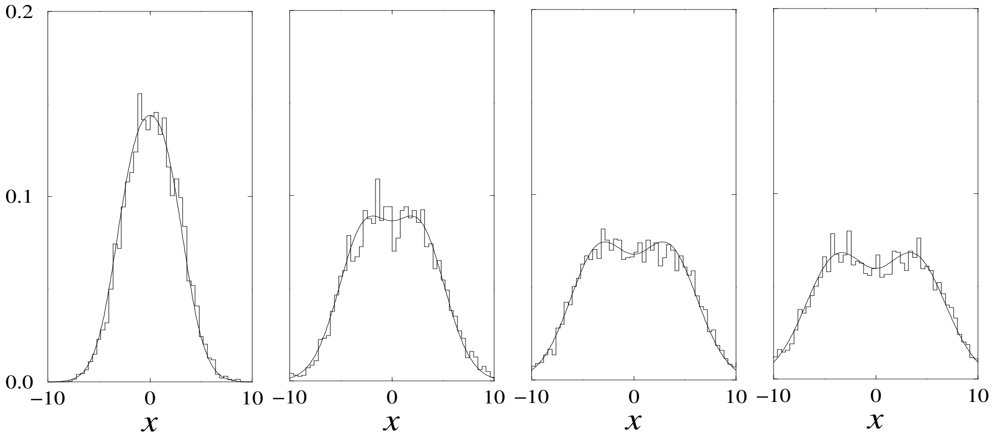

Figure 1: Student field distribution $P(x)$ for the case of output noise, at different times (left to right: $t = 1, 2, 3, 4$), for $\alpha = \gamma = \frac{1}{2}$, $J_0 = \eta = 1$, $\lambda = 0.2$. Histograms: distributions measured in simulations, ($N = 10{,}000$). Lines: theoretical predictions.

the student field $x = \boldsymbol{J} \cdot \boldsymbol{\xi}$ develops in time, starting as a Gaussian at $t = 0$ and evolving to a highly non-Gaussian distribution with a double peak by time $t = 4$. The theoretical results give an extremely satisfactory account of the numerical simulations. Figure 2 compares our predictions for the generalisation and training errors E_g and $E_{\rm tr}$ with the results of numerical simulations, for different initial conditions, $E_g(0) = 0$ and $E_g(0) = 0.5$, and for different choices of the two most important parameters λ (which controls the amount of teacher noise) and α (which measures the relative size of the training set). The theoretical results are again in excellent agreement with the simulations. The system is found to have no memory of its past (which will be different for some other learning rules), the asymptotic values of E_g and $E_{\rm tr}$ being independent of the initial student vector. In our examples E_g is consistently larger than $E_{\rm tr}$, the difference becoming less pronounced as α increases. Note, however, that in some circumstances E_{tr} can also be larger then E_g. Careful inspection shows that for Hebbian learning there are no true overfitting effects, not even in the case of large λ and small γ (for large amounts of teacher noise, without regularisation via weight decay). Minor finite time minima of the generalisation error are only found for very short times ($t < 1$), in combination with special choices for parameters and initial conditions.

7 Discussion

Starting from a microscopic description of Hebbian on-line learning in perceptrons with restricted training sets, of size $p = \alpha N$ where N is the number of inputs, we have developed an exact theory in terms of macroscopic observables which has enabled us to predict the generalisation error and the training error, as well as the probability density of the student local fields in the limit $N \to \infty$. Our results are in execellent agreement with numerical simulations (as carried out for systems of size $N = 5{,}000$) in the case of output noise; our predictions for the Gaussian input noise model are currently being compared with the results of simulations. Generalisations of our calculations to scenarios involving, for instance, time-dependent learning rates or time-dependent decay rates are straightforward. Although it will be clear that our present calculations cannot be extended to non-Hebbian rules, since they

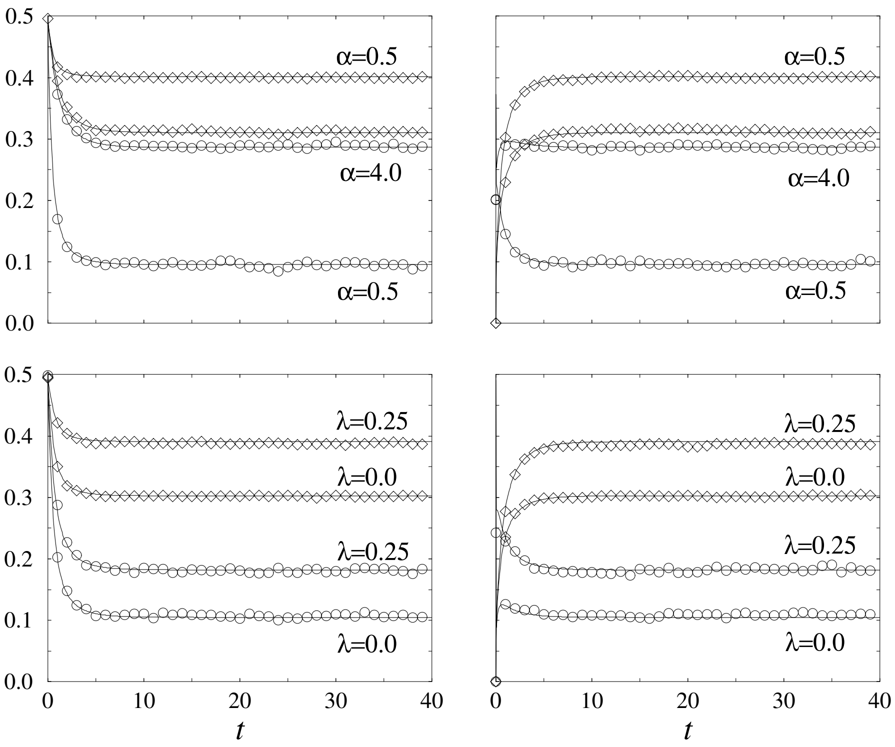

Figure 2: Generalisation errors (diamonds/lines) and training errors (circles/lines) as observed during on-line Hebbian learning, as functions of time. Upper two graphs: $\lambda = 0.2$ and $\alpha \in \{0.5, 4.0\}$ (upper left: $E_g(0) = 0.5$, upper right: $E_g(0) = 0$). Lower two graphs: $\alpha = 1$ and $\lambda \in \{0.0, 0.25\}$ (lower left: $E_g(0) = 0.5$, lower right: $E_g(0) = 0.0$). Markers: simulation results for an $N = 5{,}000$ system. Solid lines: predictions of the theory. In all cases $J_0 = \eta = 1$ and $\gamma = 0.5$.

ultimately rely on our ability to write down the microscopic weight vector $\boldsymbol{J}$ at any time in explicit form (4), they do indeed provide a significant yardstick against which more sophisticated and more general theories can be tested. In particular, they have already played a valuable role in assessing the conditions under which a recent general theory of learning with restricted training sets, based on a dynamical version of the replica formalism, is exact [6, 7].

References

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

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

[3] Krogh A. and Hertz J.A. (1992) *J.Phys. A: Math. Gen.* **25**, 1135

[4] Sollich P. and Barber D. (1997) *Europhys. Lett.* **38**, 477

[5] Sollich P. and Barber D. (1998) *Advances in Neural Information Processing Systems 10*, Eds. Jordan M., Kearns M. and Solla S. (Cambridge: MIT)

[6] Coolen A.C.C. and Saad D., King's College London preprint KCL-MTH-98-08

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