J. Phys. A: Math. Gen. **34** (2001) 9009–9026 PII: S0305-4470(01)22137-4

Supervised learning with restricted training sets: a generating functional analysis

J A F Heimel and A C C Coolen

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

Received 16 February 2001, in final form 8 June 2001
Published 12 October 2001
Online at stacks.iop.org/JPhysA/34/9009

Abstract

We study the dynamics of supervised on-line learning of realizable tasks in feed-forward neural networks. We focus on the regime where the number of examples used for training is proportional to the number of input channels N. Using generating functional techniques from spin glass theory, we are able to average over the composition of the training set and transform the problem for $N \to \infty$ to an effective single pattern system described completely by the student autocovariance, the student–teacher overlap and the student response function with exact closed equations. Our method applies to arbitrary learning rules, i.e., not necessarily of a gradient-descent type. The resulting exact macroscopic dynamical equations can be integrated without finite-size effects up to any degree of accuracy, but their main value is in providing an exact and simple starting point for analytical approximation schemes. Finally, we show how, in the region of absent anomalous response and using the hypothesis that (as in detailed balance systems) the short-time part of the various operators can be transformed away, one can describe the stationary state of the network succesfully by a set of coupled equations involving only four scalar order parameters.

PACS numbers: 87.10.+e, 02.50.−r, 05.20.−y

1. Introduction

It is now a little more than ten years since studies of the dynamics of supervised learning in artificial neural networks started appearing in the statistical physics literature. Early theoretical studies focused on on-line learning using complete training sets where the probability of the same example appearing twice during training was zero, e.g. [1–3]. This work enabled the evaluation of properties like convergence speed, generalization ability and optimal learning rates. However, such studies were still significantly removed from real-world scenarios. The most serious restriction was that one had to assume the availability of an infinite amount of training data, homogeneously distributed over the input space. In a recent article [4] it was shown that even for very simple inhomogeneity the generalization error is no longer self-averaging and deterministic. The issue of repeating examples during training is technically a much harder problem and has received much attention recently. Most of the work has focused on simple or linear learning rules [5–7] or different kinds of approximations, such as Fokker–Planck approaches [8–11] and Gaussian local field distributions [12]. Exact work on non-linear learning rules has drawn heavily on techniques from the spin glass and disordered systems community (for an early overview of these techniques, see e.g. [13]). The generating functional technique was used to study the dynamics of Gibbs learning in a perceptron with binary weights in [14, 15]. A dynamical version of the cavity method was employed in [16–18] to study gradient-descent batch learning and the methods of dynamical replica theory were applied to the problem of on-line learning in [19–22]. The on-line learning scenario in this last sequence of papers is the one that we study here, but in the present paper we adapt the generating functional method à la De Dominicis to deal with on-line learning. This paper might be the first to present exact macroscopic equations for on-line learning of restricted training sets for non-linear learning rules which are not of a gradient-descent type.

Precise definitions will be given in section 2, but the general set-up is the following. The examples presented to the student perceptron are N-dimensional vectors chosen with equal probability from a fixed training set Ω. The number of examples in Ω is $p = \alpha N$. At each presentation the student is given the teacher's classification of the pattern. The student can then decide to change its 'program', represented by the N-dimensional vector $\boldsymbol{\sigma} \in \mathbb{R}^N$, in order to resemble more the teacher's program $\boldsymbol{\tau} \in \mathbb{R}^N$. The random choice of a pattern from the training set makes the evolution of the student weight vector $\boldsymbol{\sigma}$ a stochastic process. In section 3 we write down a generating function for all the possible paths of $\boldsymbol{\sigma}$. This function can be averaged over all possible realizations of the training set Ω (a quenched disorder average). At that point we will take the limit N to infinity, to find saddle-point equations for a set of five order parameters and their conjugates. The reader who is mainly interested in results can skip section 3 and go directly to section 4, where the equations are reduced to a single exact set of three equations involving the student autocorrelation $C(t, t') = \boldsymbol{\sigma}(t) \cdot \boldsymbol{\sigma}(t')/N$, the student–teacher overlap $R(t) = \boldsymbol{\sigma}(t) \cdot \boldsymbol{\tau}/N$ and the student response function $G(t, t')$. This set gives a surprisingly simple and intuitive picture of the evolution of the order parameters and the distribution of the local fields. From that point it is easy to establish links with earlier work on infinite training sets, batch learning and linear learning rules. Numerical evidence is presented showing that the present theory is in very good agreement with the simulations.

In section 5, the stationary state of a student with constant weight decay is studied. For the stationary state one can split all the relevant order parameters into persistent and non-persistent parts. If we keep only the persistent parts and the single-time non-persistent parts, we find a closed set of equations containing just four scalar order parameters. The procedure is inspired by a similar method applied to the solution of detailed balance spin glass dynamics where it can be shown to be exact. Although the numerical evidence certainly seems to suggest that the procedure yields the correct results, we cannot prove this fact rigorously here. At the moment, it remains an interesting open question.

2. Definitions

We study on-line learning in a student perceptron characterized by a vector $\boldsymbol{\sigma} \in \mathbb{R}^N$. The student classifies patterns $\boldsymbol{\xi} \in \Omega \subset \{-1, +1\}^N$ according to $S(\boldsymbol{\xi}) = \mathrm{sgn}(\boldsymbol{\sigma} \cdot \boldsymbol{\xi})$. The student tries to learn the task set by the teacher $T(\boldsymbol{\xi}) = \mathrm{sgn}(\boldsymbol{\tau} \cdot \boldsymbol{\xi})$ with $\boldsymbol{\tau} \in \mathbb{R}^N$, i.e. we only consider linear separable classifications. The components of the weight vectors of teacher and student are assumed not to scale with N. The set Ω contains only $p = \alpha N$ examples, independently chosen with equal probability from $\{-1, +1\}^N$. Patterns will be labelled by the Greek index μ. At each iteration each pattern is equally likely to be chosen for presentation to the student, independently of previous rounds. If at step m, pattern $\mu(m)$ is presented to the learning student, the student's weight vector is slightly adjusted to converge to the desired classification according to a recipe of the general form

$$\boldsymbol{\sigma}(m+1) = \boldsymbol{\sigma}(m) + \frac{\eta}{\sqrt{N}}\boldsymbol{\xi}^{\mu(m)} F\left(\frac{\boldsymbol{\sigma}(m)\cdot\boldsymbol{\xi}^{\mu(m)}}{\sqrt{N}}, \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}^{\mu(m)}}{\sqrt{N}}\right). \tag{1}$$

The speed of the evolution is set by the learning rate η. The function $F(x, y)$ is the learning rule. Popular learning rules are, e.g.

$$F(x, y) = \begin{cases} y - x & \text{linear} \\ \mathrm{sgn}(y) & \text{Hebb} \\ \mathrm{sgn}(y) - x & \text{adaline} \\ \mathrm{sgn}(y)\Theta(-xy) & \text{perceptron} \\ |x|\mathrm{sgn}(y)\Theta(-xy) & \text{adatron} \end{cases} \tag{2}$$

where Θ is the stepfunction, $\Theta(x) = 1$ for $x \geqslant 0$ and $\Theta(x) = 0$ for $x < 0$. The first three learning rules are all linear in x, while the last two only alter the student's weights when student and teacher disagree. We restrict ourselves to learning rules which only depend on the so-called student and teacher local fields:

$$x^\mu = \frac{1}{\sqrt{N}}\boldsymbol{\sigma}\cdot\boldsymbol{\xi}^\mu \qquad y^\mu = \frac{1}{\sqrt{N}}\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\mu. \tag{3}$$

A theoretical study of perceptrons can be useful for predicting learning times, for evaluating different learning rules or for finding optimal learning rates. For this purpose one is not so much interested in predicting the specific microscopic realizations of $\boldsymbol{\sigma}$ over time, but rather in the number of errors the perceptron makes in the classification of the training set (training error E_t) and the number of errors in the classification of the complete set of examples $\{-1, +1\}^N$ (generalization error E_g):

$$E_t(\boldsymbol{\sigma}) \equiv \langle\Theta(-(\boldsymbol{\sigma}\cdot\boldsymbol{\xi})(\boldsymbol{\tau}\cdot\boldsymbol{\xi}))\rangle_\Omega = \frac{1}{p}\sum_{\boldsymbol{\xi}\in\Omega}\Theta(-(\boldsymbol{\sigma}\cdot\boldsymbol{\xi})(\boldsymbol{\tau}\cdot\boldsymbol{\xi})) \tag{4}$$

$$E_g(\boldsymbol{\sigma}) \equiv \langle\Theta(-(\boldsymbol{\sigma}\cdot\boldsymbol{\xi})(\boldsymbol{\tau}\cdot\boldsymbol{\xi}))\rangle = \frac{1}{2^N}\sum_{\boldsymbol{\xi}\in\{-1,+1\}^N}\Theta(-(\boldsymbol{\sigma}\cdot\boldsymbol{\xi})(\boldsymbol{\tau}\cdot\boldsymbol{\xi})). \tag{5}$$

Given $\boldsymbol{\sigma}$, the generalization error is independent of the training set. It is in fact a standard result in perceptron theory that this error is only dependent on the angle between student and teacher vector, i.e. the norm of $\boldsymbol{\sigma}$ and its overlap with $\boldsymbol{\tau}$.

$$E_g(\boldsymbol{\sigma}) = \frac{1}{\pi}\arccos\left(\frac{R(\boldsymbol{\sigma})}{\sqrt{C(\boldsymbol{\sigma},\boldsymbol{\sigma})}}\right). \tag{6}$$

The microscopic evolution can be used to derive expressions for the evolution of the macroscopic autocorrelation function C and the student–teacher overlap R. Taking on both sides of equation (1) the inner product with $\boldsymbol{\tau}/N$ leads to

$$R(m+1) = R(m) + \frac{\eta}{N}y^{\mu(m)}F\left(x^{\mu(m)}(m), y^{\mu(m)}\right) \tag{7}$$

while multiplying with $\boldsymbol{\sigma}(n)/N$ on both sides gives

$$C(m+1, n) = C(m, n) + \frac{\eta}{N}x^{\mu(m)}(n)F\left(x^{\mu(m)}(m), y^{\mu(m)}\right). \tag{8}$$

Finally, squaring both sides and summing over all channels gives the evolution of the squared norm of the student weight vector

$$C(m+1,m+1) = C(m,m) + \frac{2\eta}{N} x^{\mu(m)}(m) F\left(x^{\mu(m)}(m), y^{\mu(m)}\right) + \frac{\eta^2}{N} F\left(x^{\mu(m)}(m), y^{\mu(m)}\right)^2. \tag{9}$$

We see that these three equations depend only on the patterns via the local fields. The evolution of the response function G is not so easily given in this way due to cross terms like $\xi_i^\mu \xi_j^\mu$ entering the calculations. Later in the paper we will see how to deal with this. We will also give a more rigorous derivation of the continuum time limit of these equations, which were in fact already derived formally in [20], but for illustration purposes we give a short derivation at this point.

Consider the change in R after ΔN steps:

$$R(m+\Delta N) = R(m) + \frac{1}{N} \sum_{m'=m}^{m+\Delta N} \eta y^{\mu(m')} F\left(x^{\mu(m')}(m'), y^{\mu(m')}\right). \tag{10}$$

If we express R in a rescaled time $\tau = m/N$, we find

$$\frac{R(\tau+\Delta) - R(\tau)}{\Delta} = \frac{1}{\Delta N} \sum_{m'=m}^{m+\Delta N} \eta y^{\mu(m')} F\left(x^{\mu(m')}(m'), y^{\mu(m')}\right). \tag{11}$$

We now take the thermodynamic limit N just before sending Δ to 0. The probability of selecting the same pattern twice within ΔN steps vanishes when Δ goes to 0. The patterns are uncorrelated with each other (but not with $\boldsymbol{\sigma}$). This has the effect that any particular x^μ does not change very much in ΔN steps unless it is itself selected. Taking the two limits in the right order let us replace the sum over m' by an average over the local fields of all patterns in the training set:

$$\frac{dR(\tau)}{d\tau} = \eta \langle y F(x(\tau), y)\rangle. \tag{12}$$

In a very similar way we find

$$\frac{\partial C(\tau,\tau')}{\partial \tau} = \eta \langle x(\tau') F(x(\tau), y)\rangle \tag{13}$$

$$\frac{dC(\tau,\tau)}{d\tau} = 2\eta \langle x(\tau) F(x(\tau), y)\rangle + \eta^2 \langle F(x(\tau), y)^2\rangle. \tag{14}$$

Although the equations look remarkably simple, we have not dealt with the question of how to construct the local field averages which appear. This will be done using the generating functional formalism.

3. The generating functional

The random choice of a pattern $\mu(m)$ makes it more convenient to go to a description of an ensemble of students with a distribution of weight vectors, $P_m(\boldsymbol{\sigma})$, than to study the stochastic evolution of $\boldsymbol{\sigma}$ directly. The microscopic dynamics of weight vectors at time m can be written in the general form $P_{m+1}(\boldsymbol{\sigma}) = \int d\boldsymbol{\sigma}' W(\boldsymbol{\sigma} \mid \boldsymbol{\sigma}') P_m(\boldsymbol{\sigma}')$, with the transition probabilities

$$W(\boldsymbol{\sigma} \mid \boldsymbol{\sigma}') = \frac{1}{p} \sum_{\mu=1}^{p} \delta\left(\boldsymbol{\sigma} - \boldsymbol{\sigma}' - \frac{\eta}{\sqrt{N}} \boldsymbol{\xi}^\mu F\left(\frac{\boldsymbol{\sigma}' \cdot \boldsymbol{\xi}^\mu}{\sqrt{N}}, \frac{\boldsymbol{\tau} \cdot \boldsymbol{\xi}^\mu}{\sqrt{N}}\right)\right). \tag{15}$$

In order to simplify the following analysis, we first uncouple the continuum time limit from the thermodynamic limit. We do this with the often applied procedure (see e.g. [11, 20]) of making the duration of each iteration step a random variable, such that the number of steps that have been made up to time τ is given by a Poisson random variable with mean $N\tau$. This allows one to switch for finite N to a probability distribution $P_\tau(\boldsymbol{\sigma})$ depending on the continuous time τ. This distribution obeys a simple differential equation (see [11, 20] for details):

$$\frac{\partial}{\partial\tau}P_\tau(\boldsymbol{\sigma}) = N\int \mathrm{d}\boldsymbol{\sigma}'\{W(\boldsymbol{\sigma}\,|\,\boldsymbol{\sigma}') - \delta(\boldsymbol{\sigma}-\boldsymbol{\sigma}')\}P_\tau(\boldsymbol{\sigma}'). \tag{16}$$

If we prescribe the local fields $\boldsymbol{x}(\tau)$ and $\boldsymbol{y}$ and the additional fields $w^\mu(\tau) = N^{-1/2}\hat{\boldsymbol{\sigma}}\cdot\boldsymbol{\xi}^\mu$ for all patterns, the evolution of the Fourier transform $\hat{P}_\tau$ of P_τ becomes particularly simple

$$\frac{\partial}{\partial\tau}\hat{P}_\tau(\hat{\boldsymbol{\sigma}}) = L(\boldsymbol{x}(\tau),\boldsymbol{y},\boldsymbol{w}(\tau))\hat{P}_\tau(\hat{\boldsymbol{\sigma}}) \tag{17}$$

where

$$L(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}) = \frac{1}{\alpha}\sum_\mu\{\exp[-\mathrm{i}\eta w^\mu F(x^\mu, y^\mu)] - 1\}. \tag{18}$$

This can formally be integrated to

$$\hat{P}_\tau(\hat{\boldsymbol{\sigma}}) = \hat{P}_{\tau_0}(\hat{\boldsymbol{\sigma}})\exp\left[\int_{\tau_0}^{\tau}\mathrm{d}\tau' L(\boldsymbol{x}(\tau'),\boldsymbol{y},\boldsymbol{w}(\tau'))\right]. \tag{19}$$

We rediscretize time with small steps of size Δ. Later we will send Δ to zero, but this can be done completely independently of the limit $N\to\infty$. The discrete timesteps are labelled by $t=\tau/\Delta$. We find $\hat{P}_{t+1}(\hat{\boldsymbol{\sigma}}) = \hat{P}_t(\hat{\boldsymbol{\sigma}})\exp[\Delta L(\boldsymbol{x}(t),\boldsymbol{y},\boldsymbol{w}(t)) + \mathcal{O}(\Delta^2)]$. An inverse Fourier transform leads back to

$$P_{t+1}(\boldsymbol{\sigma}) = \int \mathrm{d}\boldsymbol{\sigma}' W_\Delta(\boldsymbol{\sigma}\,|\,\boldsymbol{\sigma}')P_t(\boldsymbol{\sigma}') \tag{20}$$

with a redefined transition rate W_Δ:

$$W_\Delta(\boldsymbol{\sigma}\,|\,\boldsymbol{\sigma}') = \int\frac{\mathrm{d}\hat{\boldsymbol{\sigma}}}{(2\pi)^N}\exp[\mathrm{i}\hat{\boldsymbol{\sigma}}\cdot(\boldsymbol{\sigma}-\boldsymbol{\sigma}') + \Delta L(\boldsymbol{x}(t),\boldsymbol{y},\boldsymbol{w}(t))]. \tag{21}$$

The probability for a student following a particular path can be expressed as

$$P(\boldsymbol{\sigma}(t),\boldsymbol{\sigma}(t-1),\ldots,\boldsymbol{\sigma}(0)) = W(\boldsymbol{\sigma}(t)|\boldsymbol{\sigma}(t-1))\cdots W(\boldsymbol{\sigma}(\Delta)|\boldsymbol{\sigma}(0))\,P_0(\boldsymbol{\sigma}(0)). \tag{22}$$

This distribution contains all information of the learning process and can be conveniently studied using its characteristic or moment generating function $Z[\phi]$ defined as

$$\begin{aligned}Z[\boldsymbol{\psi}] &= \left\langle \exp\left[\mathrm{i}\Delta\sum_t\sum_i \psi_i(t)\sigma_i(t)\right]\right\rangle \\ &= \int \mathrm{D}\boldsymbol{\sigma}\mathrm{D}\hat{\boldsymbol{\sigma}}\prod_t[\mathrm{d}\boldsymbol{x}(t)\mathrm{d}\boldsymbol{w}(t)]\,\mathrm{d}\boldsymbol{y}\Gamma[\boldsymbol{y},\boldsymbol{x},\boldsymbol{w},\boldsymbol{\sigma}]\exp\left[\sum_t \mathrm{i}\Delta\boldsymbol{\psi}(t)\cdot\boldsymbol{\sigma}(t)\right] \\ &\quad\times\exp\left[\sum_t \mathrm{i}\hat{\boldsymbol{\sigma}}(t)\cdot(\boldsymbol{\sigma}(t+1)-\boldsymbol{\sigma}(t)) + \Delta\sum_t L(\boldsymbol{x}(t),\boldsymbol{y},\boldsymbol{w}(t))\right]\end{aligned} \tag{23}$$

where $\mathrm{D}\boldsymbol{\sigma} = \prod_t(\mathrm{d}\boldsymbol{\sigma}(t)/\sqrt{2\pi})$. The local fields are self-consistently prescribed by the δ functions confined to the function Γ, which is given by

$$\Gamma[\boldsymbol{y},\boldsymbol{x},\boldsymbol{w},\boldsymbol{\sigma}] = \prod_\mu\delta\left[y^\mu - \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right]\prod_t\delta\left[x^\mu(t) - \frac{\boldsymbol{\sigma}(t)\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right]\delta\left[w^\mu(t) - \frac{\hat{\boldsymbol{\sigma}}(t)\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right].$$

In the thermodynamic limit ($N \to \infty$), all the macroscopic observables in this model are self-averaging with respect to the realization of the training set. To avoid the difficulty of choosing a typical training set, we can thus safely consider the disorder-averaged generating function $[Z]_{\mathrm{dis}}$. The only term involving the actual patterns is Γ. The quenched disorder average of Γ is

$$[\Gamma]_{\mathrm{dis}} = \int \prod_t \left[\frac{\mathrm{d}\hat{\boldsymbol{x}}(t)}{(2\pi)^p}\frac{\mathrm{d}\hat{\boldsymbol{w}}(t)}{(2\pi)^p}\right]\frac{\mathrm{d}\hat{\boldsymbol{y}}}{(2\pi)^p}\prod_\mu \exp\left[\mathrm{i}\hat{y}^\mu y^\mu + \sum_t \mathrm{i}\hat{x}^\mu(t)x^\mu(t) + \sum_t \mathrm{i}\hat{w}^\mu(t)w^\mu(t)\right]$$
$$\times\, 2^{-N}\sum_{\boldsymbol{\xi}^\mu\in\{\pm1\}^N} \exp\frac{-\mathrm{i}}{\sqrt{N}}\boldsymbol{\xi}^\mu\cdot\left[\hat{y}^\mu\boldsymbol{\tau} + \sum_t \hat{x}^\mu(t)\boldsymbol{\sigma}(t) + \sum_t \hat{w}^\mu(t)\hat{\boldsymbol{\sigma}}(t)\right].$$

We see that the conjugate variables $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{w}}$ will have to be of order Δ to ensure the existence of the continuum time limit, when Δ goes to 0 and the number of steps in the summations will grow as $1/\Delta$. Of the term on the second line, only the quadratic terms in $\boldsymbol{\tau}$, $\boldsymbol{\sigma}$ and $\hat{\boldsymbol{\sigma}}$ survive in the thermodynamic limit. Near this limit we find that this term containing the training patterns becomes

$$\prod_{\mu,i}\exp\left[-\frac{1}{2N}\left(\hat{y}^\mu\tau_i + \sum_t \hat{x}^\mu(t)\sigma_i(t) + \sum_t \hat{w}^\mu(t)\hat{\sigma}_i(t)\right)^2\right].$$

We assume that the initial probability distribution $P_0(\boldsymbol{\sigma})$ factorizes over sites. Full factorization of the generating function over patterns and input channels can then be achieved if we introduce the following order parameters and their conjugates via δ functions:

$$R_t = \frac{1}{N}\sum_i \sigma_i(t)\tau_i \qquad r_t = \frac{1}{N}\sum_i \hat{\sigma}_i(t)\tau_i \qquad C_{tt'} = \frac{1}{N}\sum_i \sigma_i(t)\sigma_i(t')$$
$$c_{tt'} = \frac{1}{N}\sum_i \hat{\sigma}_i(t)\hat{\sigma}_i(t') \qquad K_{tt'} = \frac{1}{N}\sum_i \sigma_i(t)\hat{\sigma}_i(t').$$

The generating function attains a form suitable for saddle-point integration:

$$[Z[\psi]]_{\mathrm{dis}} \propto \int \cdots \exp\left[N(\Psi + \Phi + \Omega)\right]. \tag{24}$$

There are three distinct leading order contributions to the exponent. The first is a 'book-keeping' term, linking the order parameters to their conjugates:

$$\Psi = \mathrm{i}\hat{R}\cdot R + \mathrm{i}\hat{r}\cdot r + \mathrm{i}\,\mathrm{Tr}[\hat{C}^T C + \hat{K}^T K + \hat{c}^T c]. \tag{25}$$

Note that the existence of a continuum time limit again implies that the single time conjugate variables $\hat{R}$ and $\hat{r}$ are actually of order Δ and the two time conjugate variables $\hat{C}$, $\hat{K}$ and $\hat{c}$ are of order Δ^2. This, with the exception of the diagonal terms, may be of order Δ. Only in this scaling, summations over τ/Δ time steps appearing in the generating function will remain finite when the continuum time limit is taken. The second term reflects the coupled dynamics of the local fields

$$
\Phi = \frac{1}{N}\sum_{\mu}\log\int \frac{\mathrm{d}y\mathrm{d}\hat{y}}{2\pi}\mathrm{D}x\mathrm{D}\hat{x}\mathrm{D}w\mathrm{D}\hat{w}\exp\left[\mathrm{i}\hat{w}\cdot w + \frac{\Delta}{\alpha}\sum_{t}\left(\mathrm{e}^{-\mathrm{i}\eta w_t F(x_t,y)} - 1\right) + \mathrm{i}\hat{y}\left(y-\theta_y^{\mu}\right)\right.
$$

$$
\left. + \mathrm{i}\hat{x}\cdot\left(x-\theta_x^{\mu}\right) - \frac{1}{2}\hat{x}C\hat{x} - \frac{1}{2}\hat{y}^2 - \hat{x}K\hat{w} - \hat{y}R\cdot\hat{x} - \frac{1}{2}\hat{w}c\hat{w} - \hat{y}r\cdot\hat{w}\right] \qquad (26)
$$

where we have added additional sources $\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_y$ to couple to $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$. These sources act as biases of teacher and student. The third term describes the evolution of the now decoupled weight components

$$
\Omega = \frac{1}{N}\sum_{i}\log\int \mathrm{D}\sigma\mathrm{D}\hat{\sigma}\,P_{0i}(\sigma_0)\exp\big[-\mathrm{i}\tau_i\hat{R}\cdot\sigma - \mathrm{i}\tau_i\hat{r}\cdot\hat{\sigma} - \mathrm{i}\sigma\hat{C}\sigma - \mathrm{i}\hat{\sigma}\hat{c}\hat{\sigma}
$$

$$
- \mathrm{i}\sigma\hat{K}\hat{\sigma} + \mathrm{i}\hat{\sigma}G_0^{-1}\sigma - \mathrm{i}\hat{\sigma}\cdot\theta_i + \mathrm{i}\sigma\cdot\psi_i\big] \qquad (27)
$$

where $\left[G_0^{-1}\right]_{tt'} = \delta_{t+1,t'} - \delta_{tt'}$ and where we include an external driving force $\theta_i(t)$ in the system. With a modest amount of foresight we write $G_{tt'} = -\mathrm{i}K_{tt'}$. Upon taking derivatives with respect to the generating fields $\{\psi_i(t), \theta_i(t)\}$, we find *at* the relevant saddle-point:

$$
R_t = \lim_{N\to\infty}\frac{1}{N}\sum_i[\langle\sigma_i(t)\tau_i\rangle]_{\mathrm{dis}}
$$

$$
C_{tt'} = \lim_{N\to\infty}\frac{1}{N}\sum_i[\langle\sigma_i(t)\sigma_i(t')\rangle]_{\mathrm{dis}}
$$

$$
G_{tt'} = \lim_{N\to\infty}\frac{1}{N}\sum_i\frac{\partial}{\partial\theta_i(t')}[\langle\sigma_i(t)\rangle]_{\mathrm{dis}}.
$$

Using the built-in normalization $[Z(0)]_{\mathrm{dis}} = 1$, we also find

$$
r_t = \lim_{N\to\infty}\frac{1}{N}\sum_i\frac{\partial}{\partial\theta_i(t)}[\langle\tau_i\rangle]_{\mathrm{dis}} = 0
$$

$$
c_{tt'} = \lim_{N\to\infty}\frac{1}{N}\sum_i\frac{\partial^2}{\partial\theta_i(t)\partial\theta_i(t')}[Z(0)]_{\mathrm{dis}} = 0.
$$

If we perform the saddle-point integration, we find in addition that

$$
\mathrm{i}\hat{R}_t = -\lim_{N\to\infty}\frac{1}{N}\sum_{\mu}\frac{\partial^2}{\partial\theta_y^{\mu}\partial\theta_x^{\mu}(t)}[Z(0)]_{\mathrm{dis}} = 0
$$

$$
\mathrm{i}\hat{C}_{tt'} = -\lim_{N\to\infty}\frac{1}{N}\sum_{\mu}\frac{\partial^2}{\partial\theta_x(t)\partial\theta_x(t')}[Z(0)]_{\mathrm{dis}} = 0.
$$

At this point we can already simplify (or remove altogether) the generating fields $\theta_i(t) = \theta_t, \theta_x^{\mu}(t) = \theta_{xt}$, $\theta_y^{\mu}(t) = \theta_{yt}$ and $\psi_i(t) = 0$. The external fields θ_x and θ_y can be interpreted as biases or thresholds of the student and teacher, respectively. Without loss of generality we may set $\tau_i = 1$. The evolution of the local fields and the weight vector are now linked only via the remaining non-zero order parameters. We proceed to evaluate the two separate processes at the saddle-point.

3.1. Pattern average Φ

Focusing on the evaluation of the pattern average Φ we find that the terms involving w can be interpreted as averages over a Poisson distribution

$$\begin{aligned}\int \frac{\mathrm{d}w_t}{2\pi}\exp\left[\mathrm{i}\hat{w}_t w_t + \frac{\Delta}{\alpha}\left(\mathrm{e}^{-\mathrm{i}\eta w_t F(x_t,y)}-1\right)\right]\\ &= \sum_{k_t=0}^{\infty}\int\frac{\mathrm{d}w_t}{2\pi}\exp\left[\mathrm{i}\hat{w}_t w_t - \mathrm{i}\eta k_t w_t F(x_t,y) - \frac{\Delta}{\alpha}\right]\frac{1}{k_t!}\left(\frac{\Delta}{\alpha}\right)^{k_t}\\ &= \sum_{k_t=0}^{\infty}\delta(\hat{w}_t - \eta k_t F(x_t,y))\mathbb{P}(k_t)\end{aligned}$$

where $\mathbb{P}(k)$ is a Poisson distribution with average Δ/α. For $\Delta N \gg 1, \mathbb{P}(k)$ gives the probability that a specific pattern is presented k times to the student in the time interval Δ. The saddle-point equations of the remaining non-zero order parameters are found to be

$$\hat{r}_t = \alpha\frac{\partial}{\partial\theta_y}\langle f_t\rangle_\Phi \qquad 2\mathrm{i}\hat{c}_{tt'} = \alpha\langle f_t f_{t'}\rangle_\Phi \qquad \mathrm{i}\hat{G}_{tt'} = -\alpha\frac{\partial}{\partial\theta_{xt}}\langle f_{t'}\rangle_\Phi \tag{28}$$

with the shorthand $f_t = \eta k_t F(x_t, y)$. The average $\langle\cdot\rangle_\Phi$ uses the measure implied by equation (26). Performing the disorder average turns the $\hat{y}$ integral into a Gaussian integral. Evaluating this integration yields

$$\begin{aligned}\Phi = \alpha\log\int\frac{\mathrm{d}y}{\sqrt{2\pi}}\mathrm{D}x\mathrm{D}\hat{x}\prod_t\left[\sum_{k_t}\mathbb{P}(k_t)\right]\\ &\times\exp\left[-\frac{1}{2}(y-\theta_y)^2 - \frac{1}{2}\hat{x}D\hat{x} + \mathrm{i}\hat{x}\cdot(x-\theta_x - Gf - R(y-\theta_y))\right]\end{aligned}\tag{29}$$

where we introduce the student autocovariance $D_{tt'} = C_{tt'} - R_t R_{t'}$. We note the operator identity $\partial/\partial\theta_y = y - R\cdot\partial/\partial\theta_x$, which in turn implies using (28) that

$$\hat{r}_t = \alpha\langle y f_t\rangle_\Phi + \sum_{t'}\mathrm{i}\hat{G}^T{}_{tt'}R_{t'}. \tag{30}$$

3.2. Weight component average Ω

The saddle-point equations involving the weight vectors are

$$R_t = \langle\sigma_t\rangle_\Omega \qquad C_{tt'} = \langle\sigma_t\sigma_{t'}\rangle_\Omega \qquad G_{tt'} = \frac{\partial}{\partial\theta_{t'}}\langle\sigma_t\rangle_\Omega \tag{31}$$

where $\langle\cdot\rangle_\Omega$ is an average with the measure induced by (27). This measure can be generated by the stochastic process: $-\hat{r} + \mathrm{i}\hat{G}^T\sigma + G_0^{-1}\sigma - \theta - \rho = 0$, where ρ_t is a Gaussian noise with zero mean and covariance $\langle\rho_t\rho_{t'}\rangle = \Lambda_{tt'} \equiv 2\mathrm{i}\hat{c}_{tt'}$. From this process, we find a simple expression for σ (upon setting $\theta = 0$)

$$\sigma = G(\hat{r} + \rho) \tag{32}$$

with the response, student–teacher overlap and student autocovariance given by

$$G = \left[G_0^{-1} + \mathrm{i}\hat{G}^T\right]^{-1} \qquad R = G\hat{r} \qquad D = G\Lambda G^T. \tag{33}$$

4. Effective single pattern process

Upon combining the results of the previous two paragraphs, we find a closed set of exact equations relating the evolution of R, D and G to the evolution of the local field distribution implied by the measure in (29). Setting $\theta_y = 0$ in this equation, we find that the distribution is generated by the stochastic process for a student-pattern overlap

$$x_t = R_t y + \sum_{t'} G_{tt'} f_{t'} + z_t + \theta_{xt} \tag{34}$$

where y and z_t are independent Gaussian random variables with zero mean and variances $\langle y^2 \rangle = 1$ and $\langle z_t z_{t'} \rangle = D_{tt'}$. In general x_t will depend on previous values of x via the term $[Gf]_t = \eta \sum_{t'} G_{tt'} k_{t'} F(x_{t'}, y)$. The conjugate response function can be given in terms of z after partial integration of the third part of equation (28) as

$$\mathrm{i}\hat{G}^T{}_{tt'} = -\alpha \left\langle f_t \sum_s D^{-1}_{t's} z_s \right\rangle \tag{35}$$

where now $\langle\cdot\rangle$ denotes the Gaussian averages over y and all z_t's.

The evolution of the order parameters is given using the bare response G_0. To ensure that the students' weight distribution will eventually reach a stationary state, we let the weights decay with rate γ. The bare response then takes the form $\left[G_0^{-1}\right]_{tt'} = \delta_{t+1,t'} - \lambda\delta_{tt'}$, where $\lambda \equiv 1 - \Delta\gamma$. In the limit of $\Delta \to 0$, this corresponds to $[G_0]_{tt'} = \Theta(t - t' - \Delta/2) \exp[-\Delta\gamma(t - t' - 1)]$. Using equation (33) along with the relation (30), we find

$$\left[G_0^{-1} R\right]_t = \hat{r}_t - \sum_s \mathrm{i}\hat{G}^T{}_{ts} R_s = \alpha\langle y f_t\rangle. \tag{36}$$

The combination of equations (33) and (35) gives the evolution of D

$$\left[G_0^{-1} D\right]_{tt'} = \alpha\langle f_t [Gf + z]_{t'}\rangle = \alpha\langle f_t (x_{t'} - R_{t'} y)\rangle \tag{37}$$

where we set θ_x to 0. For the evolution of the diagonal terms of D we use the first and third relations in equation (33) and a little algebra to find

$$D_{t+1,t+1} = \lambda^2 D_{tt} + \Lambda_{tt} - 2[\mathrm{i}\hat{G}^T G\Lambda]_{tt} + [\mathrm{i}\hat{G}^T G\Lambda G^T \mathrm{i}\hat{G}]_{tt} + 2\lambda[G\Lambda]_{tt} - 2\lambda[\mathrm{i}\hat{G}^T G\Lambda G^T]_{tt}. \tag{38}$$

Now recall that the non-diagonal terms of $\hat{G}$ and Λ (proportional to $\hat{c}$) are of the order Δ^2, while the number of steps in the summation will be τ/Δ for the proper continuum time τ. The diagonal terms of the two conjugate time order parameters are of order Δ or smaller. Using this knowledge we can determine the order of all the terms occuring in this last equation:

$$\begin{array}{lll} \Lambda_{tt} = \mathcal{O}(\Delta) & \lambda[\mathcal{G}\Lambda]_{\sqcup\sqcup} = \mathcal{O}(\Delta) & \lambda[\mathrm{i}\hat{\mathcal{G}}^{\mathcal{T}}\mathcal{G}\Lambda\mathcal{G}^{\mathcal{T}}]_{\sqcup\sqcup} = \mathcal{O}(\Delta) \\ [\mathrm{i}\hat{G}^T G\Lambda]_{tt} = \mathcal{O}(\Delta^{\in}) & [\mathrm{i}\hat{\mathcal{G}}^{\mathcal{T}}\mathcal{G}\Lambda\mathcal{G}^{\mathcal{T}}\mathrm{i}\hat{\mathcal{G}}]_{\sqcup\sqcup} = \mathcal{O}(\Delta^{\in}). & \end{array}$$

We see that in the continuum time limit, where Δ goes to 0, only the terms in the first line survive, giving

$$\begin{aligned} D_{t+1,t+1} &= \lambda^2 D_{tt} + 2\lambda[G\Lambda]_{tt} - 2\lambda[\mathrm{i}\hat{G}^T G\Lambda G^T]_{tt} + \Lambda_{tt} + \mathcal{O}(\Delta^2) \\ &= \lambda^2 D_{tt} + 2\lambda\alpha\langle f_t(x_t - yR_t)\rangle + \alpha\left\langle f_t^2\right\rangle + \mathcal{O}(\Delta^2). \end{aligned}$$

In terms of $Q_t \equiv D_{tt} - R_t^2$ this leads to

$$Q_{t+1} = \lambda^2 Q_t + 2\lambda\alpha\langle f_t x_t\rangle + \alpha\left\langle f_t^2\right\rangle + \mathcal{O}(\Delta^2). \tag{39}$$

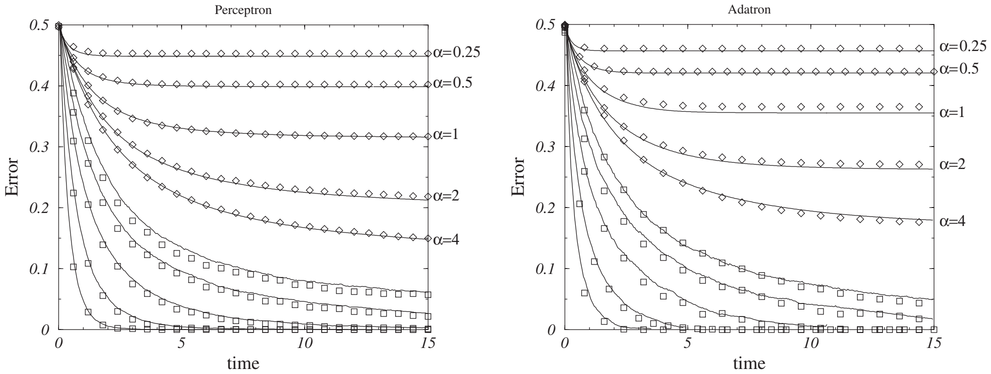

Figure 1. The evolution of the generalization (upper lines with diamonds) and training error (lower lines with squares) for the perceptron (left) and adatron (right) learning rules for various training set sizes. The lines correspond to (generalization error: top to bottom, training error: bottom to top) $\alpha = 0.25, 0.5, 1, 2, 4$. The markers correspond to single run simulations ($N = 6000$) with no decay and learning rates $\eta = 1$ (perceptron) and $\eta = 1.5$ (adatron). The solid lines are the results of numerical calculations of the effective single pattern process with $M = 20\,000$ and time step $\Delta = 0.05$.

We see that in the continuum time limit these macroscopic equations indeed coincide with the evolution equations derived in section 2. Finally, the generating functional formalism also allows us to determine the evolution of the response function. The combination of equations (33) and (28) gives

$$\left[G_0^{-1}G\right]_{tt'} = \mathbb{I}_{tt'} + \alpha \sum_s \left\langle \frac{\partial f_t}{\partial \theta_{xs}} \right\rangle G_{st'}. \tag{40}$$

The generalization error is a direct function of all these order parameters, while the training error is a slave of the local field distribution governed by them:

$$E_g(t) = \frac{1}{\pi} \arccos\left(\frac{R_t}{\sqrt{Q_t}}\right) \qquad E_t(t) = \langle \Theta(-x_t y) \rangle. \tag{41}$$

The evolution of the order parameters can be calculated numerically by a Monte Carlo procedure similar to the single-spin procedure outlined in [23]. The general idea is to follow the evolution of M pattern overlaps. For each of these patterns, one generates at time $t = 0$ a teacher overlap y from the standard normal distribution. Time is discretized with unit Δ. At each time step and for *each* pattern, one generates the Gaussian noise z_t, correlated with the previous noise values $z_{t'}$ *for that particular pattern* and a Poissonian random variable k_t. Averages over all patterns are Monte Carlo implementations of the averages occurring in the evolution equations for D, R and G. By increasing M and decreasing Δ the evolution of the $N \to \infty$-perceptron can be calculated to arbitrary precision. This is shown for various α in figure 1 with $M = 20\,000$ and $\Delta = 0.05$. The figures illustrate that the agreement of the theory with the simulations is very good.

4.1. Batch learning

So far, we have treated only the case of on-line learning. This is the most widely applied learning scenario, but much of the analytical work on learning with restricted training sets has been devoted to off-line or batch learning. In batch learning one first calculates the average effect of learning (a large sample of) the entire training set, before making a weight update.

For small learning rates, batch and on-line learning ought to generate the same macroscopic flow. For completeness we discuss here what changes when we switch from an on-line to a batch scenario. The effect on the theory as presented above is the disappearance of the extra noise term $\langle f_t^2 \rangle$ in the evolution of Q in equation (39) and the replacement of the Poisson variable k_t by its average Δ/α. The intuition behind the first change is that big changes in the student weight vector can no longer take place after a single pattern is presented; the weights undergo a much smoother evolution due to the averaging of the update over all patterns. As a result of the second change, the student training pattern overlap becomes

$$x_t = R_t y + z_t + \eta \frac{\Delta}{\alpha} \sum_s G_{ts} F(x_s, y). \tag{42}$$

This equation was derived earlier in the context of gradient-descent batch learning by Wong *et al* using an elegant application of the dynamical cavity method [18]. Again, the reason for the change in a training pattern overlap x_t is that instead of big changes when k_t times that particular pattern is presented to the student in time interval $(\Delta t, \Delta(t+1))$, now during an interval, x_t feels the average effect of the influence of the pattern. These are the only changes necessary in the present analysis when switching from on-line to batch learning.

4.2. Linear learning rules

The average occurring in the evolution of the response function G in (40) can be explicitly calculated if the student uses a learning rule that is linear in x, e.g. the linear, Hebbian or adaline rules. For these types of rules of the form $F(x, y) = g(y) - cx$, we find

$$\frac{\partial f_t}{\partial x_{t'}} = -\eta c k_t \left(\delta_{tt'} + \sum_s G_{ts} \frac{\partial f_s}{\partial x_{t'}} \right). \tag{43}$$

The causality of G allows us to perform the Poisson averages and a little matrix algebra leads to

$$G_0^{-1} G = \mathbb{I} - \Delta c \eta \left[\mathbb{I} + c\eta \frac{\Delta}{\alpha} G \right]^{-1} G. \tag{44}$$

The resulting response is translation invariant, i.e. $G_{tt'} = G(\Delta(t - t'))$ for $t \geqslant t'$. The on-line response found here for linear learning rules agrees with the batch results found for the linear rule in [5] and the adaline rule in [17]. The Fourier transform of the previous relation reads

$$G^{-1}(\omega) = \gamma - \mathrm{i}\omega + c\eta \frac{1}{1 + c\eta/\alpha G(\omega)}. \tag{45}$$

This equation is analysed in [5]. For $\gamma = 0$ and for $c \neq 0$, a transition in the behaviour of the response takes place at $\alpha_c = 1$. This position is identical for on-line and off-line learning. The nature of this transition is easily understood. Without decay, the evolution of the weight vector is confined to the linear subspace spanned by the patterns in the training set. Below $\alpha_c = 1$, the random patterns cannot span the whole N-dimensional space, resulting in a non-decaying part of the response function. This argument is valid for general rules without decay.

The student overlap with a particular pattern can also be written in a more explicit way:

$$x = [\mathbb{I} + c\eta G K]^{-1} (Ry + z + \eta g(y) G k) \qquad K_{tt'} \equiv k_t \delta_{tt'}. \tag{46}$$

The final results are rather cumbersome, but all the averages appearing in the evolution of the order parameters involving k and z can be calculated without any problems. The only remaining integrals are of the form $\langle g(y) y \rangle$ and $\langle g(y)^2 \rangle$ with the standard Gaussian measure.

4.3. Infinite training sets

To compare our results to the well-known unrestricted training set results, we take the limit $\alpha \to \infty$. In this case the probability of repeating an example is zero. This is reflected in the fact that $\langle k\rangle \to 0$ as $\alpha \to \infty$. Given y, the local fields x are random variables given by

$$x_t = yR_t + z_t \tag{47}$$

or, equivalently, x_t is a Gaussian random variable with mean yR_t and covariance $D_{tt'}$. The effects of the retarded self-interaction caused by G thus completely vanish. If we go to a continuum time description, we recover equations found in, for example, [2, 24]. The evolution of the student–teacher overlap and the student self-overlap is given by

$$\frac{\mathrm{d}R}{\mathrm{d}t} = -\gamma R + \eta\langle yF(x,y)\rangle_t \tag{48}$$

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = -2\gamma Q + 2\eta\langle xF(x,y)\rangle_t + \eta^2\langle F(x,y)^2\rangle_t \tag{49}$$

with the Gaussian single-time average defined by $\langle x\rangle_t = 0$, $\langle y\rangle_t = 0$, $\langle x^2\rangle_t = Q_t$, $\langle y^2\rangle_t = 1$, $\langle xy\rangle_t = R_t$.

5. Stationary state

Many learning rules will not reach a stationary state that is independent of the initial conditions, as soon as weight decay is absent. Weight decay, or another type of constraint, may also be necessary to bound the length of the student vector. In the Hebbian case, for example, the student weights keep on growing in the direction of the perceived teacher, regardless of the size of the training error. In order for the student to reach a stationary state, we assume that the weight decay γ is large enough to bound R and C and that the integrated response or susceptibility g is finite:

$$g \equiv \lim_{t\to\infty} \Delta \sum_{t'} G_{tt'} < \infty. \tag{50}$$

This condition is known in the disordered systems literature as *absence of anomalous response*. We also assume that for sufficiently large t the order parameters become time-translation invariant: $R_t = R$, $G_{t+\tau,t} = G_\tau$, $D_{t+\tau,t} = D_\tau$. These assumptions are related to the replica symmetry ansatz in the replica equilibrium analyses [13]. We split the covariance kernel D_t into a persistent part $d = \lim_{t\to\infty} D_t$ and a non-persistent part $\tilde{D}_t = D_t - d$. If d exists, then

$$d = \bar{D} \equiv \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} D_t.$$

Given time-translation invariance, one derives from equations (36) and (37) that

$$R = \frac{\eta}{\gamma}\langle yF(x,y)\rangle \tag{51}$$

$$d = \lim_{\tau\to\infty}\lim_{t\to\infty} \frac{\eta}{\gamma}\langle F(x_{t+\tau},y)(x_t - yR)\rangle = \frac{\eta}{\gamma}\langle \bar{F}(\bar{x} - yR)\rangle. \tag{52}$$

An earlier relation found involving the covariance (33) now yields

$$d = \alpha\overline{G\langle ff^T\rangle G^T} = \frac{\eta^2 g^2}{\alpha}\langle \bar{F}^2\rangle \tag{53}$$

while the stationary value of Q can be found from (39):

$$Q = \frac{\eta}{\gamma}\langle F(x,y)x\rangle + \frac{\eta^2}{2\gamma}\langle F(x,y)^2\rangle. \tag{54}$$

All the averages either involve a single time or two infinitely separated distant times. We lack an explicit expression for the single-time probability distribution of x_t. The probability of x_t is related to realizations of x at previous times via the response function G. This makes the evaluation of the averages as difficult as solving the dynamical equations themselves. The same problem exists in the field of (Ising) spin glasses and recurrent neural networks. In those cases where the stationary state is in detailed balance (e.g. when the dynamics are of gradient-descent type and the systems feel a Gaussian white noise) a fluctuation dissipation theorem (FDT) connects the correlation C and the response G. It is known for such systems that when calculating the persistent and the single-time parts of the correlation and the integrated response, the non-persistent parts can be chosen arbitrarily as long as the FDT is obeyed. In particular one can set them to zero and take only the persistent parts and the integrated response into account. Although there are large differences between the learning perceptron discussed here and the aforementioned spin systems (for one, the learning rule F does not have to be a gradient), we assume that this decoupling property of persistent from non-persistent parts still holds[1]. We replace the stationary distribution of x generated by equation (34) by a distribution generated by a stochastic relation containing only the integrated response g and random variables described by the persistent part of the covariance matrix D, the single-time correlation Q and the student–teacher overlap R

$$x_t = yR + z + \tilde{z}_t + \frac{\eta}{\alpha} g \bar{F} \tag{55}$$

where y, z and $\tilde{z}_t$ are all independent Gaussian random variables with zero mean and covariances $\langle y^2 \rangle = 1, \langle z^2 \rangle = d$ and $\langle \tilde{z}_t \tilde{z}_{t'} \rangle = (Q - R^2 - d)\delta_{tt'}$. The average learning term $\bar{F}$ for a specific pattern with a certain (y, z), can be expressed self-consistently as

$$\bar{F}_{yz} \equiv \lim_{T\to\infty} \frac{1}{T} \sum_{t<T} F(x_t, y) = \int \mathrm{d}\tilde{z}\, p(\tilde{z}) F\left(yR + z + \tilde{z} + \frac{\eta g}{\alpha} \bar{F}_{yz}, y \right). \tag{56}$$

For Hebbian learning one has $\bar{F} = \mathrm{sgn}(y)$, but in general (56) will be a transcendental equation, so one has to revert to numerical methods to solve it. Once $\bar{F}$ can be found for any point (y, z), the remaining two independent Gaussian integrals over y and z can be evaluated to close equations (51)–(54). The remaining closed set can be solved numerically. Results for Hebbian and perceptron learning rules and various training set sizes are presented in figure 2. For the perceptron rule, the results shown in figure 3 compare E_g and E_t for different decay strengths. Perceptron results are independent of η. The theoretical predictions seem to be in almost perfect agreement with the simulations. Although no adatron results are shown, we expect that the proposed procedure is equally valid for this latter rule. Our method of calculation is valid only when G is time-translation invariant and the integrated response is bounded. For this to happen, we need the presence of a weight decay. The complication is that any decay, however small, will cause the adatron student weight vector to vanish. An alternative way of ensuring that the student ensemble reaches a stationary state that does not exhibit this problem is by constraining $\boldsymbol{\sigma}$ to a sphere. This can be implemented by choosing $\gamma_t \propto (Q_t - 1)$. However, the adatron rule yields zero training error in this set-up. This causes other problems in numerically evaluating the stationary state equations.

5.1. Distribution of local fields

As seen earlier, an important simplifying effect of the limit $\alpha \to \infty$ is to render the local fields x and y Gaussian. This happens irrespective of the learning rule involved. As soon as

[1] Note that a rigorous proof would first require the derivation of a non-equilibrium generalization of FDT theorems.

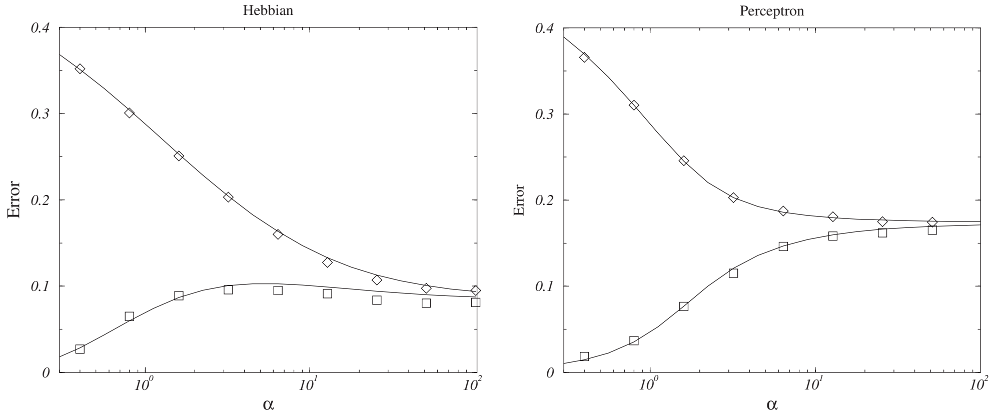

Figure 2. Stationary generalization (upper line) and training error (lower line) for Hebbian (left) and perceptron (right) learning rules. Learning rate $\eta = 1$ and decay $\gamma = 0.1$. Markers are simulation results of a single run with $N = 6000$ input channels. Solid lines are theoretical predictions obtained under the assumptions of section 5.

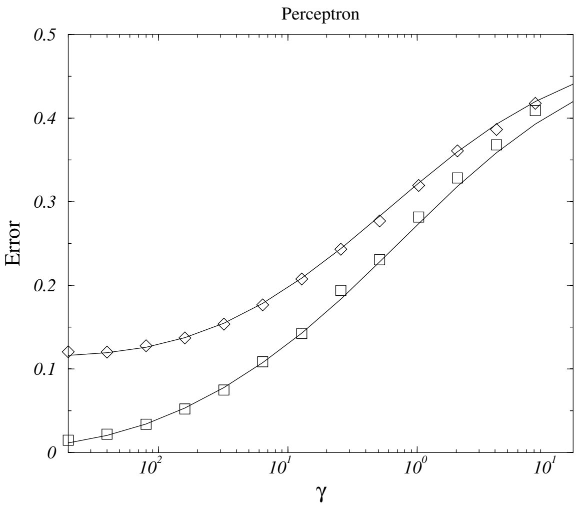

Figure 3. Stationary generalization (upper line) and training error (lower line) for perceptron learning rule with various decay rates γ. Training set size $\alpha = 4$. The data points are simulation results of a single run with $N = 6000$ input channels. The curves are theoretical predictions obtained under the assumptions of section 5.

$\alpha < \infty$, the effect of the extra term Gf in equation (55) sets in and the Gaussian form of the distribution evaporates for non-linear rules. The non-Gaussian form of the joint local field distribution has been discussed at length in [20], but equation (55) gives an intuitive idea of the origin of the deviations reported there.

For a Hebbian learning rule, $F(x, y) = \mathrm{sgn}(y)$, the conditional distribution $p(x \mid y)$ remains Gaussian with variance D, but will be shifted away from the mean yR by the amount $\eta g\,\mathrm{sgn}(y)/\alpha$. An example with $\alpha = 1$ and $\gamma = 0.1$ is shown in figure 4(middle). For the

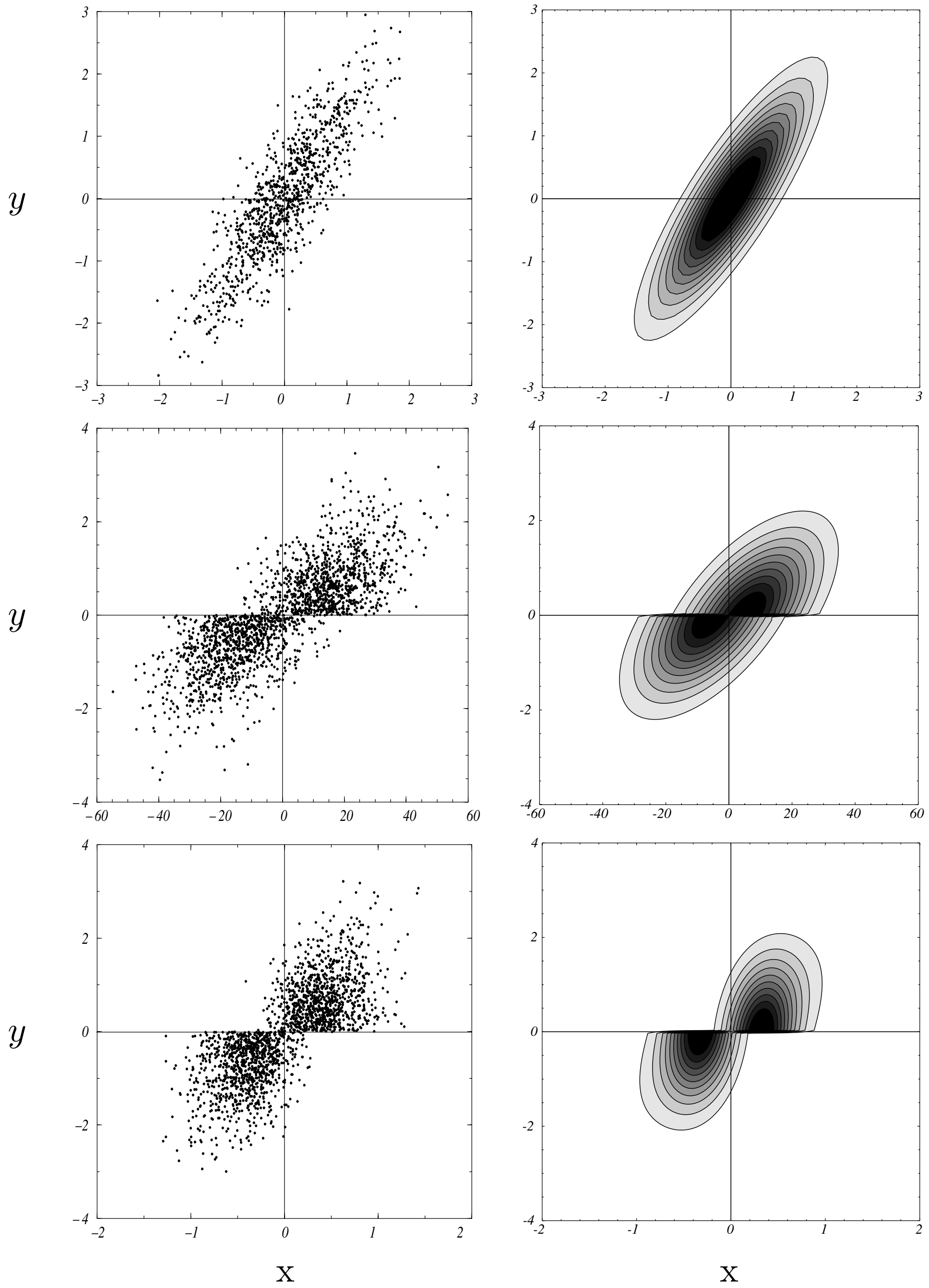

Figure 4. Stationary local field distributions $p(x, y)$ for an infinite ($\alpha = \infty$) training set after perceptron learning (top), and two finite ($\alpha = 1$) training sets after Hebbian learning (middle), or perceptron learning (bottom). The left side has simulation results, the right side contains theoretical predictions in the form of contour plots. The infinite training set yields a joint Gaussian distribution, the Hebbian rule gives only a conditionally Gaussian $p(x \mid y)$ and the perceptron rule deviates even further from the Gaussian shape. Learning rates are $\eta = 1$ and decay coefficients are $\gamma = 0.1$ in all three graphs.

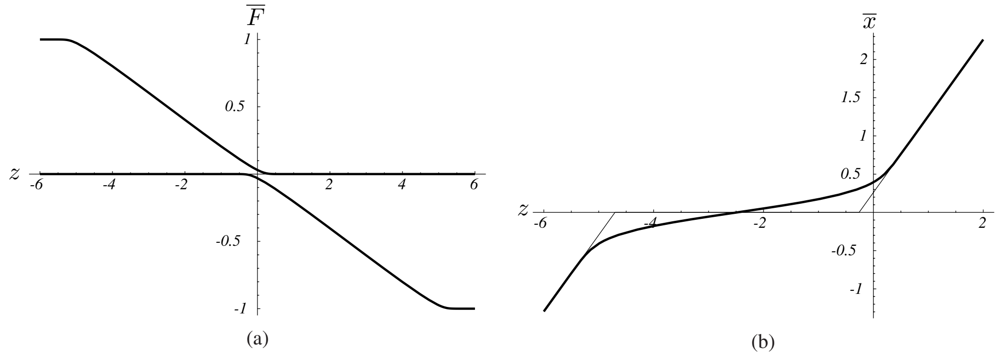

Figure 5. (a) The relationship between $\bar{F}$ (see equation (57)) and z for $y = 1$ (upper line) and $y = -1$ (lower line) for the stationary state of a perceptron with $\alpha = 1$ and $\gamma = 0.1$. The width of the sloping part is close to $\eta g/\alpha$, the abcissas are near $-yR$. (b) The time average of x_t as a function of z, given $y = 1$. The abcissas of the thin straight lines are near $-yR - \eta g/\alpha$ and $-yR$. Due to the Gaussian measure of z centred at the origin, the part close to the x-axis is the main contribution.

perceptron learning rule, this is no longer true. The random variables y and z are independently distributed Gaussian variables. From (56) we find that

$$\bar{F} = \frac{1}{2}\text{sgn}(y) - \frac{1}{2}\text{erf}\left(\frac{yR + z + \frac{\eta g}{\alpha}\bar{F}}{\sqrt{2(D-d)}}\right). \tag{57}$$

Samples of the (y, z) statistics for $\alpha = 1$ and $\gamma = 0.1$ of $\bar{F}$ as a function of z are shown in figure 5(a) for $y < 0$ (top) and $y > 0$ (bottom). The width of the sloping segment is $\eta g/\alpha$, while the size of $\sqrt{D-d}$ determines the rounding at the edges. The value of $\bar{x}$ corresponding to $y = 1$ as a function of the Gaussian disorder z is drawn in figure 5(b). For y positive and roughly $z > -yR$, one has $\bar{x} = yR + z$, whereas for $z < -yR - \eta g/\alpha$ one finds $\bar{x} = yR + z + \eta g/\alpha$. For z in the range $-yR - \eta g/\alpha < z < -yR$, we find $\bar{x} \approx 0$. In this particular example (using the same values for the order parameters as the graphs shown in figure 4(c)) $\sqrt{d} \approx 0.27$ so that the Gaussian measure confines z close to the origin. Thus the resulting local field distribution is distinctly non-Gaussian as shown in figure 4(c).

6. Conclusion

In this paper, we have studied the statics and dynamics of an ensemble of students learning on-line the classification of a large number of examples. This problem boils down to solving a large number of coupled stochastic difference equations, each corresponding to a single input channel. The situation is complicated by the existence of disorder in the form of the composition of the training set. Using the generating function method we have transformed this Markovian system of N coupled equations in the limit of $N \to \infty$ into an effective single pattern process. The price paid for this reduction is that the new process has noise which is correlated in time and the presence of a retarded self-interaction in the system, which make the dynamics non-Markovian. In principle, it is possible to calculate the evolution of the system analytically, but in general it will be impossible to pursue this after the very first few time steps. However, the process can be solved numerically up to an arbitrary precision.

Our calculation provides a solid basis for the further analytical study of linear rules. For non-linear rules the importance of our exact macroscopic dynamical equations is mainly in the insight they can provide into the behaviour of different learning rules and the possibility they create to study and solve stationary states of both on-line and batch, gradient and non-gradient learning. Until now, the stationary states of these kinds of learning processes have only been directly accessable with tools from equilibrium statistical mechanics, requiring detailed balance. This confined the analyses to batch gradient-descent learning. This restriction has now been lifted. From our macroscopic evolution equations we can extract the stationary state equations very easily if we assume time-translation invariance and the absence of anomalous response. We have not yet addressed the issue where this is likely to hold for on-line learning. To reduce the time-dependent order parameters such as the student autocorrelation and the student response to a finite set of scalar order parameters, we apply a method we know from similar spin-glass problems based on the detachment of single-time and persistent order parameters from the non-persistent ones. The procedure consists of removing all non-persistent parts of the order parameters (except for the single-time quantities), retaining only a small closed set of equations containing just four (Q, R, d, g) scalar macroscopic order parameters. Whether this last procedure is indeed exact, remains to be seen and will be the subject of a future study, but the numerical evidence clearly suggests that the underlying assumption holds.

References

[1] Kinzel W and Rujan P 1990 Improving a network generalization ability by selecting examples *Europhys. Lett.* **13** 473

[2] Biehl M and Schwarze H 1992 On-line learning of a time-dependent rule *Europhys. Lett.* **20** 773

[3] Biehl M and Riegler P 1994 On-line learning with a perceptron *Europhys. Lett.* **28** 525

[4] Rae H C, Heimel J A F and Coolen A C C 2000 Non-deterministic learning dynamics in large neural networks due to structural data bias *J. Phys. A: Math. Gen.* **33** 8703

[5] Hertz J A, Krogh A and Thorbergsson G I 1989 Phase transitions in simple learning *J. Phys. A: Math. Gen.* **22** 2133

[6] Rae H C, Sollich P and Coolen A C C 1999 On-line learning with restricted training sets: An exactly solvable case *J. Phys. A: Math. Gen.* **32** 3321

[7] Sollich P and Barber D 1998 Online learning from finite training sets and robustness to input bias *Neural Computation* **10** 2201

[8] Heskes T M and Kappen B 1991 Learning processes in neural networks *Phys. Rev.* A **44** 2718

[9] Hansen L K, Pathria R and Salamon P 1993 Stochastic dynamics of supervised learning *J. Phys. A: Math. Gen.* **26** 63

[10] Radons G 1993 On stochastic dynamics of supervised learning *J. Phys. A: Math. Gen.* **26** 3455

[11] Heskes T M 1994 On Fokker–Planck approximations of on-line learning processes *J. Phys. A: Math. Gen.* **27** 5145

[12] Barber D and Sollich P 1998 *On-Line Learning in Neural Networks* ed D Saad (Cambridge: Cambridge University Press) pp 279–302

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

[14] Horner H 1992 Dynamics of learning for the binary perceptron problem *Z. Phys.* B **86** 291

[15] Horner H 1992 Dynamics of learning and generalization in a binary perceptron model *Z. Phys.* B **87** 371

[16] Li S and Wong K Y M 2000 *Advances in Neural Information Processing* vol 12 ed S A Solla and T K Leen (Cambridge, MA: MIT Press)

[17] Wong K Y M, Li S and Tong Y W 2000 Many-body approach to the dynamics of batch learning *Phys. Rev.* E **62** 4036

[18] Wong K Y M, Li S and Luo P 2001 *Advanced Mean Field Methods—Theory and Practice* ed M Opper and D Saad (Cambridge, MA: MIT Press)

[19] Coolen A C C and Saad D 1998 *On-Line Learning in Neural Networks* ed D Saad (Cambridge: Cambridge University Press) pp 303–43

[20] Coolen A C C and Saad D 2000 Dynamics of learning with restricted training sets *Phys. Rev.* E **62** 5444

[21] Coolen A C C, Saad D and Xiong Y S 2000 On-line learning from restricted training sets in multilayer neural networks *Europhys. Lett.* **51** 691

[22] Mace C W H and Coolen A C C 2000 *Advances in Neural Information Processing* vol 12 ed S A Solla, T K Leen and K Müller (Cambridge, MA: MIT Press) p 237

[23] Eissfeller H and Opper M 1992 New method for studying the dynamics of disordered spin systems without finite-size effects *Phys. Rev. Lett.* **68** (13) 2094

[24] Kinouchi O and Caticha N 1992 Optimal generalization in perceptrons *J. Phys. A: Math. Gen.* **25** 6243