Statistics and Computing (1998) **8**, 55–88

Statistical mechanical analysis of the dynamics of learning in perceptrons

C. W. H. MACE and A. C. C. COOLEN

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

Received May 1997; and accepted September 1997

We describe the application of tools from statistical mechanics to analyse the dynamics of various classes of supervised learning rules in perceptrons. The character of this paper is mostly that of a cross between a biased non-encyclopaedic review and lecture notes: we try to present a coherent and self-contained picture of the basics of this field, to explain the ideas and tricks, to show how the predictions of the theory compare with (simulation) experiments, and to bring together scattered results. Technical details are given explicitly in an appendix. In order to avoid distraction we concentrate the references in a final section. In addition this paper contains some new results: (i) explicit solutions of the macroscopic equations that describe the error evolution for on-line and batch learning rules; (ii) an analysis of the dynamics of arbitrary macroscopic observables (for complete and incomplete training sets), leading to a general Fokker–Planck equation; and (iii) the macroscopic laws describing batch learning with complete training sets. We close the paper with a preliminary exposé of ongoing research on the dynamics of learning for the case where the training set is incomplete (i.e. where the number of examples scales linearly with the network size).

Keywords: Learning dynamics, generalization, statistical mechanics

Contents

1. Introduction

1.1. *Supervised learning in neural networks*

In this paper we study the dynamics of supervised learning in artificial neural networks. The basic scenario is as follows. A 'student' neural network executes a certain known operation $S : D \rightarrow R$, which is parameterized by a vector $\boldsymbol{J}$, usually representing synaptic weights and/or neuronal thresholds. Here D denotes the set of all possible 'questions' and R denotes the set of all possible 'answers'. The student is being trained to emulate a given 'teacher', which executes some as yet unknown operation $T : D \rightarrow R$. In order to achieve the objective, the student network S tries to improve its performance gradually by adapting its parameters $\boldsymbol{J}$ according to an iterative procedure, using only examples of input vectors (or 'questions') $\boldsymbol{\xi} \in \mathscr{R}^N$ which are drawn at random from a fixed training set $\tilde{D} \subseteq D$ of size $|\tilde{D}|$, and the corresponding values of the teacher outputs $T(\boldsymbol{\xi})$ (the 'correct answers'). The iterative procedure (the 'learning rule') is not allowed to involve any further knowledge of the operation T. As far as the student is concerned the teacher is an 'oracle', or 'black box'; the only information available about the inner workings of the black box is contained in the various answers, $T(\boldsymbol{\xi})$, it provides (see Fig. 1). For simplicity we will assume each 'question' $\boldsymbol{\xi}$ to be equally likely to occur, and we will consider only 'perfect' (noise- free) teachers (generalization of what follows to cases where the questions $\boldsymbol{\xi}$ carry non-uniform probabilities or probability densities $p(\boldsymbol{\xi})$, or where the teachers have themselves a non-zero probability of giving an incorrect answer is straightforward).

We will consider the following two classes of learning rules, i.e. of recipes for the iterative modification of the student's control parameters $\boldsymbol{J}$, which we will refer to as on-line learning rules and batch learning rules, respectively:

$$\begin{array}{ll}\text{On-Line:} & \boldsymbol{J}(t+1) = \boldsymbol{J}(t) + \boldsymbol{F}[\boldsymbol{\xi}(t), \boldsymbol{J}(t), T(\boldsymbol{\xi}(t))] \\ \text{Batch:} & \boldsymbol{J}(t+1) = \boldsymbol{J}(t) + \langle \boldsymbol{F}[\boldsymbol{\xi}, \boldsymbol{J}(t), T(\boldsymbol{\xi})]\rangle_{\tilde{D}}\end{array} \tag{1}$$

The integer variable $t = 0, 1, 2, 3, \ldots$ labels the iteration steps. In the case of on-line learning an input vector $\boldsymbol{\xi}(t)$ is drawn independently at each iteration step from the training set $\tilde{D}$, followed by a modification of the control parameters $\boldsymbol{J}$. Therefore this process is stochastic (Markovian). In the case of batch learning the modification that would have been made in the on-line version is averaged over the input vectors in the training set $\tilde{D}$, at each iteration step. This process is therefore a deterministic iterative map.* Both rules in (1) can formally be written in the general form of a Markovian stochastic process. We introduce the probability density $\hat{p}_t(\boldsymbol{J})$ to find parameter vector $\boldsymbol{J}$ at discrete iteration step t. In terms of this microscopic probability density the processes (1) can be written as:

$$\hat{p}_{t+1}(\boldsymbol{J}) = \int \mathrm{d}\boldsymbol{J}' W[\boldsymbol{J}; \boldsymbol{J}'] \hat{p}_t(\boldsymbol{J}') \tag{2}$$

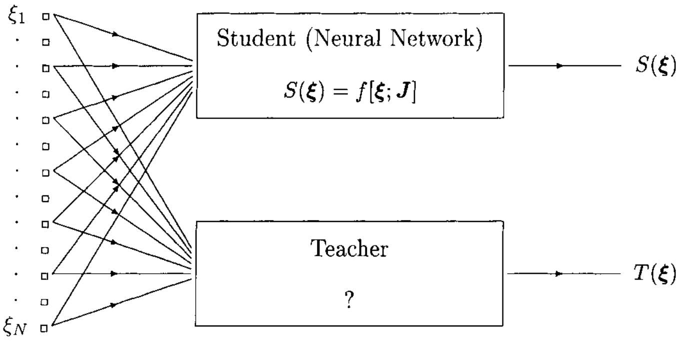

Fig. 1. *The general scenario of supervised learning: a 'student network' S, executing the parametrized operation $S(\boldsymbol{\xi}) = f[\boldsymbol{\xi}; \boldsymbol{J}]$, is being 'trained' to perform an operation $T : D \rightarrow R$ by updating its control parameters $\boldsymbol{J}$ according to an iterative procedure, the 'learning rule'. This rule is allowed to make use only of examples of 'question/answer pairs' $(\boldsymbol{\xi}, T(\boldsymbol{\xi}))$, where $\boldsymbol{\xi} \in \tilde{D} \subseteq D$. The actual 'teacher operation' T that generated the answers $T(\boldsymbol{\xi})$, on the other hand, cannot be observed directly. The goal is to arrive at a situation where $S(\boldsymbol{\xi}) = T(\boldsymbol{\xi})$ for all $\boldsymbol{\xi} \in D$*

with the transition probability densities

$$\begin{array}{ll}\text{On-Line:} & W[\boldsymbol{J}; \boldsymbol{J}'] = \langle \delta\{\boldsymbol{J} - \boldsymbol{J}' - \boldsymbol{F}[\boldsymbol{\xi}, \boldsymbol{J}', T(\boldsymbol{\xi})]\}\rangle_{\tilde{D}} \\ \text{Batch:} & W[\boldsymbol{J}; \boldsymbol{J}'] = \delta\{\boldsymbol{J} - \boldsymbol{J}' - \langle \boldsymbol{F}[\boldsymbol{\xi}, \boldsymbol{J}', T(\boldsymbol{\xi})]\rangle_{\tilde{D}}\}\end{array} \tag{3}$$

(in which $\delta[z]$ denotes the delta-distribution). The advantage of using the on-line version of the learning rule is a reduction in the amount of calculations that have to be done at each iteration step; the price paid for this reduction is the presence of fluctuations, with as yet unknown impact on the performance of the system.

We will denote averages over the probability density $\hat{p}_t(\boldsymbol{J})$, averages over the full set D of possible input vectors and averages over the training set $\tilde{D}$ in the following way:

$$\langle g(\boldsymbol{J})\rangle = \int \mathrm{d}\boldsymbol{J} \hat{p}_t(\boldsymbol{J}) g(\boldsymbol{J}) \qquad \langle K(\boldsymbol{\xi})\rangle_D = \frac{1}{|D|}\sum_{\boldsymbol{\xi} \in D} K(\boldsymbol{\xi})$$

$$\langle K(\boldsymbol{\xi})\rangle_{\tilde{D}} = \frac{1}{|\tilde{D}|}\sum_{\boldsymbol{\xi} \in \tilde{D}} K(\boldsymbol{\xi})$$

The average $\langle K(\boldsymbol{\xi})\rangle_{\tilde{D}}$ will in general depend on the microscopic realization of the training set $\tilde{D}$. To quantify the goal and the progress of the student one finally defines an error $E[T(\boldsymbol{\xi}), S(\boldsymbol{\xi})] = E[T(\boldsymbol{\xi}), f[\boldsymbol{\xi}; \boldsymbol{J}]]$, which measures the mismatch between student answers and correct (teacher) answers for individual questions. The two key quantities of interest in supervised learning are the (time-dependent) averages of this error measure, calculated over the training set $\tilde{D}$ and the full question set D, respectively:

$$\begin{array}{ll}\text{Training error:} & E_{\mathrm{t}}(\boldsymbol{J}) = \langle E[T(\boldsymbol{\xi}), f[\boldsymbol{\xi}; \boldsymbol{J}]]\rangle_{\tilde{D}} \\ \text{Generalization error:} & E_{\mathrm{g}}(\boldsymbol{J}) = \langle E[T(\boldsymbol{\xi}), f[\boldsymbol{\xi}; \boldsymbol{J}]]\rangle_D\end{array} \tag{4}$$

These quantities are stochastic observables, since they are functions of the stochastically evolving vector $\boldsymbol{J}$. Their expectation values over the stochastic process (2) are given by

[*] Clearly one could define an infinite number of intermediate classes of learning rules (e.g. learning with 'momentum'); the present two are just the extreme cases. Note also that the term 'batch' unfortunately means different things to different scientists. The definition used here is sometimes described as 'off-line'.

$$\text{Mean training error:} \quad \langle E_{\mathrm{t}}\rangle = \langle\langle E[T(\xi), f[\xi; \boldsymbol{J}]]\rangle_{\tilde{D}}\rangle$$
$$\text{Mean generalization error:} \quad \langle E_{\mathrm{g}}\rangle = \langle\langle E[T(\xi), f[\xi; \boldsymbol{J}]]\rangle_{D}\rangle \tag{5}$$

Note that the prefix 'mean' refers to the stochasticity in the vector $\boldsymbol{J}$; both $\langle E_{\mathrm{t}}\rangle$ and $\langle E_{\mathrm{g}}\rangle$ will in general still depend on the realization of the training set $\tilde{D}$.

The training error measures the performance of the student on the questions it could have been confronted with during the learning stage (in the case of on-line learning the student need not have seen all of them). The generalization error measures the student's performance on the full question set and its minimization is therefore the main target of the process. The quality of a theory describing the dynamics of supervised learning can be measured by the degree to which it succeeds in predicting the values of $\langle E_{\mathrm{t}}\rangle$ and $\langle E_{\mathrm{g}}\rangle$ as a function of the iteration time t and for arbitrary choices made for the function $F[\ldots]$ that determines the details of the learning rules (1).

1.2. *Statistical mechanics and its applicability*

Statistical mechanics deals with large systems of stochastically interacting microscopic elements (particles, magnets, polymers, etc.). The general strategy of statistical mechanics is to abandon any ambition to solve models of such systems at the microscopic level of individual elements, but to use the microscopic laws to calculate laws describing the behaviour of a suitably chosen set of *macroscopic* observables. The toolbox of statistical mechanics consists of various methods and tricks to perform this reduction from the microscopic to a macroscopic level, which are based on clever ways to do the bookkeeping of probabilities. The experience and intuition that has been built up over the last century tells us what to expect (e.g. phase transitions), and serves as a guide in choosing the macroscopic observables and in seeing the difference between relevant mathematical subtleties and irrelevant ones. As in any statistical theory, clean and transparent mathematical laws can be expected to emerge only for large (preferably infinitely large) systems.

Supervised learning processes as described in the previous subsection appear to meet the criteria for statistical mechanics to apply, provided we are happy to restrict ourselves to large systems ($N \to \infty$). Here the microscopic stochastic dynamical variables are the components of the vector $\boldsymbol{J}$, and one is as little interested in knowing all individual components of $\boldsymbol{J}$ as one would be in knowing all position coordinates of the molecules in a bucket of water. We are, rather, after the generalization and training errors, which are indeed *macroscopic* observables.

Further support for the applicability of statistical mechanics is provided by numerical simulations. Consider, for instance, the example of the ordinary perceptron learning rule. For simplicity we choose $\tilde{D} = D = \{-1, 1\}^N$, with a task T generated by a 'teacher perceptron' corresponding to the following choices in the language of the previous subsection:

$$S(\xi) = \mathrm{sgn}(\boldsymbol{J}\cdot\xi) \qquad T(\xi) = \mathrm{sgn}(\boldsymbol{B}\cdot\xi)$$

with $\boldsymbol{J}, \boldsymbol{B} \in \mathscr{R}^N$. The teacher weight vector $\boldsymbol{B}$ is chosen at random, and normalized according to $|\boldsymbol{B}| = 1$. The (on-line) perceptron learning rule is

$$\boldsymbol{J}(t+1) = \boldsymbol{J}(t) + \xi(t)T(\xi(t))\theta[-T(\xi(t))(\boldsymbol{J}(t)\cdot\xi(t))]$$

with the step function $\theta[z > 0] = 1$, $\theta[z < 0] = 0$. An educated guess for a possibly relevant macroscopic observable is the object that also played a central role in the original perceptron convergence proof (Minsky and Papert, 1969): $\omega(t) = \boldsymbol{J}(t)\cdot\boldsymbol{B}/|\boldsymbol{J}(t)|$. The result of measuring the value of $\omega(t)$ during the execution of the above learning rule is shown in Fig. 2. These experiments clearly suggest (keeping

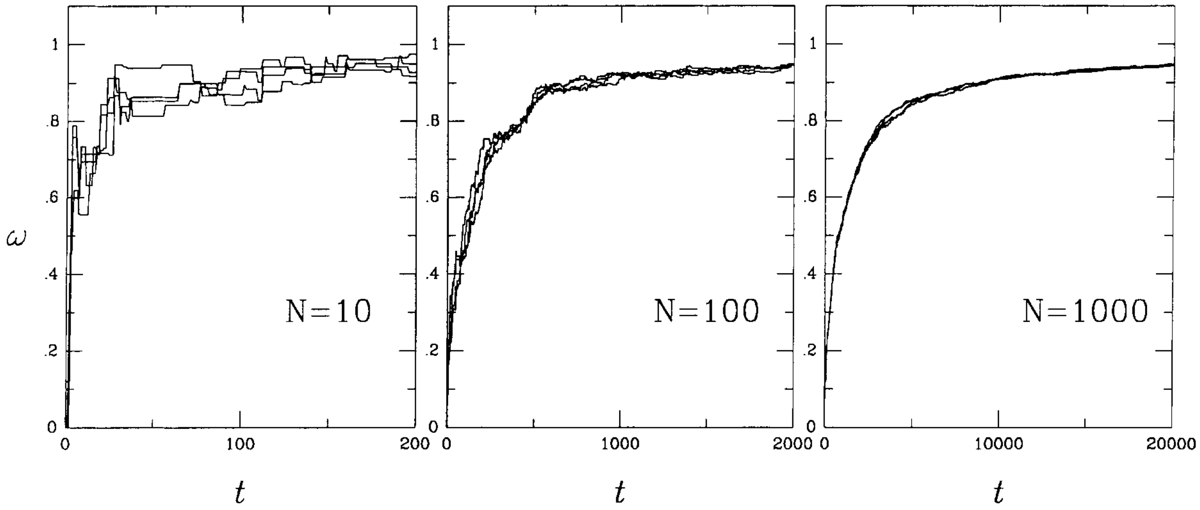

Fig. 2. *Evolution in time of the observable $\omega = \boldsymbol{J}\cdot\boldsymbol{B}/|\boldsymbol{J}|$ during numerical simulations of the standard perceptron learning rule (with a randomly drawn normalized teacher weight vector $\boldsymbol{B}$), following random initializations of the student weight vector $\boldsymbol{J}$*

in mind that for specifically constructed pathological teacher vectors the picture might be different), that if viewed on the relevant N-dependent time-scale (as in the figure), the fluctuations in ω become negligible as $N \to \infty$, and a clean deterministic law emerges. This is the type of situation we need in order to use statistical mechanics, and finding an analytical expression for this deterministic law will be our goal.

As a second example we will choose a two-layer network, trained according to the error back-propagation rule. Here the microscopic stochastic variables are both the weights $\{W_{ij}\}$ from the input to the hidden layer (of L neurons) and the weights $\{J_i\}$ from the hidden layer to the output layer. We define $\tilde{D} = D = \{-1, 1\}^N$ and

$$S(\xi) = \tanh\left[\sum_{i=1}^{L} J_i y_i(\xi)\right] \qquad y_i(\xi) = \tanh\left[\sum_{j=1}^{N} W_{ij}\xi_j\right]$$

We consider two types of tasks, a linearly separable one (which is learnable by the present student to any required accuracy), and the parity operation (which is learnable only for $L \geq N$):

$$\xi \in \{-1,1\}^N: \quad \begin{array}{ll} \text{task I:} & T(\xi) = \text{sgn}(\boldsymbol{B}\cdot\xi) \in \{-1,1\} \\ \text{task II:} & T(\xi) = \prod_{i=1}^{N} \xi_i \in \{-1,1\} \end{array}$$

Our macroscopic observable will be the mean error (since $\tilde{D} = D$ the generalization error and the training error are here identical):

$$E = \langle E[T(\xi), S(\xi)]\rangle_D, \quad E[T(\xi), S(\xi)] = \tfrac{1}{2}[T(\xi) - S(\xi)]^2$$

Perfect performance would correspond to $E = 0$. On the other hand, a trivial perceptron with zero weights throughout would give $S(\xi) = 0$ so $E = \frac{1}{2}\langle T^2(\xi)\rangle_D = \frac{1}{2}$. The learning rule used is the discretized on-line version (with learning rate ϵ) of the error backpropagation rule:

$$J_i(t+\epsilon) = J_i(t) - \epsilon \frac{\partial}{\partial J_i} E[T(\xi(t)), S(\xi(t))]$$

$$W_{ij}(t+\epsilon) = W_{ij}(t) - \epsilon \frac{\partial}{\partial W_{ij}} E[T(\xi(t)), S(\xi(t))]$$

The results of doing several such simulations, for $N = 15$ and $L = 10$ (so that the parity operation is an unlearnable task for the student network) and following random initialization of the various weights, are shown in Fig. 3. These experiments again clearly suggest that for multi-layer networks also, statistical mechanics will be a natural tool to analyse the dynamics of learning. Provided we scale our parameters appropriately and take a suitable limit (there will be different equivalent ways of doing this), the fluctuations in suitably chosen macroscopic observables can be made to vanish, such that transparent deterministic laws emerge.

1.3. *A preview*

There are two main classes of situations in the supervised learning arena, which differ fundamentally in their dynamics and in the degree to which we can analyse them mathematically. The first class is the one where the training set $\tilde{D}$ is what we call 'complete': sufficiently large and sufficiently diverse to lead to a learning dynamics which in the limit $N \to \infty$ is identical to that of the situation where $\tilde{D} = D$. For example: in single perceptrons and in multilayer perceptrons with a finite number of hidden nodes one finds, for the case where $D = \{-1, 1\}^N$ and where the members of the training set $\tilde{D}$ are drawn at random from D, that completeness of the training set amounts to

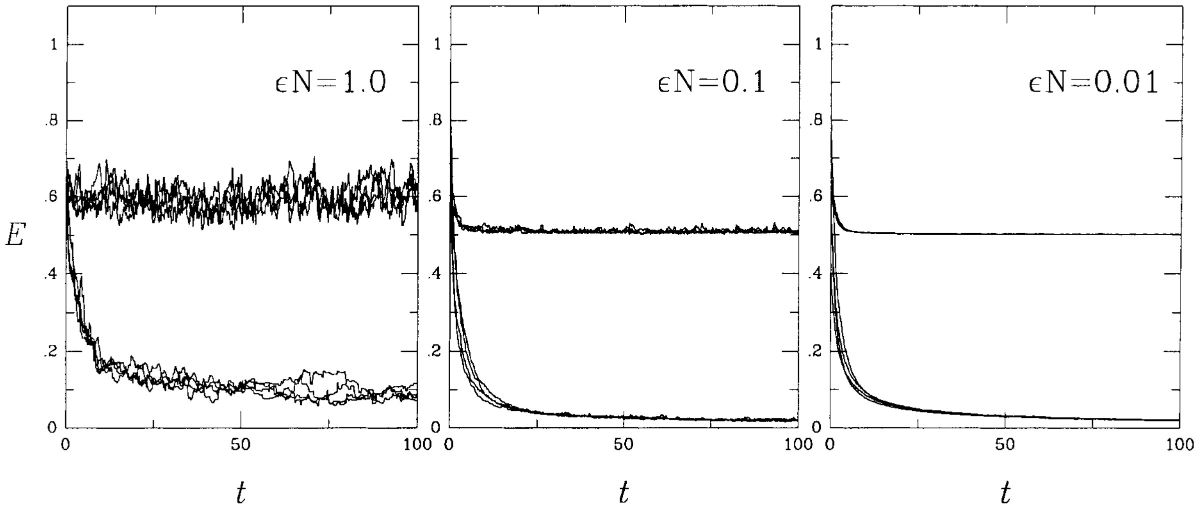

Fig. 3. *Evolution of the overall error E in a two-layer feed-forward network, trained by error backpropagation (with N =15 input neurons, L = 10 hidden neurons, and a single output neuron). The results refer to independent experiments involving either a linearly separable task (with random teacher vector, lower curves) or the parity operation (upper curves), following random initialization*

$\lim_{N\to\infty} N/|\tilde{D}| = 0$. This makes sense: it means that for $N \to \infty$ there will be an infinite number of training examples per degree of freedom. For this class of models it is fair to say that the dynamics of learning can be fully analysed in a reasonably simple way.* Because this situation is now so nicely under control and admits for analytical solutions, it is a nice area to describe in a self-contained way in a paper such as the present one. We will restrict ourselves to single perceptrons with various types of learning rules, since they form the most transparent playground for explaining how the mathematical techniques work. For multi-layer perceptrons with a finite number of hidden neurons and complete training sets the procedure to be followed is very similar.‡

The picture changes dramatically if we move away from complete training sets and consider those where the number of training examples is proportional to the number of degrees of freedom, i.e. in simple perceptrons and in two-layer perceptrons with a finite number of hidden neurons this implies $|\tilde{D}| = \alpha N$ $(0 < \alpha < \infty)$. Now the dependence of the microscopic variables $\boldsymbol{J}$ on the realization of the training set $\tilde{D}$ is non-negligible. However, if the questions in the training set are drawn at random from the full question set D one often finds that in the $N \to \infty$ limit the values of the *macroscopic* observables only depend on the size $|\tilde{D}|$ of the training set, not on its microscopic realization. For those familiar with the statistical mechanical analysis of the operation of recurrent neural networks: learning dynamics with complete training sets is mathematically similar to the dynamics of attractor networks away from saturation, whereas learning dynamics with incomplete training sets is similar, if non-equivalent, to the dynamics of attractor networks close to saturation (in turn equivalent to the complex dynamics of spin-glasses). Here one needs much more powerful mathematical tools, which are as yet only partly available. This class of problems is therefore only beginning to be studied, and we cannot yet give a well-rounded overview with a happy ending (as for the case of complete training sets). We will do the next best thing: we try to explain as clearly as possible what the problem is and describe (in an inevitably more condensed way) which techniques are currently being employed to solve it.

No review is unbiased and complete and one always has to strike a balance between broadness and depth (equivalently: between being encyclopaedic and being self-contained). Here we have opted for the latter. As a result, the references we give are intended to serve as a guide only, not as a true reflection of all the work that has been done; for each paper mentioned at least 50 will have been left out, and we wish to apologise beforehand to the authors of the papers in the latter category. We aim to explain the ideas and techniques only for a subset of the field, in the hope that the text can then be sufficiently self-contained to serve not just the interested spectator but also those who wish to become actively involved.

2. On-line learning: complete training sets and explicit rules

We will now derive explicitly macroscopic dynamical equations that describe the evolution in time for the error in large perceptrons, trained with several on-line learning rules to perform linearly separable tasks. In this section we restrict ourselves* to complete training sets $\tilde{D} = D = \{-1, 1\}^N$. There is consequently no difference between training and generalization error, and we can simply define $E = \lim_{N\to\infty}\langle E_{\mathrm{g}}\rangle = \lim_{N\to\infty}\langle E_{\mathrm{t}}\rangle$.

2.1. *General on-line learning rules*

Consider a linearly separable binary classification task $T : \{-1, 1\}^N \to \{-1, 1\}$. It can be regarded as generated by a teacher perceptron with some unknown weight vector $\boldsymbol{B} \in \mathscr{R}^N$, i.e. $T(\boldsymbol{\xi}) = \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})$, normalized according to $|\boldsymbol{B}| = 1$ (with the sign function $\mathrm{sgn}(z > 0) = 1$, $\mathrm{sgn}(z < 0) = -1$). A student perceptron with output $S(\boldsymbol{\xi}) = \mathrm{sgn}(\boldsymbol{J}\cdot\boldsymbol{\xi})$ (where $\boldsymbol{J} \in \mathscr{R}^N$) is being trained in an on-line fashion using randomly drawn examples of input vectors $\boldsymbol{\xi} \in \{-1, 1\}^N$ with corresponding teacher answers $T(\boldsymbol{\xi})$. The general picture of Fig. 1 thus specializes to Fig. 4. We exploit our knowledge of the perceptron's scaling properties (see Fig. 2) and distinguish between the discrete time unit in terms of iteration steps, from now on to be denoted by $\mu = 1, 2, 3, \ldots$, and the scale-invariant time unit $t_\mu = \mu/N$. Our goal is to derive well-behaved differential equations in the limit $N \to \infty$, so we require weight changes occurring in intervals $\Delta t = 1/N$ to be of order $\mathcal{O}(1/N)$ as well. In terms of Equation 1 this implies that $F[\ldots] = \mathcal{O}(1/N)$. In view of $J_i = \mathcal{O}(N^{-\frac{1}{2}})$, the changes made in each single iteration step to the components of the weight vector are small relative to their values. If, finally, we restrict ourselves to those rules where weight changes are made in the direction of the example vectors (which includes most popular rules), we obtain the generic‡ recipe

* In these models one can still study interesting new phenomena, such as the effects of having noisy teachers etc., but at least the route to be followed to solve them is well-defined and guaranteed to work.

‡ The situation is different if we try to deal with multi-layer perceptrons with a number of hidden neurons which scales linearly with the number of input channels N. As far as we are aware, this still poses an unsolved problem, even in the case of complete training sets.

* The binary nature of the input vectors is of no particular relevance. The methods described in this paper will apply equally well to the case where the input vectors are real-valued, and will give similar results.

‡ One can obviously write down more general rules, and also write the present recipe (6) in different ways.

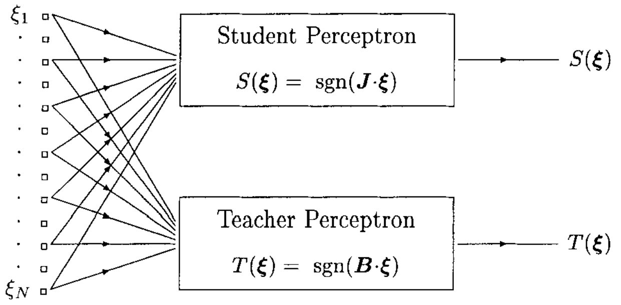
ξ_1
ξ_N
Student Perceptron
$S(\boldsymbol{\xi}) = \mathrm{sgn}(\boldsymbol{J}\cdot\boldsymbol{\xi})$
$S(\boldsymbol{\xi})$
Teacher Perceptron
$T(\boldsymbol{\xi}) = \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})$
$T(\boldsymbol{\xi})$

Fig. 4. *A student perceptron S is being trained according to on-line learning rules to perform a linearly separable operation, generated by some unknown teacher perceptron T*

$$\boldsymbol{J}\left(t_\mu + \frac{1}{N}\right) = \boldsymbol{J}(t_\mu) + \frac{1}{N}\eta(t_\mu)\boldsymbol{\xi}^\mu \ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu) \times \mathscr{F}\left[|\boldsymbol{J}(t_\mu)|; \boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu, \boldsymbol{B}\cdot\boldsymbol{\xi}^\mu\right] \tag{6}$$

Here $\eta(t_\mu)$ denotes a (possibly time-dependent) learning rate and $\boldsymbol{\xi}^\mu$ is the input vector selected at iteration step μ. $\mathscr{F}[\ldots]$ is an as yet arbitrary function of the length of the student weight vector and of the local fields u and v of student and teacher (note: $\mathscr{F}$ can depend on the sign of the teacher field only, not on its magnitude). For example, for $\mathscr{F}[J;u,v] = 1$ we obtain a Hebbian rule, for $\mathscr{F}[J;u,v] = \theta[-uv]$ we obtain the perceptron learning rule, etc.

We now try to solve the dynamics of the learning process in terms of the two macroscopic observables that play a special role in the perceptron convergence proof (Minsky and Papert, 1969):

$$Q[\boldsymbol{J}] = \boldsymbol{J}^2 \qquad R[\boldsymbol{J}] = \boldsymbol{J}\cdot\boldsymbol{B} \tag{7}$$

(at this stage the selection of observables is still no more than intuition-driven guesswork). The formal approach would now be to derive an expression for the (time-dependent) probability density $P(Q,R) = \langle\delta[Q - Q[\boldsymbol{J}]]\ \delta[R - R[\boldsymbol{J}]]\rangle$, however, it turns out that in the present case[‡] there is a short cut. Squaring (6) and taking the inner product of (6) with the teacher vector $\boldsymbol{B}$ gives, respectively

$$Q\left[\boldsymbol{J}\left(t_\mu + \frac{1}{N}\right)\right] = Q[\boldsymbol{J}(t_\mu)] + \frac{2}{N}\eta(t_\mu)(\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu) \times \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\mathscr{F}\left[|\boldsymbol{J}(t_\mu)|; \boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu, \boldsymbol{B}\cdot\boldsymbol{\xi}^\mu\right] + \frac{1}{N}\eta^2(t_\mu)\mathscr{F}^2\left[|\boldsymbol{J}(t_\mu)|; \boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu, \boldsymbol{B}\cdot\boldsymbol{\xi}^\mu\right]$$

$$R\left[\boldsymbol{J}\left(t_\mu + \frac{1}{N}\right)\right] = R[\boldsymbol{J}(t_\mu)] + \frac{1}{N}\eta(t_\mu)|\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu| \times \mathscr{F}\left[|\boldsymbol{J}(t_\mu)|; \boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu, \boldsymbol{B}\cdot\boldsymbol{\xi}^\mu\right]$$

(note: $\boldsymbol{\xi}^\mu\cdot\boldsymbol{\xi}^\mu = N$). After ℓ discrete update steps we will have accumulated ℓ such modifications, and will thus arrive at:

$$\frac{Q[\boldsymbol{J}(t_\mu + \ell/N)] - Q[\boldsymbol{J}(t_\mu)]}{\ell/N} = \frac{1}{\ell}\sum_{m=0}^{\ell-1}\left\{2\eta\left(t_\mu + \frac{m}{N}\right)\left(\boldsymbol{J}\left(t_\mu + \frac{m}{N}\right)\cdot\boldsymbol{\xi}^{\mu+m}\right)\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^{\mu+m}) \times \mathscr{F}\left[\left|\boldsymbol{J}\left(t_\mu + \frac{m}{N}\right)\right|; \boldsymbol{J}\left(t_\mu + \frac{m}{N}\right)\cdot\boldsymbol{\xi}^{\mu+m}, \boldsymbol{B}\cdot\boldsymbol{\xi}^{\mu+m}\right] + \eta^2\left(t_\mu + \frac{m}{N}\right) \times \mathscr{F}^2\left[\left|\boldsymbol{J}\left(t_\mu + \frac{m}{N}\right)\right|; \boldsymbol{J}\left(t_\mu + \frac{m}{N}\right)\cdot\boldsymbol{\xi}^{\mu+m}, \boldsymbol{B}\cdot\boldsymbol{\xi}^{\mu+m}\right]\right\}$$

$$\frac{R\left[\boldsymbol{J}(t_\mu + \ell/N)\right] - R[\boldsymbol{J}(t_\mu)]}{\ell/N} = \frac{1}{\ell}\sum_{m=0}^{\ell-1}\left\{\eta\left(t_\mu + \frac{m}{N}\right)|\boldsymbol{B}\cdot\boldsymbol{\xi}^{\mu+m}| \times \mathscr{F}\left[\left|\boldsymbol{J}\left(t_\mu + \frac{m}{N}\right)\right|; \boldsymbol{J}\left(t_\mu + \frac{m}{N}\right)\cdot\boldsymbol{\xi}^{\mu+m}, \boldsymbol{B}\cdot\boldsymbol{\xi}^{\mu+m}\right]\right\}$$

All is still exact, but at this stage we will have to make an assumption which is not entirely satisfactory.[*] We assume that $\boldsymbol{J}(t_\mu + (m/N))\cdot\boldsymbol{\xi}^{\mu+m} \to \boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^{\mu+m}$ if $N\to\infty$ for finite m. This is only true in a probabilistic sense, since, although $J_i(t_\mu + (m/N)) = J_i(t_\mu) + \mathcal{O}(m/N)$, the inner product is a sum of N terms. If for now, however, we accept this step and also choose learning rates which vary sufficiently slowly over time to guarantee existence of the limit $\lim_{N\to\infty}\eta(t_\mu)$, we find that by taking the limit $N\to\infty$, followed by the limit $\ell\to\infty$, three pleasant simplifications occur: (i) the time unit $t_\mu = \mu/N$ becomes a continuous variable; (ii) the left-hand sides of the above equations for the evolution of the observables Q and R become temporal derivatives; and (iii) the summations in the right-hand sides of these equations become averages of the training set. Upon putting $Q(t) = Q[\boldsymbol{J}(t)]$ and $R(t) = R[\boldsymbol{J}(t)]$ the result can be written as:

$$\frac{\mathrm{d}}{\mathrm{d}t}Q(t) = 2\eta(t)\left\langle(\boldsymbol{J}(t)\cdot\boldsymbol{\xi})\,\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}\left[Q^{\frac{1}{2}}(t); \boldsymbol{J}(t)\cdot\boldsymbol{\xi}, \boldsymbol{B}\cdot\boldsymbol{\xi}\right]\right\rangle_{\tilde{D}} + \eta^2(t)\langle\mathscr{F}^2[Q^{\frac{1}{2}}(t); \boldsymbol{J}(t)\cdot\boldsymbol{\xi}, \boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde{D}}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}R(t) = \eta(t)\left\langle|\boldsymbol{B}\cdot\boldsymbol{\xi}|\mathscr{F}\left[Q^{\frac{1}{2}}(t); \boldsymbol{J}(t)\cdot\boldsymbol{\xi}, \boldsymbol{B}\cdot\boldsymbol{\xi}\right]\right\rangle_{\tilde{D}}$$

The only dependence of the right-hand sides of these expressions on the microscopic variables $\boldsymbol{J}$ is via the student fields[‡] $\boldsymbol{J}(t)\cdot\boldsymbol{\xi} = Q^{\frac{1}{2}}(t)\hat{\boldsymbol{J}}(t)\cdot\boldsymbol{\xi}$, with $\hat{\boldsymbol{J}} = \boldsymbol{J}/|\boldsymbol{J}|$. We therefore define the stochastic variables $x = \hat{\boldsymbol{J}}\cdot\boldsymbol{\xi}$ and $y = \boldsymbol{B}\cdot\boldsymbol{\xi}$ and their joint probability distribution $P_t(x,y)$:

[‡] This will be different in the case of incomplete training sets.

[*] We will later find out that a more careful probabilistic analysis gives the same results.

[‡] This property of course depends crucially on our choice (6) made for the form of the learning rules.

$$P_t(x,y) = \langle \delta[x - \hat{\boldsymbol{J}}(t)\cdot\boldsymbol{\xi}]\delta[y - \boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde{D}} \qquad \langle f(x,y)\rangle = \int \mathrm{d}x\,\mathrm{d}y P_t(x,y) f(x,y) \tag{8}$$

Using angular brackets without subscripts for joint field averages cannot cause confusion, since such expressions always *replace* averages over $\boldsymbol{J}$, rather than occur simultaneously. Our previous result now takes the form

$$\frac{\mathrm{d}}{\mathrm{d}t}Q(t) = 2\eta(t)Q^{\frac{1}{2}}(t)\left\langle x\,\mathrm{sgn}\,(y)\mathscr{F}\left[Q^{\frac{1}{2}}(t);Q^{\frac{1}{2}}(t)x,y\right]\right\rangle + \eta^2(t)\left\langle\mathscr{F}^2\left[Q^{\frac{1}{2}}(t);Q^{\frac{1}{2}}(t)x,y\right]\right\rangle \tag{9}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}R(t) = \eta(t)\left\langle |y|\mathscr{F}\left[Q^{\frac{1}{2}}(t);Q^{\frac{1}{2}}(t)x,y\right]\right\rangle \tag{10}$$

Since the operation performed by the student does not depend on the length $|\boldsymbol{J}|$ of its weight vector, and since both Q and R involve $|\boldsymbol{J}|$, it will be convenient at this stage to switch to another (equivalent) pair of observables:

$$J(t) = |\boldsymbol{J}(t)| \qquad \omega(t) = \boldsymbol{B}\cdot\hat{\boldsymbol{J}}(t) \tag{11}$$

Using the relations $\frac{\mathrm{d}}{\mathrm{d}t}Q = 2J\frac{\mathrm{d}}{\mathrm{d}t}J$ and $\frac{\mathrm{d}}{\mathrm{d}t}R = J\frac{\mathrm{d}}{\mathrm{d}t}\omega + \omega\frac{\mathrm{d}}{\mathrm{d}t}J$, and upon dropping the various explicit time arguments (for notational convenience) we then find the compact expressions

$$\frac{\mathrm{d}}{\mathrm{d}t}J = \eta\langle x\,\mathrm{sgn}(y)\mathscr{F}[J;Jx,y]\rangle + \frac{\eta^2}{2J}\langle\mathscr{F}^2[J;Jx,y]\rangle \tag{12}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\omega = \frac{\eta}{J}\langle[\,|y| - \omega x\,\mathrm{sgn}(y)]\mathscr{F}[J;Jx,y]\rangle - \frac{\omega\eta^2}{2J^2}\langle\mathscr{F}^2[J;Jx,y]\rangle \tag{13}$$

Unless we manage to express $P(x,y)$ in terms of the pair (J,ω), however, Equations 12 and 13 do not constitute a solution of our problem, since we would still be forced to solve the original microscopic dynamical equations in order to find $P(x,y)$ as a function of time and work out (12, 13).

The final stage of the argument is to assume that the joint probability distribution (8) has a Gaussian shape, since $\tilde{D} = \{-1,1\}^N$ and since all $\boldsymbol{\xi}\in\tilde{D}$ contribute equally to the average in (8). This will be true in the vast majority of cases, e.g. it is true with probability one if the vectors $\boldsymbol{J}$ and $\boldsymbol{B}$ are drawn at random from compact sets like $[-1,1]^N$, due to the central limit theorem.* If we were to choose the components ξ_i of the input vectors to be themselves independent Gaussian random variables (as opposed to binary) the joint distribution $P_t(x,y)$ would, of course, always be Gaussian. Gaussian distributions are fully specified by their first and second-order moments, which are here calculated trivially using $\langle\xi_i\rangle = 0$ and $\langle\xi_i\xi_j\rangle = \delta_{ij}$:

$$\langle x\rangle = \sum_i \hat{J}_i\langle\xi_i\rangle = 0 \qquad \langle y\rangle = \sum_i B_i\langle\xi_i\rangle = 0$$
$$\langle x^2\rangle = \sum_{ij}\hat{J}_i\hat{J}_j\langle\xi_i\xi_j\rangle = 1 \qquad \langle y^2\rangle = \sum_{ij}B_iB_j\langle\xi_i\xi_j\rangle = 1$$
$$\langle xy\rangle = \sum_{ij}\hat{J}_iB_j\langle\xi_i\xi_j\rangle = \omega$$

giving

$$P(x,y) = \frac{\mathrm{e}^{-\frac{1}{2}\left[x^2+y^2-2xy\omega\right]/(1-\omega^2)}}{2\pi\sqrt{1-\omega^2}} \tag{14}$$

Note that $P(x,y) = P(y,x)$. The simple fact that $P(x,y)$ depends on time only through ω ensures that the two Equations 12 and 13 are a *closed* set. Note also that now (12, 13) are deterministic equations; the fluctuations in the macroscopic observables $Q[\boldsymbol{J}]$ and $R[\boldsymbol{J}]$ vanish in the $N\to\infty$ limit.

Finally, the generalization error E_g (here identical to the training error E_t due to $\tilde{D} = D$) can be expressed in terms of our macroscopic observables. We define the error made in a single classification of an input $\boldsymbol{\xi}$ as $E[T(\boldsymbol{\xi}),S(\boldsymbol{\xi})] = \theta[-(\boldsymbol{B}\cdot\boldsymbol{\xi})(\boldsymbol{J}\cdot\boldsymbol{\xi})]\in\{0,1\}$. Averaged over D this gives the probability of a misclassification for randomly drawn questions $\boldsymbol{\xi}\in D$:

$$\lim_{N\to\infty}E_\mathrm{g}(\boldsymbol{J}(t)) = \lim_{N\to\infty}\langle[\theta[-(\boldsymbol{B}\cdot\boldsymbol{\xi})(\boldsymbol{J}(t)\cdot\boldsymbol{\xi})]\rangle_D = \langle\theta[-xy]\rangle = \int_0^\infty\int_0^\infty \mathrm{d}x\,\mathrm{d}y[P(x,-y) + P(-x,y)]$$

The generalization error (from this stage onwards to be denoted simply by E) also evolves deterministically for $N\to\infty$, and can be expressed purely in terms of the observable ω. The integral (with the distribution (14)) can even be done analytically (see the appendix) and produces the simple result.

$$E = \tfrac{1}{\pi}\arccos(\omega) \tag{15}$$

The macroscopic Equations 12 and 13 can now equivalently be written in terms of the pair (J,E). We have hereby achieved our goal: we have derived a closed set of deterministic equations for a small number (two) of macroscopic observables, valid for $N\to\infty$, and we know the generalization error at any time.

2.2. *Hebbian learning with constant learning rate*

We will now work out our general result (12, 13, 14) for specific members of the general class (6) of on-line learning rules. The simplest non-trivial choice to be made is the Hebbian rule, obtained by choosing $\mathscr{F}[J;Jx,y] = 1$, with a constant learning rate η:

[*] It is not true for all choices of $\boldsymbol{J}$ and $\boldsymbol{B}$. A trivial counter-example is $J_k = \delta_{k1}$, less trivial counter-examples are $J_k = \mathrm{e}^{-k}$ and $J_k = k^{-\gamma}$ with $\gamma > \frac{1}{2}$.

$$\boldsymbol{J}\left(t_\mu+\frac{1}{N}\right)=\boldsymbol{J}(t_\mu)+\frac{\eta}{N}\xi^\mu\,\mathrm{sgn}\,(\boldsymbol{B}\cdot\xi^\mu) \qquad (16)$$

Equations 12 and 13, describing the macroscopic dynamics generated by (16) in the limit $N\to\infty$, now become

$$\frac{\mathrm{d}}{\mathrm{d}t}J=\eta\langle x\,\mathrm{sgn}(y)\rangle+\frac{\eta^2}{2J} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\omega=\frac{\eta}{J}\langle|y|-\omega x\,\mathrm{sgn}(y)\rangle-\frac{\omega\eta^2}{2J^2}$$

The integrals in these equations can be calculated analytically (see the appendix) and we get

$$\frac{\mathrm{d}}{\mathrm{d}t}J=\omega\eta\sqrt{\frac{2}{\pi}}+\frac{\eta^2}{2J} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\omega=(1-\omega^2)\frac{\eta}{J}\sqrt{\frac{2}{\pi}}-\frac{\omega\eta^2}{2J^2}$$

Thus, upon elimination of the observable ω using Equation 15, we arrive at the following closed differential equations in terms of J and E:

$$\frac{\mathrm{d}}{\mathrm{d}t}J=\eta\cos(\pi E)\sqrt{\frac{2}{\pi}}+\frac{\eta^2}{2J} \qquad (17)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}E=-\frac{\eta\sin(\pi E)}{\pi J}\sqrt{\frac{2}{\pi}}+\frac{\eta^2}{2\pi J^2\tan(\pi E)} \qquad (18)$$

The flow in the (E,J) plane described by these equations is drawn in Fig. 5 (which is obtained by numerical solution of (17, 18)). From (17) it follows that $\frac{\mathrm{d}}{\mathrm{d}t}J>0\ \forall t\geq 0$. From (18) it follows that $\frac{\mathrm{d}}{\mathrm{d}t}E=0$ along the line

$$J_c(E)=\frac{\eta\cos(\pi E)}{2\sin^2(\pi E)}\sqrt{\frac{\pi}{2}}$$

(drawn as a dashed line in Fig. 5).

Let us now investigate the temporal properties of the solution (17, 18), and work out their predictions for the asymptotic decay of the generalization error. For small values of E Equations 17 and 18 yield

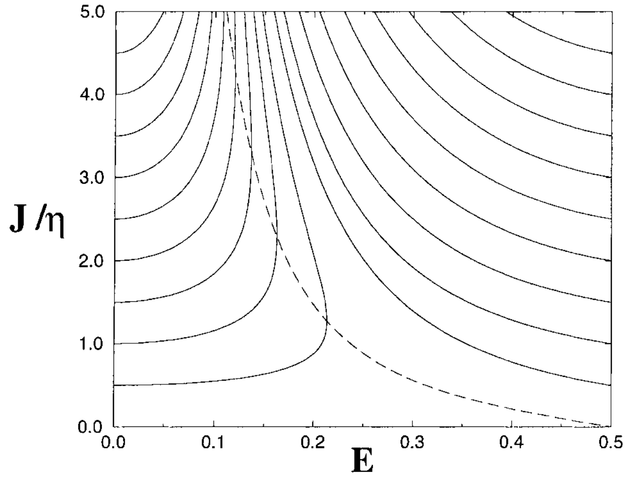

Fig. 5. *Flow in the (E,J) plane generated by the Hebbian learning rule with constant learning rate η, in the limit $N\to\infty$. Dashed: the line where $\mathrm{d}E/\mathrm{d}t=0$ ($\mathrm{d}J/\mathrm{d}t>0$ for any (E,J)). Note that the flow asymptotically gives $E\to 0$ and $J\to\infty$*

$$\frac{\mathrm{d}}{\mathrm{d}t}J=\eta\sqrt{\frac{2}{\pi}}+\frac{\eta^2}{2J}+\mathcal{O}(E^2) \qquad (19)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}E=-\frac{\eta E}{J}\sqrt{\frac{2}{\pi}}+\frac{\eta^2}{2\pi^2J^2E}+\mathcal{O}(E^3/J,E/J^2) \qquad (20)$$

From (19) we infer that $J\sim\eta t\sqrt{2/\pi}$ for $t\to\infty$. Substitution of this asymptotic solution into Equation 20 gives

$$\frac{\mathrm{d}}{\mathrm{d}t}E=-\frac{E}{t}+\frac{1}{4\pi Et^2}+\mathcal{O}(E^3/t,E/t^2) \qquad (t\to\infty) \qquad (21)$$

We insert the ansatz $E=At^{-\alpha}$ into Equation 21 and get the solution $A=1/\sqrt{2\pi}$, $\alpha=1/2$. This implies that (in the $N\to\infty$ limit) on-line Hebbian learning with complete training sets produces an asymptotic decay of the generalization error of the form

$$E\sim\frac{1}{\sqrt{2\pi t}} \qquad (t\to\infty) \qquad (22)$$

The reason that this expression does not depend on the learning rate η is that the latter can be eliminated from the macroscopic dynamic equations by a simple rescaling of the length J of the weight vector. Figures 8, 9 and 10 will show the theoretical results of this section together with the results of doing numerical simulations of the learning rule (16) and with similar results for other on-line learning rules with constant learning rates. The agreement between theory and simulations is quite convincing.

2.3. *Perceptron learning with constant learning rate*

Our second application of (12, 13, 14) is making the choice $\mathcal{F}[J;Jx,y]=\theta[-xy]$ in Equation 6, with constant learning rate η, which produces the perceptron learning algorithm:

$$\boldsymbol{J}\left(t_\mu+\frac{1}{N}\right)=\boldsymbol{J}(t_\mu)+\frac{\eta}{N}\xi^\mu\mathrm{sgn}(\boldsymbol{B}\cdot\xi^\mu) \\ \times\theta[-(\boldsymbol{B}\cdot\xi^\mu)(\boldsymbol{J}(t_\mu)\cdot\xi^\mu)] \qquad (23)$$

In other words, the student weights are updated in accordance with the Hebbian rule only when $\mathrm{sgn}(\boldsymbol{B}\cdot\xi)=-\mathrm{sgn}(\boldsymbol{J}\cdot\xi)$, i.e. when student and teacher are not in agreement. Equations 12 and 13 now become

$$\frac{\mathrm{d}}{\mathrm{d}t}J=\eta\langle x\,\mathrm{sgn}\,(y)\theta[-xy]\rangle+\frac{\eta^2}{2J}\langle\theta[-xy]\rangle$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\omega=\frac{\eta}{J}\langle[\,|y|-\omega x\,\mathrm{sgn}\,(y)]\theta[-xy]\rangle-\frac{\omega\eta^2}{2J^2}\langle\theta[-xy]\rangle$$

As before the various Gaussian integrals occurring in these expressions can be done analytically (see the appendix), which results in

$$\frac{\mathrm{d}}{\mathrm{d}t}J=-\frac{\eta(1-\omega)}{\sqrt{2\pi}}+\frac{\eta^2}{2\pi J}\arccos(\omega)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\omega=\frac{\eta(1-\omega^2)}{\sqrt{2\pi}J}-\frac{\omega\eta^2}{2\pi J^2}\arccos(\omega)$$

Elimination of ω using (15) then gives us the dynamical equations in terms of the pair (J, E):

$$\frac{\mathrm{d}}{\mathrm{d}t}J = -\frac{\eta(1-\cos(\pi E))}{\sqrt{2\pi}} + \frac{\eta^2 E}{2J} \tag{24}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{\eta\sin(\pi E)}{\pi\sqrt{2\pi}J} + \frac{\eta^2 E}{2\pi J^2\tan(\pi E)} \tag{25}$$

Figure 6 shows the flow in the (E, J) plane, obtained by numerical solution of (24, 25). The two lines where $\frac{\mathrm{d}}{\mathrm{d}t}J = 0$ and where $\frac{\mathrm{d}}{\mathrm{d}t}E = 0$ are found to be $J_{c,1}(E)$ and $J_{c,2}(E)$, respectively:

$$J_{c,1}(E) = \eta\sqrt{\frac{\pi}{2}}\frac{E}{1-\cos(\pi E)}$$

$$J_{c,2}(E) = \eta\sqrt{\frac{\pi}{2}}\frac{E\cos(\pi E)}{1-\cos^2(\pi E)}$$

For $E \in [0, 1/2]$ one always has $J_{c,1}(E) \geq J_{c,2}(E)$, with equality only if $(J, E) = (\infty, 0)$. Figure 6 shows that the flow is drawn into the gully between the curves $J_{c,1}(E)$ and $J_{c,2}(E)$.

As with the Hebbian rule we now wish to investigate the asymptotic behaviour of the generalization error. To do this we expand Equations 24 and 25 for small E:

$$\frac{\mathrm{d}}{\mathrm{d}t}J = -\frac{\eta\pi^2 E^2}{2\sqrt{2\pi}} + \frac{\eta^2 E}{2J} + \mathcal{O}(E^4)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{\eta E}{\sqrt{2\pi}J} + \frac{\eta^2}{2\pi^2 J^2} - \frac{\eta^2 E^2}{6J^2} + \mathcal{O}(E^3)$$

For small E and large t we know that $J \sim J_{c,1}(E) \sim 1/E$. Making the ansatz $J = A/E$ (and hence $\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{E^2}{A}\frac{\mathrm{d}}{\mathrm{d}t}J$) leads to a situation where we have two equivalent differential equations for E:

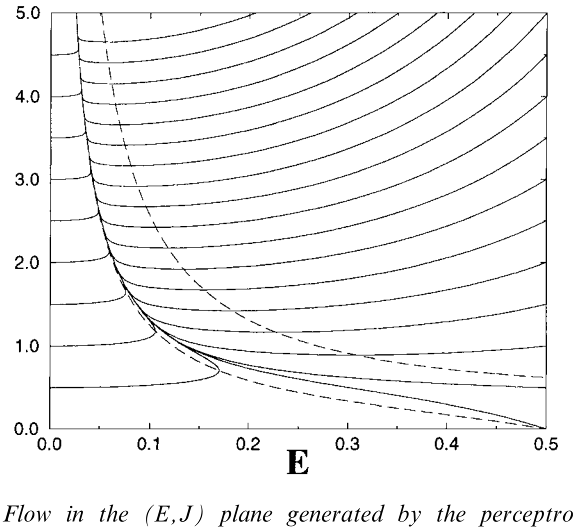

Fig. 6. *Flow in the (E,J) plane generated by the perceptron learning rule with constant learning rate η, in the limit $N \to \infty$. Dashed: the two lines where $\mathrm{d}E/\mathrm{d}t = 0$ and $\mathrm{d}J/\mathrm{d}t = 0$, respectively. Note that the flow is attracted into the gully between these two dashed lines and asymptotically gives $E \to 0$ and $J \to \infty$*

$$\frac{\mathrm{d}}{\mathrm{d}t}E = \frac{\eta\pi^2 E^4}{2\sqrt{2\pi}A} - \frac{\eta^2 E^4}{2A^2} + \mathcal{O}(E^6)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{\eta E^2}{\sqrt{2\pi}A} + \frac{\eta^2 E^2}{2\pi^2 A^2} + \mathcal{O}(E^4)$$

Since both describe the same dynamics, the leading term of the second expression should be identical to that of the first, i.e. $\mathcal{O}(E^4)$, giving us the condition $A = \eta\sqrt{2\pi}/(2\pi^2)$. Substitution of this condition into the first expression for $\frac{\mathrm{d}}{\mathrm{d}t}E$ then gives

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\tfrac{1}{2}\pi^3 E^4 + \mathcal{O}(E^5) \qquad (t \to \infty)$$

which has the solution

$$E \sim \left(\tfrac{2}{3}\right)^{1/3}\pi^{-1}t^{-1/3} \qquad (t \to \infty) \tag{26}$$

As with the Hebbian learning rule, this expression does not depend on the learning rate since the latter can be eliminated from the macroscopic equations by rescaling J. We find, somewhat surprisingly, that in large systems $(N \to \infty)$ the on-line perceptron learning rule is asymptotically much slower in converging towards the desired $E = 0$ state than the simpler Hebbian rule. This will be different if we allow for time-dependent learning rates. Figures 8, 9 and 10 will show the theoretical results on the perceptron rule together with the results of doing numerical simulations and together with similar results for other on-line learning rules. Again the agreement between theory and experiment is quite satisfactory.

2.4. *AdaTron learning with constant learning rate*

As our third application we analyse the macroscopic dynamics of the AdaTron learning rule, corresponding to the choice $\mathscr{F}[J; Jx, y] = |Jx|\theta[-xy]$ in the general recipe (6). As in the perceptron rule, modifications are made only when student and teacher are in disagreement; however, here the modification made is proportional to the magnitude of the student's local field. Students are punished in proportion to their confidence in the wrong answer. The rationale is that wrong student answers $S(\boldsymbol{\xi}) = \mathrm{sgn}(\boldsymbol{J}\cdot\boldsymbol{\xi})$ with large values of $|\boldsymbol{J}\cdot\boldsymbol{\xi}|$ require more rigorous corrections to $\boldsymbol{J}$ to be remedied than those with small values of $|\boldsymbol{J}\cdot\boldsymbol{\xi}|$.

$$\begin{aligned}\boldsymbol{J}\left(t_\mu + \frac{1}{N}\right) &= \boldsymbol{J}(t_\mu) + \frac{\eta}{N}\boldsymbol{\xi}^\mu\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu) \\ &\quad\times|\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu|\theta[-(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)(\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu)]\end{aligned} \tag{27}$$

Working out the general Equations 12 and 13 for the learning rule (27) gives

$$\frac{\mathrm{d}}{\mathrm{d}t}J = \eta J\langle x|x|\mathrm{sgn}(y)\theta[-xy]\rangle + \tfrac{1}{2}\eta^2 J\langle x^2\theta[-xy]\rangle$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\omega &= \eta\langle|xy|\theta[-xy]\rangle - \eta\omega\langle x|x|\mathrm{sgn}(y)\theta[-xy]\rangle \\ &\quad - \tfrac{1}{2}\omega\eta^2\langle x^2\theta[-xy]\rangle\end{aligned}$$

All integrals can again be done analytically (see the appendix) so that we obtain explicit macroscopic flow equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}J = \frac{J}{\omega}\left[\eta - \frac{\eta^2}{2}\right]I_2(\omega)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\omega = \eta I_1(\omega) - \left[\eta - \frac{\eta^2}{2}\right]I_2(\omega)$$

with the short-hands

$$I_1(\omega) = \frac{(1-\omega^2)^{3/2}}{\pi} - \frac{\omega(1-\omega^2)}{\pi}\arccos(\omega) + \frac{\omega^2\sqrt{1-\omega^2}}{\pi} - \frac{\omega^3}{\pi}\arccos(\omega)$$

$$I_2(\omega) = -\frac{\omega(1-\omega^2)}{\pi}\arccos(\omega) + \frac{\omega^2\sqrt{1-\omega^2}}{\pi} - \frac{\omega^3}{\pi}\arccos(\omega)$$

The usual translation from equations for the pair (J, ω) into one involving the pair (J, E), following (15), turns out to simplify matters considerably, since it gives

$$\frac{\mathrm{d}}{\mathrm{d}t}J = J\left[\frac{\eta^2}{2} - \eta\right]\left[E - \frac{\cos(\pi E)\sin(\pi E)}{\pi}\right] \tag{28}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{\eta\sin^2(\pi E)}{\pi^2} + \frac{\eta^2 E}{2\pi\tan(\pi E)} - \frac{\eta^2\cos^2(\pi E)}{2\pi^2} \tag{29}$$

The flow described by Equations 28 and 29 is shown in Fig. 7, for the case $\eta = 1$. In contrast with the Hebbian and the perceptron learning rules we here observe from Equations 28 and 29 that the learning rate η cannot be eliminated from the macroscopic laws by a rescaling of the weight vector length J. Moreover, the state $E = 0$ is stable only for $\eta < 3$, in which case $\frac{\mathrm{d}}{\mathrm{d}t}E < 0$ for all t. For $\eta < 2$ one has $\frac{\mathrm{d}}{\mathrm{d}t}J < 0$ for all t, for $\eta = 2$ one has $J(t) = J(0)$ for all t, and for $2 < \eta < 3$ we have $\frac{\mathrm{d}}{\mathrm{d}t}J > 0$ for all t.

For small E Equation 29 reduces to

$$\frac{\mathrm{d}}{\mathrm{d}t}E = \left[\frac{\eta^2}{3} - \eta\right]E^2 + \mathcal{O}(E^4)$$

giving

$$E \sim \frac{3t^{-1}}{\eta(3-\eta)} \qquad (t \to \infty) \tag{30}$$

For $\eta = 1$, which gives the standard representation of the AdaTron algorithm, we find $E \sim \frac{3}{2}t^{-1}$. Note from Equation 28 that for the AdaTron rule there is a specific value for η, namely $\eta = 2$, for which the length J of the student's weight vector would remain constant; this again gives $E \sim \frac{3}{2}t^{-1}$. The optimal value for η, however, is $\eta = \frac{3}{2}$ in which case we find $E \sim \frac{4}{3}t^{-1}$ (see (30)).

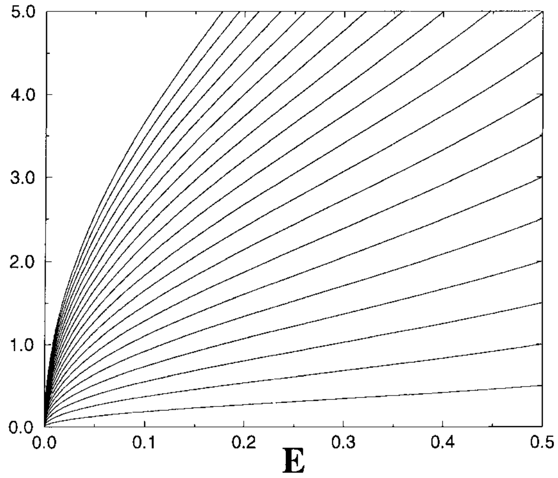

Fig. 7. *Flow in the (E, J) plane generated by the AdaTron learning rule with constant learning rate $\eta = 1$, in the limit $N \to \infty$ (in this case the influence of the value of the learning rate on the flow is more than just a rescaling of the length J)*

2.5. *Theory versus simulations*

We close this section with results of the comparison of the dynamics described by the various macroscopic flow equations with the results of measuring the error E during numerical simulations of the various (microscopic) learning rules discussed so far. This will serve to support the analysis and its implicit and explicit assumptions, but also illustrates how the three learning rules compare among one another. Figures 8 and 9 show the initial stage of the learning processes for initializations corresponding to random guessing ($E = 0.5$) and almost correct classification (E small), respectively (note that for the perceptron and AdaTron rules, starting at precisely $E = 0$ produces a stationary state in finite systems). Here the solutions of the flow equations (solid lines) were obtained by numerical iteration. The initial increase in the error E, as observed for the Hebbian and perceptron rule, following initialization with small values of E can be understood as follows. The error depends only on the angle of the weight vector $\boldsymbol{J}$, not on its length J, this means that the modifications generated by the Hebbian and perceptron learning rules (which are of uniform magnitude) generate large changes in E when J is small, but small changes in E when J is large, with corresponding effects on the stability of low E states. The AdaTron rule, in contrast, involves weight changes which scale with the length J, so that the stability of the $E = 0$ state does not depend on the value of J. Figure 10 shows the asymptotic relaxation of the

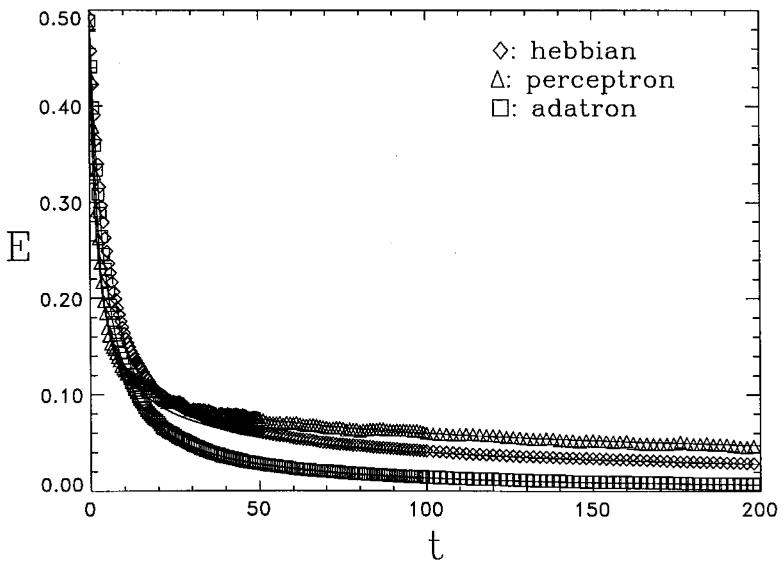

Fig. 8. *Evolution in time of the generalization error E as measured during numerical simulations (with N =1000 neurons) of three different learning rules: Hebbian (diamonds), perceptron (triangles) and AdaTron (squares). Initial state: E (0) = 1/2 (random guessing) and J(0) = 1. Learning rate: $\eta = 1$. The solid lines give for each learning rule the prediction of the $N = \infty$ theory, obtained by numerical solution of the flow equations for (E, J)*

error E, in a log–log plot, together with the three corresponding asymptotic (power law) predictions (22, 26, 30). All simulations were carried out with networks of $N = 1000$ neurons, which, it can be seen, is already sufficiently large for the $N = \infty$ theory to apply. The teacher weight vectors $\boldsymbol{B}$ were in all cases drawn at random from $[-1, 1]^N$. We conclude that the theory describes the simulations essentially perfectly.

3. On-line learning: complete training sets and optimized rules

We now set out to use our macroscopic equations in 're-verse mode'. Rather than calculate the macroscopic dynamics for a given choice of learning rule, we will try to find learning rules that optimize the macroscopic dynamical laws in the sense that they produce the fastest decay to-

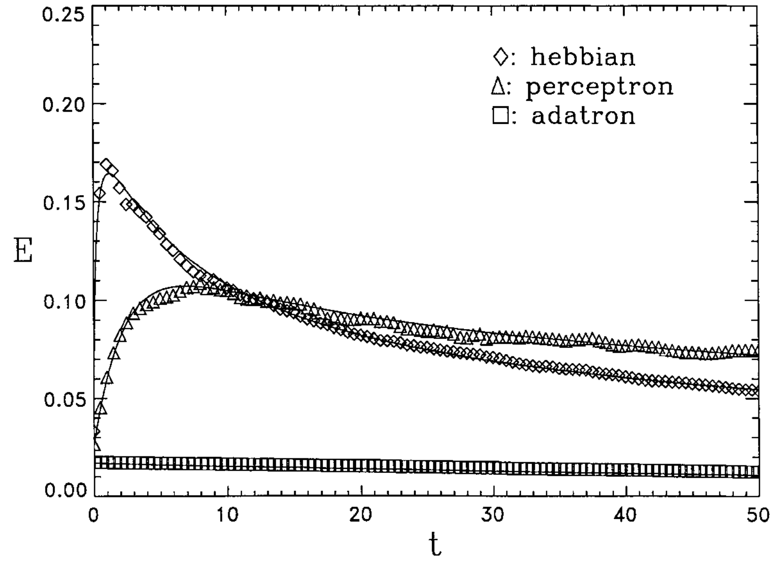

Fig. 9. *Evolution in time of the generalization error E as measured during numerical simulations (with N = 1000 neurons) of three different learning rules: Hebbian (diamonds), perceptron (triangles) and AdaTron (squares). Initial state: E(0) ≈ 0.025 and J(0) = 1. Learning rate: $\eta = 1$. The solid lines give, for each learning rule, the prediction of the $N = \infty$ theory, obtained by numerical solution of the flow equations for (E, J)*

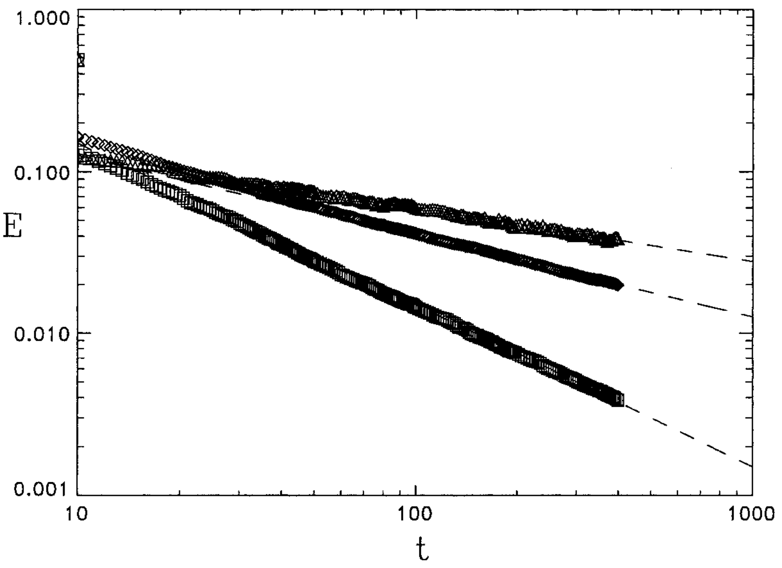

Fig. 10. *Asymptotic behaviour of the generalization error E measured during numerical simulations (with $N = 1000$) of three different learning rules: Hebbian (diamonds, middle curve), perceptron (triangles, upper curve) and AdaTron (squares, lower curve). Initial state: $E(0) = 1/2$ and $J(0) = 1$. Learning rate: $\eta = 1$. The dashed lines give, for each learning rule, the corresponding asymptotic power law predicted by the $N = \infty$ theory (Equations 22, 26 and 30, respectively)*

wards the desired $E = 0$ state. As a bonus it will turn out that in many cases we can even solve the corresponding macroscopic differential equations analytically, and find explicit expressions for $E(t)$, or rather its inverse $t(E)$.

3.1. *Time-dependent learning rates*

First we illustrate how modifying existing learning rules in a simple way, by just allowing for suitably chosen time-dependent learning rates $\eta(t)$, can already lead to a drastic improvement in the asymptotic behaviour of the error E.

We will inspect two specific choices of time-dependent learning rates for the perceptron rule. Without loss of generality we can always put $\eta(t) = K(t)J(t)$ in our dynamic equations (for notational convenience we will drop the explicit time argument of K). This choice will enable us to decouple the dynamics of J from that of the generalization error E. For the perceptron rule we subsequently find Equation 25 being replaced by

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{K\sin(\pi E)}{\pi\sqrt{2\pi}} + \frac{K^2 E}{2\pi\tan(\pi E)}$$

giving for small E

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{KE}{\sqrt{2\pi}} + \frac{K^2}{2\pi^2} + \mathcal{O}(K^2E^2)$$

In order to obtain $E \to 0$ for $t \to \infty$ it is clear that we need $K \to 0$. Applying the ansatz $E = A/t^{\alpha}$, $K = B/t^{\beta}$ for the asymptotic forms in the previous equation produces

$$-At^{-\alpha-1} = \frac{-ABt^{-\alpha-\beta}}{\sqrt{2\pi}} + \frac{B^2t^{-2\beta}}{2\pi^2} + \mathcal{O}(t^{-2\alpha-2\beta})$$

and so: $\alpha = \beta = 1$ and $A = B^2/(\pi\sqrt{2\pi})(B - \sqrt{2\pi})$. Our aim is to obtain the fastest approach of the $E = 0$ state, i.e. we wish to maximize α (for which we found $\alpha = 1$) and subsequently minimize A. We find the value of B for which A is minimized is $B = 2\sqrt{2\pi}$, in which case we obtain the error decay given by

$$\eta \sim \frac{2J\sqrt{2\pi}}{t}: \qquad E \sim \frac{4}{\pi t} \qquad (t \to \infty) \tag{31}$$

This is clearly a great improvement upon the result for the perceptron rule with constant η, i.e. Equation 26; in fact it is the fastest relaxation we have derived so far.

Let us now move to an alternative choice for the time-dependent learning rate for the perceptron. According to Equation 24, there is one specific recipe for $\eta(t)$ such that the length J of the student's weight vector will remain constant, given by

$$\eta = \sqrt{\frac{2}{\pi}}\frac{J}{E}(1 - \cos(\pi E)) \tag{32}$$

Making this choice converts Equation 25 for the evolution of E into

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{(1 - \cos(\pi E))^2}{\pi^2 E\sin(\pi E)} \tag{33}$$

Equation 33 can be written in the form $\frac{\mathrm{d}}{\mathrm{d}E}t = g(E)$, so that $t(E)$ becomes a simple integral which can be done analytically, with the result

$$t(E) = \frac{\pi E + \sin(\pi E)}{1 - \cos(\pi E)} - \frac{\pi E_0 + \sin(\pi E_0)}{1 - \cos(\pi E_0)} \qquad (34)$$

(which can also be verified directly by substitution into (33)). Expansion of (34) and (32) for small E gives the asymptotic behaviour also encountered in (31):

$$\eta \sim \frac{2J\sqrt{2\pi}}{t}, \qquad E \sim \frac{4}{\pi t} \qquad (t \to \infty) \qquad (35)$$

It might appear that implementation of the recipe (32) is in practice impossible, since it involves information which is not available to the student perceptron (namely the instantaneous error E). However, since we know (34) we can simply calculate the required $\eta(t)$ explicitly as a function of time.

One has to be somewhat careful in extrapolating results such as those obtained in this section. For instance, choosing the time-dependent learning rate (32) enforces the constraint $J(t) = J(0)$ in the macroscopic equations for $N \to \infty$. This is not identical to choosing $\eta(t_\mu)$ in the original Equation 6 such as to enforce $\boldsymbol{J}^2(t_\mu + (1/N)) = \boldsymbol{J}^2(t_\mu)$ at the level of individual iteration steps, as can be seen by working out the dynamical laws. The latter case would correspond to the *microscopically fluctuating* choice

$$\eta(t_\mu) = -2\frac{(\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu)\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)}{\mathscr{F}[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]}$$

if

$$\mathscr{F}[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu] \neq 0$$

If we now choose for example $\mathscr{F}[J; Jx, y] = \theta[-xy]$, implying $\eta(t_\mu) = 2|\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu|$, we find by insertion into (6) that the perceptron rule with 'hard' weight normalization at each iteration step via adaptation of the learning rate is identical to the AdaTron rule with constant learning rate $\eta = 2$. We know therefore that in this case one obtains $E \sim 3/2t$, whereas for the perceptron rule with 'soft' weight normalization via (32) (see the analysis above) one obtains $E \sim 4/\pi t$. Clearly the two procedures are not equivalent.

3.2. *Spherical on-line learning rules*

We arrive in a natural way at the question of how to find the optimal time-dependent learning rate for any given learning rule, or more generally: of how to find the optimal learning rule. This involves variational calculations in two-dimensional flows (since our macroscopic equations are defined in terms of the evolving pair (J, E)). Such calculations would be much simpler if our macroscopic equations were just one-dimensional, e.g. describing only the evolution of the error E with a stationary (or simply irrelevant) value of the length J. Often it will turn out that for finding the optimal learning rate or the optimal learning rule the problem can indeed be reduced to a one-dimensional one. To be able to obtain results also for those cases where this reduction does not happen we will now construct so-called spherical learning rules, where $\boldsymbol{J}^2(t) = 1$ for all t. This can be arranged in several equivalent ways.

The first method is to add the general rule (6) a term proportional to the instantaneous weight vector $\boldsymbol{J}$, whose sole purpose is to achieve the constraint $\boldsymbol{J}^2 = 1$:

$$\boldsymbol{J}(t_\mu + \tfrac{1}{N}) = \boldsymbol{J}(t_\mu) + \tfrac{1}{N}\{\eta(t_\mu)\boldsymbol{\xi}^\mu\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu) \times \mathscr{F}[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu] - \lambda(t_\mu)\boldsymbol{J}(t_\mu)\} \qquad (36)$$

The evolution of the two observables $Q[\boldsymbol{J}]$ and $R[\boldsymbol{J}]$ (7) is now given by

$$\begin{aligned} Q\left[\boldsymbol{J}\left(t_\mu + \frac{1}{N}\right)\right] &= Q[\boldsymbol{J}(t_\mu)]\left(1 - \frac{2\lambda(t_\mu)}{N}\right) + \frac{2}{N}\eta(t_\mu)(\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu) \\ &\quad \times \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\mathscr{F}[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu] \\ &\quad + \frac{1}{N}\eta^2(t_\mu)\mathscr{F}^2[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu] \\ &\quad + \mathcal{O}(N^{-2}) \\ R\left[\boldsymbol{J}\left(t_\mu + \frac{1}{N}\right)\right] &= R[\boldsymbol{J}(t_\mu)]\left(1 - \frac{\lambda(t_\mu)}{N}\right) + \frac{1}{N}\eta(t_\mu)|\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu| \\ &\quad \times \mathscr{F}[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu] \end{aligned}$$

Following the procedure of Section 1.2 to arrive at the $N \to \infty$ limit of the dynamical equations for Q and R then leads to (we drop explicit time arguments for notational convenience):

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}Q &= 2\eta Q^{\frac{1}{2}}\left\langle x\ \mathrm{sgn}(y)\mathscr{F}\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right]\right\rangle \\ &\quad + \eta^2\left\langle\mathscr{F}^2\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right]\right\rangle - 2\lambda Q \\ \frac{\mathrm{d}}{\mathrm{d}t}R &= \eta\left\langle|y|\mathscr{F}\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right]\right\rangle - \lambda R \end{aligned}$$

We now choose the function $\lambda(t)$ such that $Q(t) = 1$ for all $t \geq 0$. This ensures that $R(t) = \omega(t) = \hat{\boldsymbol{J}}(t)\cdot\boldsymbol{B}$, and gives (via $\frac{\mathrm{d}}{\mathrm{d}t}Q = 0$) a recipe for $\lambda(t)$

$$\lambda = \eta\langle x\ \mathrm{sgn}(y)\mathscr{F}[1; x, y]\rangle + \tfrac{1}{2}\eta^2\langle\mathscr{F}^2[1; x, y]\rangle$$

which can then be substituted into our equation for $\frac{\mathrm{d}}{\mathrm{d}t}\omega$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\omega = \eta\langle[|y| - \omega x\ \mathrm{sgn}(y)]\mathscr{F}[1; x, y]\rangle - \tfrac{1}{2}\omega\eta^2\langle\mathscr{F}^2[1; x, y]\rangle \qquad (37)$$

with averages as usual defined with respect to the Gaussian joint field distribution (14), which depends only on ω, so that Equation 37 is indeed autonomous.

The second method to arrange the constraint $\boldsymbol{J}^2 = 1$ is to explicitly normalize the weight vector $\boldsymbol{J}$ after each modification step, i.e.

$$
\begin{aligned}
&\boldsymbol{J}\left(t_\mu+\frac{1}{N}\right)\\
&=\frac{\boldsymbol{J}(t_\mu)+\frac{1}{N}\eta(t_\mu)\boldsymbol{\xi}^\mu\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\mathscr{F}[1;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]}{|\boldsymbol{J}(t_\mu)+\frac{1}{N}\eta(t_\mu)\boldsymbol{\xi}^\mu\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\mathscr{F}[1;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]|}\\
&=\hat{\boldsymbol{J}}(t_\mu)+\frac{1}{N}\eta(t_\mu)\left\{\left[\boldsymbol{\xi}^\mu-\hat{\boldsymbol{J}}(t_\mu)(\hat{\boldsymbol{J}}(t_\mu)\cdot\boldsymbol{\xi}^\mu)\right]\right.\\
&\quad\times\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\mathscr{F}[1;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]\\
&\quad\left.-\frac{1}{2}\eta(t_\mu)\boldsymbol{J}(t_\mu)\mathscr{F}^2\left[1;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu\right]\right\}+\mathcal{O}(N^{-2})
\end{aligned}
\tag{38}
$$

The evolution of the observable $\omega[\boldsymbol{J}]=\hat{\boldsymbol{J}}\cdot\boldsymbol{B}$ is thus given by

$$
\begin{aligned}
\omega\left[\boldsymbol{J}\left(t_\mu+\frac{1}{N}\right)\right]&=\omega[\boldsymbol{J}(t_\mu)]+\frac{1}{N}\eta(t_\mu)\\
&\quad\times\left\{\left[|\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu|-\omega[\boldsymbol{J}(t_\mu)](\hat{\boldsymbol{J}}(t_\mu)\cdot\boldsymbol{\xi}^\mu)\right.\right.\\
&\quad\left.\times\,\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\right]\mathscr{F}[1;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]\\
&\quad\left.-\frac{1}{2}\omega\eta(t_\mu)\mathscr{F}^2[1;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]\right\}\\
&\quad+\mathcal{O}(N^{-2})
\end{aligned}
$$

Following the procedure of Section 1.2 then leads to

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\omega&=\eta\langle[|y|-\omega x\ \mathrm{sgn}(y)]\mathscr{F}[1;x,y]\rangle\\
&\quad-\frac{1}{2}\omega\eta^2\langle\mathscr{F}^2[1;x,y]\rangle
\end{aligned}
\tag{39}
$$

which is identical to Equation 37.

Finally we convert Equation 37 into a dynamical equation for the error E, using (15), which gives the final result

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}E&=-\frac{\eta}{\pi\sin(\pi E)}\langle[|y|-\cos(\pi E)x\ \mathrm{sgn}(y)]\mathscr{F}[1;x,y]\rangle\\
&\quad+\frac{\eta^2}{2\pi\tan(\pi E)}\langle\mathscr{F}^2[1;x,y]\rangle
\end{aligned}
\tag{40}
$$

with averages defined with respect to the distribution (14), in which $\omega=\cos(\pi E)$.

For spherical models described by either of the equivalent classes of on-line rules (36) or (38), the evolution of the error is described by a single first-order non-linear differential equation, rather than a pair of coupled non-linear differential equations. This will allow us to push the analysis further, but the price we pay is that of a loss in generality.

3.3. *Optimal time-dependent learning rates*

We wish to optimize the approach to the $E=0$ state of our macroscopic equations, by choosing a suitable time-dependent learning rate. Let us distinguish between the possible situations we can find ourselves in. If our learning rule is of the general form (6), without spherical normalization, we have two coupled macroscopic equations:

$$
\frac{\mathrm{d}}{\mathrm{d}t}J=\eta\langle x\ \mathrm{sgn}(y)\mathscr{F}[J;Jx,y]\rangle+\frac{\eta^2}{2J}\langle\mathscr{F}^2[J;Jx,y]\rangle \tag{41}
$$

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}E&=-\frac{\eta}{J\pi\sin(\pi E)}\langle[|y|-\cos(\pi E)x\,\mathrm{sgn}(y)]\mathscr{F}[J;Jx,y]\rangle\\
&\quad+\frac{\eta^2}{2\pi J^2\tan(\pi E)}\langle\mathscr{F}^2[J;Jx,y]\rangle
\end{aligned}
\tag{42}
$$

which are obtained by combining (12, 13) with (15). The probability distribution (14) with which the averages are computed depends on E only, not on J. If, on the other hand, we complement the rule (6) with weight vector normalization as in (36) or (38) (the spherical rules), we obtain a single equation for E only:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}E&=-\frac{\eta}{\pi\sin(\pi E)}\langle[|y|-\cos(\pi E)x\ \mathrm{sgn}(y)]\mathscr{F}[1;x,y]\rangle\\
&\quad+\frac{\eta^2}{2\pi\tan(\pi E)}\langle\mathscr{F}^2[1;x,y]\rangle
\end{aligned}
\tag{43}
$$

Since Equation 43 is autonomous (there are no dynamical variables other than E), the optimal choice of the function $\tilde{\eta}(t)$ (i.e. the one that generates the fastest decay of the error E) is obtained by simply minimizing the temporal derivative of the error *at each time-step*:

$$
\forall t\geq 0:\qquad \frac{\partial}{\partial\tilde{\eta}(t)}\left[\frac{\mathrm{d}}{\mathrm{d}t}E\right]=0 \tag{44}
$$

which is called the 'greedy' recipe. Note, however, that the same is true for Equation 42 if we restrict ourselves to rules with the property that $\mathscr{F}[J;Jx,y]=\gamma(J)\mathscr{F}[1;x,y]$ for some function $\gamma(J)$, such as the Hebbian ($\gamma(J)=1$), perceptron ($\gamma(J)=1$) and AdaTron ($\gamma(J)=J$) rules. This property can also be written as

$$
\frac{\partial}{\partial x}\frac{\mathscr{F}[J;Jx,y]}{\mathscr{F}[1;x,y]}=\frac{\partial}{\partial y}\frac{\mathscr{F}[J;Jx,y]}{\mathscr{F}[1;x,y]}=0 \tag{45}
$$

For rules which obey (45) we can simply write the time-dependent learning rate as $\eta=\tilde{\eta}J/\gamma(J)$, such that Equations 41 and 42 acquire the form:

$$
\frac{\mathrm{d}}{\mathrm{d}t}\log J=\tilde{\eta}\langle x\,\mathrm{sgn}(y)\mathscr{F}[1;x,y]\rangle+\frac{1}{2}\tilde{\eta}^2\langle\mathscr{F}^2[1;x,y]\rangle \tag{46}
$$

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}E&=-\frac{\tilde{\eta}}{\pi\sin(\pi E)}\langle[|y|-\cos(\pi E)x\,\mathrm{sgn}(y)]\mathscr{F}[1;x,y]\rangle\\
&\quad+\frac{\tilde{\eta}^2}{2\pi\tan(\pi E)}\langle\mathscr{F}^2[1;x,y]\rangle
\end{aligned}
\tag{47}
$$

In these cases, precisely since we are free to choose the function $\tilde{\eta}(t)$ as we wish, the evolution of J decouples from our problem of optimizing the evolution of E. For learning rules where $\mathscr{F}[J;Jx,y]$ truly depends on J, on the other hand (i.e. where (45) does not hold), optimization of the error relaxation is considerably more difficult, and is likely to depend on the particular time t for which one wants to minimize $E(t)$. We will not deal with such cases here.

If the 'greedy' recipe applies (for spherical rules and for ordinary ones with the property (45)) working out the derivative in (44) immediately gives us

$$\tilde{\eta}(t)_{\mathrm{opt}} = \frac{\langle\{|y| - \cos(\pi E)x\,\mathrm{sgn}(y)\}\mathcal{F}[1;x,y]\rangle}{\cos(\pi E)\langle\mathcal{F}^2[1;x,y]\rangle} \tag{48}$$

Insertion of this choice into Equation 40 subsequently leads to

$$\left.\frac{\mathrm{d}E}{\mathrm{d}t}\right|_{\mathrm{opt}} = -\frac{\langle\{|y| - \cos(\pi E)x\,\mathrm{sgn}(y)\}\mathcal{F}[1;x,y]\rangle^2}{2\pi\sin(\pi E)\cos(\pi E)\langle\mathcal{F}^2[1;x,y]\rangle} \tag{49}$$

These and subsequent expressions we will write in terms of $\tilde{\eta}$, defined as $\tilde{\eta}(t) = \eta(t)$ for the spherical learning rules and as $\tilde{\eta}(t) = \eta(t)J(t)/\gamma(J(t))$ for the non-spherical learning rules. We will now work out the details of the results (48, 49) upon making the familiar choices for the function $\mathcal{F}[\ldots]$: the Hebbian, perceptron and AdaTron rules.

For the (ordinary and spherical) Hebbian rules, corresponding to $\mathcal{F}[J;Jx,y] = 1$, the various Gaussian integrals in (48, 49) are the same as those we already met (analytically) in the case of constant learning rate η. Substitution of the outcomes of the integrals (see the appendix) into Equations 48 and 49 gives

$$\tilde{\eta}_{\mathrm{opt}} = \sqrt{\frac{2}{\pi}}\frac{\sin^2(\pi E)}{\cos(\pi E)} \qquad \left.\frac{\mathrm{d}E}{\mathrm{d}t}\right|_{\mathrm{opt}} = -\frac{\sin^3(\pi E)}{\pi^2\cos(\pi E)}$$

The equation for the error E can be solved explicitly, giving (to be verified by substitution):

$$t(E) = \tfrac{1}{2}\pi\sin^{-2}(\pi E) - \tfrac{1}{2}\pi\sin^{-2}(\pi E_0) \tag{50}$$

The asymptotic behaviour of the process follows from expansion of (50) for small E, and gives

$$E_{\mathrm{opt}} \sim \frac{1}{\sqrt{2\pi t}} \qquad \tilde{\eta}_{\mathrm{opt}} \sim \sqrt{\frac{\pi}{2}}\frac{1}{t} \qquad (t \to \infty)$$

Asymptotically there is nothing to be gained by choosing the optimal time-dependent learning rate, since the same asymptotic form for E was also obtained for constant η (see (22)). Note that the property $\mathcal{F}[J;Jx,y] = \mathcal{F}[1;x,y]$ of the Hebbian recipe guarantees that the result (50) applies to both the ordinary and the spherical Hebbian rule. The only difference between the two cases is in the definition of $\tilde{\eta}$: for the ordinary (non-spherical) version $\tilde{\eta}(t) = \eta(t)/J(t)$, whereas for the spherical version $\tilde{\eta}(t) = \eta(t)$.

We move on to the (ordinary and spherical) perceptron learning rules, where $\mathcal{F}[J;Jx,y] = \theta[-xy]$, with time-dependent learning rates $\eta(t)$ which we aim to optimize. As in the Hebbian case all integrals occurring in (48, 49) upon substitution of the present choice $\mathcal{F}[J;Jx,y] = \theta[-xy]$ have been done already (see the appendix). Insertion of the outcomes of these integrals into (48, 49) gives

$$\tilde{\eta}_{\mathrm{opt}} = \frac{\sin^2(\pi E)}{\sqrt{2\pi}E\cos(\pi E)} \qquad \left.\frac{\mathrm{d}E}{\mathrm{d}t}\right|_{\mathrm{opt}} = -\frac{\sin^3(\pi E)}{4\pi^2 E\cos(\pi E)}$$

Again the non-linear differential equation describing the evolution of the error E can be solved exactly:

$$t(E) = \frac{2[\pi E + \sin(\pi E)\cos(\pi E)]}{\sin^2(\pi E)} - \frac{2[\pi E_0 + \sin(\pi E_0)\cos(\pi E_0)]}{\sin^2(\pi E_0)} \tag{51}$$

Expansion of (51) for small E gives the asymptotic behaviour

$$E_{\mathrm{opt}} \sim \frac{4}{\pi t} \qquad \tilde{\eta}_{\mathrm{opt}} \sim \frac{2\sqrt{2\pi}}{t} \qquad (t \to \infty)$$

which is identical to that found in the beginning of this section, i.e. Equations 31 and 35, upon exploring the consequences of making two simple *ad hoc* choices for the time-dependent learning rate (since $\tilde{\eta} = \eta/J$). As with the Hebbian rule, the property $\mathcal{F}[J;Jx,y] = \mathcal{F}[1;x,y]$ of the perceptron recipe guarantees that the result (51) applies to both the ordinary and the spherical version.

Finally we try to optimize the learning rate for the spherical AdaTron learning rule, corresponding to the choice $\mathcal{F}[J;Jx,y] = |Jx|\theta[-xy]$. Working out the averages in (48, 49) again does not require doing any new integrals. Using those already encountered in analysing the AdaTron rule with constant learning rate (to be found in the appendix), we obtain

$$\tilde{\eta}_{\mathrm{opt}} = \frac{\sin^3(\pi E)}{\pi}\left[E\cos(\pi E) - \frac{\cos^2(\pi E)\sin(\pi E)}{\pi}\right]^{-1}$$

$$\left.\frac{\mathrm{d}E}{\mathrm{d}t}\right|_{\mathrm{opt}} = -\frac{\sin^5(\pi E)}{2\pi^2\cos(\pi E)}\left[\frac{1}{\pi E - \cos(\pi E)\sin(\pi E)}\right]$$

(note that in both versions, ordinary and spherical, of the AdaTron rule we simply have $\tilde{\eta}(t) = \eta(t)$). It will no longer come as a surprise that this equation for the evolution of the error also allows for an analytical solution:

$$t(E) = \frac{\pi}{8}\left[\frac{4\pi E - \sin(4\pi E)}{\sin^4(\pi E)} - \frac{4\pi E_0 - \sin(4\pi E_0)}{\sin^4(\pi E_0)}\right] \tag{52}$$

Asymptotically, we find, upon expanding (52) for small E, a relaxation of the form

$$E_{\mathrm{opt}} \sim \tfrac{4}{3t} \qquad \tilde{\eta}_{\mathrm{opt}} \sim \tfrac{3}{2} \qquad (t \to \infty)$$

So for the AdaTron rule the asymptotic behaviour for optimal time-dependent learning rate η is identical to that found for optimal *constant* learning rate η (which is indeed $\eta = \frac{3}{2}$, see (30)). As with the previous two rules, the property $\mathcal{F}[J;Jx,y] = J\mathcal{F}[1;x,y]$ of the AdaTron recipe guarantees that the result (50) applies to both the ordinary and the spherical version.

It is quite remarkable that the simple perceptron learning rule, which came out at the bottom of the league among the three learning rules considered so far in the case of having constant learning rates, shoots to the top as soon as we allow for optimized time-dependent learning rates. It is in addition quite satisfactory that in a number of cases one can actually find an explicit expression for the relation $t(E)$ between the duration of the learning stage and the generalization error achieved, i.e. Equations 34, 50, 51 and 52.

3.4. *Optimal on-line learning rules*

We need not restrict our optimization attempts to varying the learning rate η only, but we can also vary the full form $\eta\mathscr{F}[J;Jx,y]$ of the learning rule. The aim, as always, is to minimize the generalization error, but there will be limits to what is achievable. So far all examples of on-line learning rules we have studied gave an asymptotic relaxation of the error of the form $E \sim t^{-q}$ with $q \leq 1$. It can be shown using general probabilistic arguments that if one only has $p = \alpha N$ examples of randomly drawn question/answer pairs $\{\xi^\mu, T(\xi^\mu)\}$ with which to calculate the weight vector $\boldsymbol{J}$ of an N-neuron binary student perceptron (whether in an on-line or a batch fashion), the generalization error $E_{\mathrm{g}}(\boldsymbol{J})$ obeys the inequality $E_{\mathrm{g}}(\boldsymbol{J}) \geq 0.44\ldots/\alpha$ for $N \to \infty$ (this is the one result we will mention without derivation). For on-line learning rules of the class (6) or (36, 38) we have used at time t a number of examples $p \leq tN$, so this inequality translates into

$$\lim_{t\to\infty} tE(t) \geq 0.44\ldots \tag{53}$$

No on-line learning rule can violate (53).* On the other hand, we have already encountered several rules with at least the optimal power $E \sim t^{-1}$. The optimal on-line learning rule is thus one which gives asymptotically $E \sim A/t$, but with the smallest value of A possible.

The function $\mathscr{F}[J;Jx,y]$ in the learning rules is allowed to depend only on the *sign* of the teacher field $y = \boldsymbol{B}\cdot\boldsymbol{\xi}$, not on its magnitude, since otherwise it would describe a situation where considerably more than just the answers $T(\boldsymbol{\xi}) = \mathrm{sgn}[\boldsymbol{B}\cdot\boldsymbol{\xi}]$ of the teacher are used for updating the parameters of the student. One can easily see that using unavailable information indeed violates (53). Suppose, for instance, we would consider spherical on-line rules, i.e. (36) or (38), and make the forbidden choice

$$\eta\mathscr{F}[1;x,y] = \frac{|y| - \cos(\pi E)x\,\mathrm{sgn}(y)}{\cos(\pi E)}$$

We would then find for the corresponding Equation 43 describing the evolution of the error E for $N \to \infty$:

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{\langle[|y| - \cos(\pi E)x\,\mathrm{sgn}(y)]^2\rangle}{2\pi\sin(\pi E)\cos(\pi E)}$$

(with averages as always calculated with the distribution (14)) from which it follows, upon using the Gaussian integrals done in the appendix:

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{\tan(\pi E)}{2\pi}$$

This produces exponential decay of the error, and thus indeed violates (53).

Taking into account the restrictions on available information, and anticipating the form subsequent expressions will take, we write the function $\mathscr{F}[J;Jx,y]$ (which we will be varying, and which we will also allow to have an explicit time-dependence*) in the following form

$$\eta\mathscr{F}[J;Jx,y] = \begin{cases} J\mathscr{F}_+(x,t) & \text{if} \quad y > 0 \\ J\mathscr{F}_-(x,t) & \text{if} \quad y < 0 \end{cases} \tag{54}$$

If our learning rule is of the general form (6), without spherical normalization, the coupled Equations 41 and 42 describe the macroscopic dynamics. For the spherical rules (36, 38) we have the single macroscopic Equation 43. Both (42) and (43) now acquire the form

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{1}{\pi\sin(\pi E)}\Big\{&\langle(y-\omega x)\theta[y]\mathscr{F}_+(x,t)\rangle \\ &- \langle(y-\omega x)\theta[-y]\mathscr{F}_-(x,t)\rangle \\ &- \frac{1}{2}\omega\langle\theta[y]\mathscr{F}_+^2(x,t)\rangle \\ &- \frac{1}{2}\omega\langle\theta[-y]\mathscr{F}_-^2(x,t)\rangle\Big\}\end{aligned} \tag{55}$$

with the usual short-hand $\omega = \cos(\pi E)$ and with averages calculated with the (time-dependent) distribution (14). To simplify notation we now introduce the two functions

$$\int \mathrm{d}y\,\theta[y]P(x,y) = \Omega(x,t)$$
$$\int \mathrm{d}y\,\theta[y](y-\omega x)P(x,y) = \Delta(x,t)$$

and hence, using the symmetry $P_t(x,y) = P_t(-x,-y)$, Equation 55 acquires the compact form

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}E = &-\frac{1}{\pi\sin(\pi E)}\int\mathrm{d}x\Big\{\Delta(x,t)\mathscr{F}_+(x,t) \\ &\qquad - \frac{1}{2}\omega\Omega(x,t)\mathscr{F}_+^2(x,t)\Big\} \\ &-\frac{1}{\pi\sin(\pi E)}\int\mathrm{d}x\Big\{\Delta(-x,t)\mathscr{F}_-(x,t) \\ &\qquad - \frac{1}{2}\omega\Omega(-x,t)\mathscr{F}_-^2(x,t)\Big\}\end{aligned} \tag{56}$$

* This will be different for graded-response perceptrons.

* By allowing for an explicit time-dependence, we can drop the dependence on J in $\mathscr{F}[J;Jx,y]$ if we wish, without loss of generality, since J is itself just some function of time.

Since there is only one dynamical variable, the error E, our optimization problem is solved by the 'greedy' recipe which here involves functional derivatives:

$$\forall x, \forall t: \quad \frac{\delta}{\delta\mathscr{F}_+(x,t)}\left[\frac{\mathrm{d}E}{\mathrm{d}t}\right] = \frac{\delta}{\delta\mathscr{F}_-(x,t)}\left[\frac{\mathrm{d}E}{\mathrm{d}t}\right] = 0$$

with the solution

$$\mathscr{F}_+(x,t) = \frac{\Delta(x,t)}{\omega\Omega(x,t)} \qquad \mathscr{F}_-(x,t) = \frac{\Delta(-x,t)}{\omega\Omega(-x,t)} = \mathscr{F}_+(-x,t)$$

Substitution of this solution into (56) gives the corresponding law describing the optimal error evolution of (ordinary and spherical) on-line rules:

$$\left.\frac{\mathrm{d}E}{\mathrm{d}t}\right|_{\mathrm{opt}} = -\frac{1}{\pi\sin(\pi E)\cos(\pi E)}\int\!\mathrm{d}x\,\frac{\Delta^2(x,t)}{\Omega(x,t)}$$

Explicit calculation of the integrals $\Delta(x,t)$ and $\Omega(x,t)$ (see the appendix) gives:

$$\Delta(x,t) = \frac{\sin(\pi E)}{2\pi}\mathrm{e}^{-\frac{1}{2}x^2/\sin^2(\pi E)}$$

$$\Omega(x,t) = \frac{\mathrm{e}^{-\frac{1}{2}x^2}}{2\sqrt{2\pi}}\left[1 + \mathrm{erf}\left(x/\sqrt{2}\tan(\pi E)\right)\right]$$

with which we finally obtain an explicit expression for the optimal form of the learning rule, via (54), as well as for the dynamical law describing the corresponding error evolution:

$$\eta\mathscr{F}[J; Jx, y]_{\mathrm{opt}} = \sqrt{\frac{2}{\pi}}\frac{J\tan(\pi E)\mathrm{e}^{-\frac{1}{2}x^2/\tan^2(\pi E)}}{1 + \mathrm{sgn}(xy)\mathrm{erf}\left(|x|/\sqrt{2}\tan(\pi E)\right)} \qquad (57)$$

$$\left.\frac{\mathrm{d}E}{\mathrm{d}t}\right|_{\mathrm{opt}} = -\frac{\tan^2(\pi E)}{\pi^2\sqrt{2\pi}} \times \int\!\mathrm{d}x\,\frac{\mathrm{e}^{-\frac{1}{2}x^2[1+\cos^2(\pi E)]/\cos^2(\pi E)}}{1 + \mathrm{erf}(x/\sqrt{2})} \qquad (58)$$

The asymptotic form of the error relaxation towards the $E = 0$ state follows from expansion of Equation 58 for small E, which gives

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -E^2\int\!\mathrm{d}y\,\frac{\mathrm{e}^{-y^2}}{\sqrt{2\pi}[1 + \mathrm{erf}(y/\sqrt{2})]} + \mathcal{O}(E^4)$$

so that we can conclude that the optimum asymptotic decay for on-line learning rules (whether spherical or non-spherical) is given by $E \sim A/t$ for $t \to \infty$, with

$$A^{-1} = \int\!\mathrm{d}x\,\frac{\mathrm{e}^{-x^2}}{\sqrt{2\pi}[1 + \mathrm{erf}(x/\sqrt{2})]}$$

Numerical evaluation of this integral (which is somewhat delicate due to the behaviour of the integrand for $x \to -\infty$) finally gives

$$E \sim \frac{0.883\ldots}{t} \qquad (t \to \infty)$$

It is instructive to investigate briefly the form of the optimal learning rule (57) for large values of E (as in the initial stages of learning processes) and for small values of E (as in the final stages of learning processes). Initially we find

$$\lim_{E\uparrow\frac{1}{2}}\frac{\eta\mathscr{F}[J; Jx, y]_{\mathrm{opt}}}{\tan(\pi E)} = J\sqrt{\frac{2}{\pi}}$$

which describes a Hebbian-type learning rule with diverging learning rate (note that $\tan(\pi E) \to \infty$ for $E \uparrow \frac{1}{2}$). In contrast, in the final stages the optimal learning rule (57) acquires the form

$$\begin{aligned}\lim_{E\downarrow 0}\eta\mathscr{F}[J; Jx, y]_{\mathrm{opt}} &= \frac{J|x|}{\sqrt{\pi}}\theta[-xy]\lim_{z\to\infty}\frac{\mathrm{e}^{-z^2}}{z[1 - \mathrm{erf}(z)]} \\ &= J|x|\theta[-xy]\end{aligned}$$

which is the AdaTron learning rule with learning rate $\eta = 1$.*

In Figs. 11 (short times and ordinary axes) and 12 (large times and log–log axes) we finally compare the evolution of the error for the optimal on-line learning rule (57) with the two on-line learning rules which so far were found to give the fastest relaxation: the perceptron rule with normalizing time-dependent learning rate (giving the error of (34)), and the perceptron rule with optimal time-dependent learning rate (giving the error of (51)). This, in order to assess whether choosing the optimal on-line learning rule (57) rather than its simpler competitors, is actually worth the effort. The curves for the optimal on-line rule were obtained by numerical solution of Equation 58.

3.5. *Summary in a table*

We close this section with an overview of some of the results on on-line learning in perceptrons described/derived so far. The upper part of Table 1 contains results for specific learning rules with either arbitrary constant learning rates η (first column), optimal constant learning rate η (second column), and where possible, a time-dependent learning rate $\eta(t)$ chosen such as to realize the normalization $J(t) = 1$ for all t. The lower part of the table gives results for specific learning rules with optimized time-dependent learning rates $\eta(t)$, as well as lower bounds on the asymptotic generalization error.

* The reason that, in spite of the asymptotic equivalence of the two rules, the optimal rule does not asymptotically give the same relaxation of the error E as the AdaTron rule is that in order to determine the asymptotics one has to take the limit $E \to 0$ in the full macroscopic differential equation for E, which, in addition to the function $\mathscr{F}[\ldots]$ defining the learning rule, involves the Gaussian probability distribution (14) which depends on E in a non-trivial way, especially near $E = 0$.

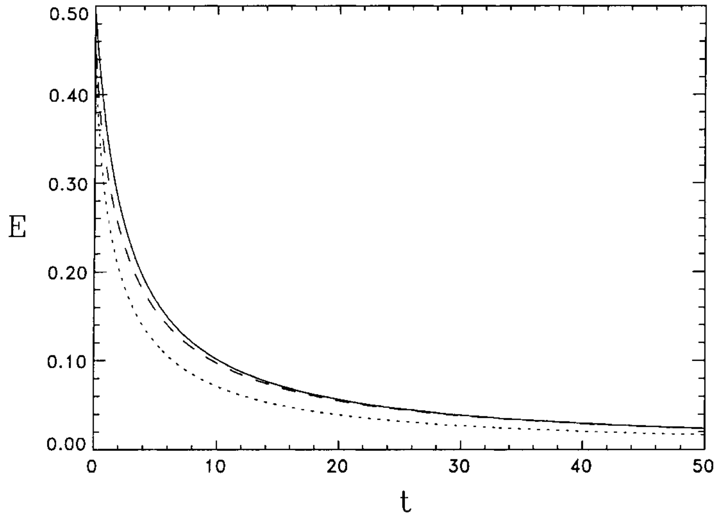

Fig. 11. *Evolution of the error E for three on-line learning rules: perceptron rule with a learning rate such that $J(t) = 1$ for all $t \geq 0$ (solid line), perceptron rule with optimal learning rate (dashed line) and the optimal spherical learning rule (dotted line). Initial state: $E(0) = 1/2$ and $J(0) = 1$. The curves for the perceptron rules are given by (34) and (51). The curve for the optimal spherical rule was obtained by numerical solution of Equation 58*

4. The formal approach

The main reason for developing a more formal approach to learning dynamics is that in the complicated cases of incomplete training sets or layered systems with large numbers of hidden neurons we can no longer get away with the relatively simple methods used so far. For perceptrons with N inputs the situation of incomplete training sets arises when the number of 'questions' scales as $|\tilde{D}| = \alpha N$; here the training error is no longer identical to the generalization error, see Fig. 13. We show how in the limit $N \to \infty$ the dynamics of any finite set of mean-field observables will be described by a (macroscopic) Fokker–Planck equation, of which the flow and diffusion terms can be calculated explicitly. In addition, the more formal analysis will allow us to recover the previous results on on-line learning in a more rigorous way, and will clarify the relation between the macroscopic laws for the on-line and batch scenarios.

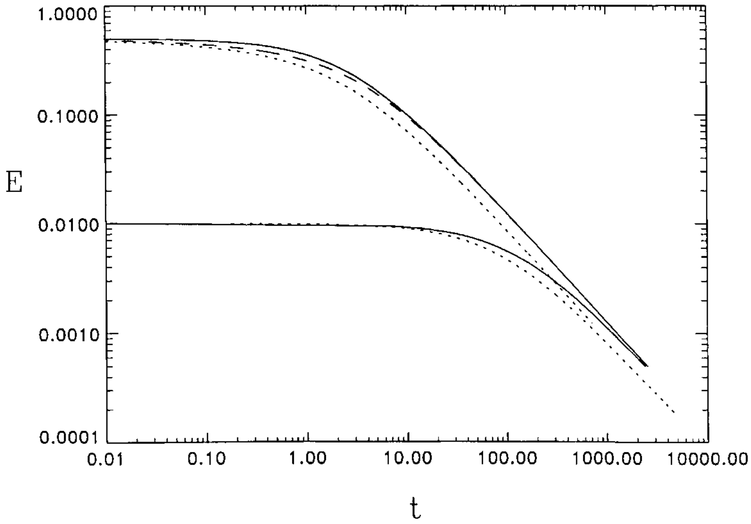

Fig. 12. *Evolution of the error E for the on-line perceptron rule with a learning rate such that $J(t) = 1$ for all $t \geq 0$ (solid line), the on-line perceptron rule with optimal learning rate (dashed line) and the optimal spherical on-line learning rule (dotted line). Initial states: $(J,E) = (1,1/2)$ (upper curves), and $(J,E) = (1,1/100)$ (lower curves). The curves for the perceptron rules are given by (34) and (51). The curves for the optimal spherical rule were obtained by numerical solution of Equation 58*

Table 1. *Overview of exact results on on-line learning rules for perceptrons with complete training sets, in the limit $N \to \infty$ (an infinite number of inputs)*

Generalization error in perceptrons with on-line learning rules

	Constant learning rate η		Variable η
Rule	Asymptotic decay for constant η	Optimal asymptotic decay for constant η	(η chosen to normalize J)
Hebbian	$E \sim \frac{1}{\sqrt{2\pi}} t^{-1/2}$ for $\eta > 0$	$E \sim \frac{1}{\sqrt{2\pi}} t^{-1/2}$ for $\eta > 0$	N/A
Perceptron	$E \sim \left(\frac{2}{3}\right)^{1/3} \pi^{-1} t^{-1/3}$ for $\eta > 0$	$E \sim \left(\frac{2}{3}\right)^{1/3} \pi^{-1} t^{-1/3}$ for $\eta > 0$	$E \sim \frac{4}{\pi} t^{-1}$
AdaTron	$E \sim \left(\frac{3}{3\eta-\eta^2}\right) t^{-1}$ for $0 < \eta < 3$	$E \sim \frac{4}{3} t^{-1}$ for $\eta = \frac{3}{2}$	$E \sim \frac{3}{2} t^{-1}$

Optimal generalization

	Optimal time-dependent learning rate η	
Rule	Generalization error for optimal time-dependent η	Asymptotics
Hebbian	$t = \frac{\pi}{2}\left[\frac{1}{\sin^2(\pi E)} - \frac{1}{\sin^2(\pi E_0)}\right]$	$E \sim \frac{1}{\sqrt{2\pi}} t^{-1/2}$
Perceptron	$t = 2\left[\frac{\pi E + \sin(\pi E)\cos(\pi E)}{\sin^2(\pi E)} - \frac{\pi E_0 + \sin(\pi E_0)\cos(\pi E_0)}{\sin^2(\pi E_0)}\right]$	$E \sim \frac{4}{\pi} t^{-1}$
AdaTron	$t = \frac{\pi}{8}\left[\frac{4\pi E - \sin(4\pi E)}{\sin^4(\pi E)} - \frac{4\pi E_0 - \sin(4\pi E_0)}{\sin^4(\pi E_0)}\right]$	$E \sim \frac{4}{3} t^{-1}$
Lower bound for on-line learning (asymptotics of the optimal learning rule)		$E \sim 0.88t^{-1}$
Lower bound for any learning rule		$E \sim 0.44t^{-1}$

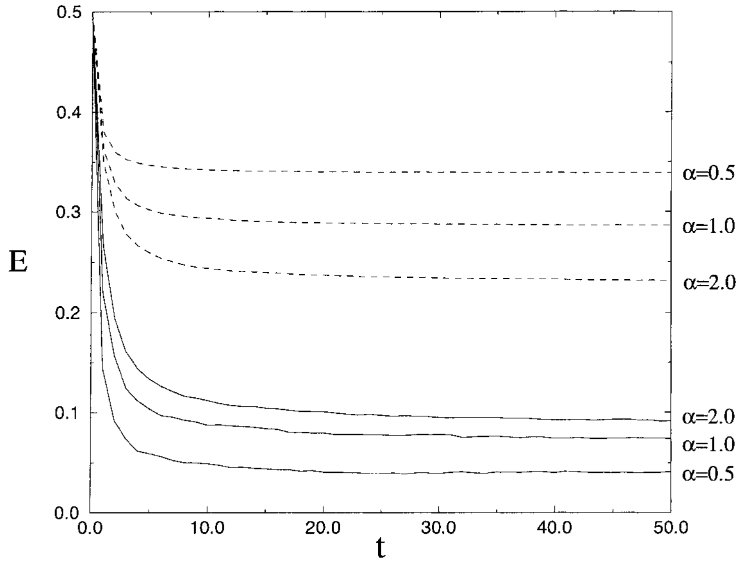

Fig. 13. *Evolution in time of the generalization errors E_g (dashed lines) and the training errors E_t (solid lines) as measured during numerical simulations (with N =10 000 neurons) of the Hebbian learning rule, for three different sizes of the (restricted) training set (α =0.5,1.0,2.0). Initial state: $E(0) = 1/2$ and $J(0) = 1$. Learning rate: $\eta = 1$*

4.1. *From discrete to continuous times*

We will describe the formal procedure for calculating macroscopic dynamical laws from the microscopic ones for on-line and batch learning process in simple perceptrons. It involves several distinct stages. Our starting point is the formulation (2) in terms of a Markov process:

$$\hat{p}_{m+1}(\boldsymbol{J}) = \int \mathrm{d}\boldsymbol{J}' W[\boldsymbol{J};\boldsymbol{J}']\hat{p}_m(\boldsymbol{J}') \tag{59}$$

with transition probability densities corresponding to the class (6) of generic learning rules*

$$\begin{aligned} \text{On-Line:} \quad W[\boldsymbol{J};\boldsymbol{J}'] &= \Big\langle \delta\Big\{\boldsymbol{J} - \boldsymbol{J}' - \frac{\eta}{N}\xi\,\mathrm{sgn}(\boldsymbol{B}\cdot\xi) \\ &\qquad \times\mathscr{F}[|\boldsymbol{J}'|;\boldsymbol{J}'\cdot\xi,\boldsymbol{B}\cdot\xi]\Big\}\Big\rangle_{\tilde{D}} \\ \text{Batch:} \quad W[\boldsymbol{J};\boldsymbol{J}'] &= \delta\Big\{\boldsymbol{J} - \boldsymbol{J}' - \frac{\eta}{N}\langle\xi\,\mathrm{sgn}(\boldsymbol{B}\cdot\xi) \\ &\qquad \times\mathscr{F}[|\boldsymbol{J}'|;\boldsymbol{J}'\cdot\xi,\boldsymbol{B}\cdot\xi]\rangle_{\tilde{D}}\Big\} \end{aligned} \tag{60}$$

Note that in the previous approach the limit $N \to \infty$ realized several simplifications at once (continuous versus discrete time, stochastic versus deterministic macroscopic evolution) which for technical reasons we would prefer to control independently.

We will first describe a method to make the transition from the discrete-time process (59) to a description involving real-valued times in a more transparent and exact way. The idea is to choose the duration of each discrete iteration step in the process (59) to be a real-valued random number, such that the probability that at time t precisely m steps have been made is given by the Poisson expression

$$\pi_m(t) = \frac{1}{m!}(Nt)^m \mathrm{e}^{-Nt}$$

with the properties

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\pi_{m>0}(t) &= N[\pi_{m-1}(t) - \pi_m(t)] \\ \frac{\mathrm{d}}{\mathrm{d}t}\pi_0(t) &= -N\pi_0(t) \end{aligned} \tag{61}$$

$$\langle m\rangle = Nt \qquad \langle m^2\rangle - \langle m\rangle^2 = Nt \tag{62}$$

This move at first sight appears to make the problem more complicated, but will turn out to do precisely the opposite. From (62) it follows that for times $t \ll N$ one has $t = m/N + \mathcal{O}(N^{-\frac{1}{2}})$, the usual time unit. Due to the random durations of the iteration steps we also have to replace the microscopic probability distribution $\hat{p}_m(\boldsymbol{J})$ in (59) by one that takes the variations in numbers of iteration steps performed at a given time t into account:

$$p_t(\boldsymbol{J}) = \sum_{m\geq 0}\pi_m(t)\hat{p}_m(\boldsymbol{J}) \tag{63}$$

This distribution obeys a simple differential equation, which follows from combining Equations 59, 61, 62, and 63:

$$\frac{\mathrm{d}}{\mathrm{d}t}p_t(\boldsymbol{J}) = N\int \mathrm{d}\boldsymbol{J}'\{W[\boldsymbol{J};\boldsymbol{J}'] - \delta[\boldsymbol{J}-\boldsymbol{J}']\}\,p_t(\boldsymbol{J}') \tag{64}$$

So far no approximations have been made, Equation 64 which replaces (59) is exact for any N. We have made the transition from discrete-time iterations to differential equations (which are usually much easier to handle) without invoking the limit $N \to \infty$, but at the price of an uncertainty in where we are on the time axis. This uncertainty, however, is guaranteed to vanish in the limit $N \to \infty$.

4.2. *From microscopic to macroscopic laws*

We next wish to investigate the dynamics of a number of as yet arbitrary *macroscopic* observables $\boldsymbol{\Omega}[\boldsymbol{J}] = (\Omega_1[\boldsymbol{J}],\ldots,\Omega_k[\boldsymbol{J}])$. They are assumed to be $\mathcal{O}(1)$ each for $N \to \infty$, and finite in number. To do so we introduce the associated macroscopic probability distribution

$$P_t(\boldsymbol{\Omega}) = \int \mathrm{d}\boldsymbol{J}p_t(\boldsymbol{J})\delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}[\boldsymbol{J}]] \tag{65}$$

Its time derivative immediately follows from that in (64):

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}p_t(\boldsymbol{\Omega}) = N\int &\mathrm{d}\boldsymbol{J}\,\mathrm{d}\boldsymbol{J}'\,\delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}[\boldsymbol{J}]] \\ &\times\{W[\boldsymbol{J};\boldsymbol{J}'] - \delta[\boldsymbol{J}-\boldsymbol{J}']\}\,p_t(\boldsymbol{J}') \end{aligned}$$

This equation can be written in the standard form

$$\frac{\mathrm{d}}{\mathrm{d}t}p_t(\boldsymbol{\Omega}) = \int \mathrm{d}\boldsymbol{\Omega}'\mathscr{W}_t[\boldsymbol{\Omega};\boldsymbol{\Omega}']\,p_t(\boldsymbol{\Omega}') \tag{66}$$

where

$$\begin{aligned} \mathscr{W}_t[\boldsymbol{\Omega};\boldsymbol{\Omega}'] = {}&[\textstyle\int \mathrm{d}\boldsymbol{J}'\,p_t(\boldsymbol{J}')\delta[\boldsymbol{\Omega}' - \boldsymbol{\Omega}[\boldsymbol{J}']]]^{-1} \\ &\times[\textstyle\int \mathrm{d}\boldsymbol{J}'p_t(\boldsymbol{J}')\delta[\boldsymbol{\Omega}' - \boldsymbol{\Omega}[\boldsymbol{J}']]\int \mathrm{d}J\delta[\Omega - \Omega[J]] \\ &\times N\{W[J;J'] - \delta[J - J']\}] \end{aligned}$$

(this statement can be verified by substitution of $\mathscr{W}_t[\boldsymbol{\Omega};\boldsymbol{\Omega}']$ into (66)). Note that the macroscopic process (66) need not be Markovian, since the explicit time-dependence of the macroscopic transition density $\mathscr{W}_t[\boldsymbol{\Omega};\boldsymbol{\Omega}']$ requires knowledge of the microscopic probability distribution $p_t(\boldsymbol{J})$. If we now insert the relevant expressions (60) for $W[\boldsymbol{J};\boldsymbol{J}']$, we can perform the $\boldsymbol{J}$-integrations, and obtain an expression in terms of so-called sub-shell averages (or conditional averages) $\langle f(\boldsymbol{J})\rangle_{\boldsymbol{\Omega};t}$, which are defined as

$$\langle f(\boldsymbol{J})\rangle_{\boldsymbol{\Omega};t} = \frac{\int \mathrm{d}\boldsymbol{J}p_t(\boldsymbol{J})\delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}[\boldsymbol{J}]]f(\boldsymbol{J})}{\int \mathrm{d}\boldsymbol{J}p_t(\boldsymbol{J})\delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}[\boldsymbol{J}]]}$$

For the two types of learning rules at hand (on-line and batch) we obtain (upon replacing the remaining dummy variables $\boldsymbol{J}'$ by $\boldsymbol{J}$):

[*] As before this is just one choice of many. We could, for example, easily add a term of the form $(\eta/N)\boldsymbol{J}'\mathscr{K}[|\boldsymbol{J}'|;\boldsymbol{J}'\cdot\xi,\boldsymbol{B}\cdot\xi]$ to account for weight decay (constant, ‘hard’ spherical, ‘soft’ spherical, or otherwise), without making the analysis significantly more difficult.

$$\mathscr{W}_t^{\rm onl}[\mathbf{\Omega};\mathbf{\Omega}'] = N\Big\langle\Big\langle\delta\Big[\mathbf{\Omega}-\mathbf{\Omega}\Big[\boldsymbol{J}+\frac{\eta}{N}\xi\,{\rm sgn}\,(\boldsymbol{B}\cdot\xi)\times\mathscr{F}[|\boldsymbol{J}|;\boldsymbol{J}\cdot\xi,\boldsymbol{B}\cdot\xi]\Big]\Big]\Big\rangle_{\tilde{D}} - \delta[\mathbf{\Omega}-\mathbf{\Omega}[\boldsymbol{J}]]\Big\rangle_{\mathbf{\Omega}';t}$$

$$\mathscr{W}_t^{\rm bat}[\mathbf{\Omega};\mathbf{\Omega}'] = N\Big\langle\delta\Big[\mathbf{\Omega}-\mathbf{\Omega}\Big[\boldsymbol{J}+\frac{\eta}{N}\times\Big\langle\xi\,{\rm sgn}\,(\boldsymbol{B}\cdot\xi)\mathscr{F}[|\boldsymbol{J}|;\boldsymbol{J}\cdot\xi,\boldsymbol{B}\cdot\xi]\Big\rangle_{\tilde{D}}\Big]\Big] - \delta[\mathbf{\Omega}-\mathbf{\Omega}[\boldsymbol{J}]]\Big\rangle_{\mathbf{\Omega}';t}$$

We now insert integral representations for the δ-distributions

$$\delta[\mathbf{\Omega}-\boldsymbol{Q}] = \int\frac{{\rm d}\hat{\mathbf{\Omega}}}{(2\pi)^k}{\rm e}^{i\hat{\mathbf{\Omega}}\cdot[\mathbf{\Omega}-\boldsymbol{Q}]}$$

which gives for our two learning scenarios:

$$\mathscr{W}_t^{\rm onl}[\mathbf{\Omega};\mathbf{\Omega}'] = \int\frac{{\rm d}\hat{\mathbf{\Omega}}}{(2\pi)^k}{\rm e}^{i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}} \times N\Big\langle\Big\langle{\rm e}^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}\left[\boldsymbol{J}+\frac{\eta}{N}\xi\,{\rm sgn}(\boldsymbol{B}\cdot\xi)\mathscr{F}[|\boldsymbol{J}|;\boldsymbol{J}\cdot\xi,\boldsymbol{B}\cdot\xi]\right]}\Big\rangle_{\tilde{D}} - {\rm e}^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}]}\Big\rangle_{\mathbf{\Omega}';t} \tag{67}$$

$$\mathscr{W}_t^{\rm bat}[\mathbf{\Omega};\mathbf{\Omega}'] = \int\frac{{\rm d}\hat{\mathbf{\Omega}}}{(2\pi)^k}{\rm e}^{i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}} \times N\Big\langle{\rm e}^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}\left[\boldsymbol{J}+\frac{\eta}{N}\langle\xi\,{\rm sgn}(\boldsymbol{B}\cdot\xi)\mathscr{F}[|\boldsymbol{J}|;\boldsymbol{J}\cdot\xi,\boldsymbol{B}\cdot\xi]\rangle_{\tilde{D}}\right]} - {\rm e}^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}]}\Big\rangle_{\mathbf{\Omega}';t} \tag{68}$$

Still no approximations have been made. The above two expressions differ only in at which stage the averaging over the training set $\tilde{D}$ occurs.

Our aim is to obtain from (66) an autonomous set of macroscopic dynamic equations, i.e. we want to choose the observables $\mathbf{\Omega}[\boldsymbol{J}]$ such that for $N\to\infty$ the explicit time-dependence in $\mathscr{W}_t[\mathbf{\Omega};\mathbf{\Omega}']$, induced by the appearance of the microscopic distribution $p_t(\boldsymbol{J})$ will vanish. This can happen either because $p_t(\boldsymbol{J})$ drops out, or because $p_t(\boldsymbol{J})$ depends on $\boldsymbol{J}$ only via $\mathbf{\Omega}[\boldsymbol{J}]$, or even through combinations of these mechanisms. In expanding Equations 67 and 68 for large N we have to be somewhat careful, since the system size N enters both as a small parameter to control the magnitude of the modification of individual components of the weight vector $\boldsymbol{J}$, but also determines the dimensions and lengths of various vectors. Upon inspection of the general Taylor expansion

$$\Omega[\boldsymbol{J}+\boldsymbol{k}] = \sum_{\ell\geq 0}\frac{1}{\ell!}\sum_{i_1=1}^{N}\cdots\sum_{i_\ell=1}^{N}k_{i_1}\cdots k_{i_\ell}\frac{\partial^\ell\Omega[\boldsymbol{J}]}{\partial J_{i_1}\cdots\partial J_{i_\ell}}$$

we see that if all derivatives were to be treated as $\mathcal{O}(1)$ (i.e. if we only take into account the dependence of the components of $\boldsymbol{k}$ on N), we end up in trouble, since in the cases of interest (where $\boldsymbol{k}^2=\mathcal{O}(N^{-1})$) this series could give $\Omega[\boldsymbol{J}+\boldsymbol{k}]=\sum_{\ell\geq 0}(\sum_i k_i)^\ell=\sum_{\ell\geq 0}\mathcal{O}(1)$. We need to restrict ourselves to observables $\Omega_\mu[\boldsymbol{J}]$ of the mean-field type, where all components J_i play an equivalent role in determining the overall scaling with respect to N (which makes sense). For instance:

$$\begin{aligned}
\Omega[\boldsymbol{J}] &= \sum_k B_k J_k: & \mathcal{O}(\partial_i\Omega[\boldsymbol{J}]) &= \mathcal{O}(B_i)\\
& & &= N^{-1}\mathcal{O}(\Omega[\boldsymbol{J}])/\mathcal{O}(J_i)\\
\Omega[\boldsymbol{J}] &= \sum_k J_k^2: & \mathcal{O}(\partial_i\Omega[\boldsymbol{J}]) &= \mathcal{O}(J_i)\\
& & &= N^{-1}\mathcal{O}(\Omega[\boldsymbol{J}])/\mathcal{O}(J_i)\\
\Omega[\boldsymbol{J}] &= \sum_{kl} J_k A_{kl} J_l: & \mathcal{O}(\partial_i\Omega[\boldsymbol{J}]) &= \mathcal{O}\left(\sum_k A_{ik}J_k\right)\\
& & &= N^{-1}\mathcal{O}(\Omega[\boldsymbol{J}])/\mathcal{O}(J_i)
\end{aligned}$$

The pattern is clear. The only additional point to be taken into account is that in the case of multiple derivatives with respect to the *same* component J_i, our scaling requirement will be less severe due to the fact that such terms occur less frequently than multiple derivatives with respect to different components (i.e. in $\sum_{ij}J_iA_{ij}J_j$ we have $N(N-1)$ terms with $i\neq j$, but just N with $i=j$). We thus define mean-field observables $\Omega[\boldsymbol{J}]$ as mean-field observables:

$$\frac{\partial^\ell\Omega[\boldsymbol{J}]}{\partial J_{i_1}\cdots\partial J_{i_\ell}} = \mathcal{O}\left(N^{-\frac{1}{2}\ell}\frac{\Omega[\boldsymbol{J}]}{|\boldsymbol{J}|^\ell}\cdot N^{\ell-d}\right) \qquad (N\to\infty) \tag{69}$$

in which d is the number of *different* elements in the set $\{i_1,\ldots,i_\ell\}$. For mean-field observables we can estimate the scaling of the various terms in the Taylor expansion:

$$\Omega[\boldsymbol{J}+\boldsymbol{k}] = \Omega[\boldsymbol{J}]+\sum_i k_i\frac{\partial\Omega[\boldsymbol{J}]}{\partial J_i} + \frac{1}{2}\sum_{ij}k_ik_j\frac{\partial^2\Omega[\boldsymbol{J}]}{\partial J_i\partial J_j} + \sum_{\ell\geq 3}\mathcal{O}\left(\Omega[\boldsymbol{J}]\left[\frac{|\boldsymbol{k}|}{|\boldsymbol{J}|}\right]^\ell\right) \tag{70}$$

Here we have used $\sum_i k_i=\mathcal{O}(\sqrt{N}|\boldsymbol{k}|)$ (note that the order symbols describe worst-case scaling properties).

We now apply (70) to our Equations 67 and 68, restricting ourselves henceforth to mean-field observables $\Omega_\mu[\boldsymbol{J}]$ in the sense of (69). The shifts $\boldsymbol{k}$, being either $(\eta/N)\xi\ \ {\rm sgn}(\boldsymbol{B}\cdot\xi)\mathscr{F}[|\boldsymbol{J}|;\boldsymbol{J}\cdot\xi,\boldsymbol{B}\cdot\xi]$ or $(\eta/N)\langle\xi\ \ {\rm sgn}(\boldsymbol{B}\cdot\xi)\mathscr{F}[|\boldsymbol{J}|;\boldsymbol{J}\cdot\xi,\boldsymbol{B}\cdot\xi]\rangle_{\tilde{D}}$, scale as $|\boldsymbol{k}|=\mathcal{O}(N^{-\frac{1}{2}})$. Furthermore, if we choose one of our observables to be $\Omega_1[\boldsymbol{J}]=\boldsymbol{J}^2$, the subshells in (67, 68) will ensure $\boldsymbol{J}^2=\mathcal{O}(1)$, so that the ℓ-th order term in the expansions (70) will be of order $N^{-\frac{1}{2}\ell}$ in both cases. This allows us to expand:

$$
\begin{aligned}
&\mathrm{e}^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}+\boldsymbol{k}]}-\mathrm{e}^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}]}\\
&\quad=\mathrm{e}^{-i\hat{\mathbf{\Omega}}\cdot\left\{\mathbf{\Omega}[\boldsymbol{J}]+\sum_i k_i\frac{\partial}{\partial J_i}\mathbf{\Omega}[\boldsymbol{J}]+\frac{1}{2}\sum_{ij}k_ik_j\frac{\partial^2}{\partial J_i\partial J_j}\mathbf{\Omega}[\boldsymbol{J}]+\mathcal{O}\left(N^{-\frac{3}{2}}\right)\right\}}\\
&\qquad-\mathrm{e}^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}]}\\
&\quad=-\mathrm{e}^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}]}\left\{i\sum_i k_i\frac{\partial}{\partial J_i}(\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}])\right.\\
&\qquad\qquad+\frac{i}{2}\sum_{ij}k_ik_j\frac{\partial^2}{\partial J_i\partial J_j}(\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}])\\
&\qquad\qquad\left.+\frac{1}{2}\left[\sum_i k_i\frac{\partial}{\partial J_i}(\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}])\right]^2\right\}\\
&\qquad\qquad+\mathcal{O}\left(N^{-\frac{3}{2}}\right)
\end{aligned}
$$

so that

$$
\begin{aligned}
&N\int\frac{\mathrm{d}\hat{\mathbf{\Omega}}}{(2\pi)^k}\mathrm{e}^{i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}}\left[\mathrm{e}^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}+\boldsymbol{k}]}-\mathrm{e}^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}]}\right]\\
&\quad=-N\int\frac{\mathrm{d}\hat{\mathbf{\Omega}}}{(2\pi)^k}\mathrm{e}^{i\hat{\mathbf{\Omega}}\cdot[\mathbf{\Omega}-\mathbf{\Omega}[\boldsymbol{J}]]}\left\{i\sum_\mu\hat{\Omega}_\mu\sum_i k_i\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}\right.\\
&\qquad+\frac{i}{2}\sum_\mu\hat{\Omega}_\mu\sum_{ij}k_ik_j\frac{\partial^2\Omega_\mu[\boldsymbol{J}]}{\partial J_i\partial J_j}\\
&\qquad\left.+\frac{1}{2}\sum_{\mu\nu}\hat{\Omega}_\mu\hat{\Omega}_\nu\sum_{ij}k_ik_j\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}\frac{\partial\Omega_\nu[\boldsymbol{J}]}{\partial J_j}\right\}\\
&\qquad+\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\\
&\quad=-N\int\frac{\mathrm{d}\hat{\mathbf{\Omega}}}{(2\pi)^k}\left\{\sum_\mu\left[\sum_i k_i\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}\right.\right.\\
&\qquad\left.+\frac{1}{2}\sum_{ij}k_ik_j\frac{\partial^2\Omega_\mu[\boldsymbol{J}]}{\partial J_i\partial J_j}\right]\frac{\partial}{\partial\Omega_\mu}\\
&\qquad\left.-\frac{1}{2}\sum_{\mu\nu}\sum_{ij}k_ik_j\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}\frac{\partial\Omega_\nu[\boldsymbol{J}]}{\partial J_j}\frac{\partial^2}{\partial\Omega_\mu\partial\Omega_\nu}\right\}\\
&\qquad\times\mathrm{e}^{i\hat{\mathbf{\Omega}}\cdot[\mathbf{\Omega}-\mathbf{\Omega}[\boldsymbol{J}]]}+\mathcal{O}\left(N^{-\frac{1}{2}}\right)\\
&\quad=-N\left\{\sum_\mu\left[\sum_i k_i\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}+\frac{1}{2}\sum_{ij}k_ik_j\frac{\partial^2\Omega_\mu[\boldsymbol{J}]}{\partial J_i\partial J_j}\right]\frac{\partial}{\partial\Omega_\mu}\right.\\
&\qquad\left.-\frac{1}{2}\sum_{\mu\nu}\sum_{ij}k_ik_j\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}\frac{\partial\Omega_\nu[\boldsymbol{J}]}{\partial J_j}\frac{\partial^2}{\partial\Omega_\mu\partial\Omega_\nu}\right\}\\
&\qquad\times\delta[\mathbf{\Omega}-\mathbf{\Omega}[\boldsymbol{J}]]+\mathcal{O}\left(N^{-\frac{1}{2}}\right)
\end{aligned}
$$

We now find, upon insertion of this expansion into the expressions (67) and (68), that both types of learning dynamics (on-line and batch) are described by macroscopic laws with transition probabilities of the general form

$$
\begin{aligned}
\mathcal{W}_t^{\star\star\star}[\mathbf{\Omega};\mathbf{\Omega}']=&\left\{-\sum_\mu F_\mu[\mathbf{\Omega}';t]\frac{\partial}{\partial\Omega_\mu}\right.\\
&\left.+\frac{1}{2}\sum_{\mu\nu}G_{\mu\nu}[\mathbf{\Omega}';t]\frac{\partial^2}{\partial\Omega_\mu\partial\Omega_\nu}\right\}\delta[\mathbf{\Omega}-\mathbf{\Omega}']
\end{aligned}
$$

which, in combination with the dynamic Equation 66, leads to convenient and transparent description of the macroscopic dynamics in the form of a Fokker-Planck equation:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}P_t(\mathbf{\Omega})=&-\sum_{\mu=1}^k\frac{\partial}{\partial\Omega_\mu}\{F_\mu[\mathbf{\Omega};t]\,P_t(\mathbf{\Omega})\}\\
&+\frac{1}{2}\sum_{\mu\nu=1}^k\frac{\partial^2}{\partial\Omega_\mu\partial\Omega_\nu}\{G_{\mu\nu}[\mathbf{\Omega};t]\,P_t(\mathbf{\Omega})\}
\end{aligned}
\tag{71}
$$

(modulo contributions which vanish for $N\to\infty$). The differences between on-line and batch learning are in the explicit expressions for the functions $F_\mu[\mathbf{\Omega};t]$ and $G_{\mu\nu}[\mathbf{\Omega};t]$ in the flow and diffusion terms. Upon introducing the short hand $\mathscr{F}[\ldots]$ for $\mathscr{F}[|\boldsymbol{J}|;\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]$ these can be written as:

$$
\begin{aligned}
F_\mu^{\mathrm{onl}}[\mathbf{\Omega};t]=&\,\eta\left\langle\left\langle\sum_i\xi_i\,\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}\right\rangle_{\tilde{D}}\right\rangle_{\mathbf{\Omega};t}\\
&+\frac{\eta^2}{2N}\left\langle\left\langle\sum_{ij}\xi_i\xi_j\mathscr{F}^2[\ldots]\frac{\partial^2\Omega_\mu[\boldsymbol{J}]}{\partial J_i\partial J_j}\right\rangle_{\tilde{D}}\right\rangle_{\mathbf{\Omega};t}
\end{aligned}
\tag{72}
$$

$$
\begin{aligned}
G_{\mu\nu}^{\mathrm{onl}}[\mathbf{\Omega};t]=&\,\frac{\eta^2}{N}\left\langle\left\langle\sum_{ij}\xi_i\xi_j\mathscr{F}^2[\ldots]\left[\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}\right]\right.\right.\\
&\left.\left.\times\left[\frac{\partial\Omega_\nu[\boldsymbol{J}]}{\partial J_j}\right]\right\rangle_{\tilde{D}}\right\rangle_{\mathbf{\Omega};t}
\end{aligned}
\tag{73}
$$

$$
\begin{aligned}
F_\mu^{\mathrm{bat}}[\mathbf{\Omega};t]=&\,\eta\left\langle\sum_i\langle\xi_i\,\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\rangle_{\tilde{D}}\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}\right\rangle_{\mathbf{\Omega};t}\\
&+\frac{\eta^2}{2N}\left\langle\sum_{ij}\langle\xi_i\,\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\rangle_{\tilde{D}}\right.\\
&\left.\times\langle\xi_j\,\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\rangle_{\tilde{D}}\frac{\partial^2\Omega_\mu[\boldsymbol{J}]}{\partial J_i\partial J_j}\right\rangle_{\mathbf{\Omega};t}
\end{aligned}
\tag{74}
$$

$$
\begin{aligned}
G_{\mu\nu}^{\mathrm{bat}}[\mathbf{\Omega};t]=&\,\frac{\eta^2}{N}\left\langle\sum_{ij}\langle\xi_i\,\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\rangle\right.\\
&\times\langle\xi_j\,\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\rangle_{\tilde{D}}\\
&\left.\times\left[\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}\right]\left[\frac{\partial\Omega_\nu[\boldsymbol{J}]}{\partial J_j}\right]\right\rangle_{\mathbf{\Omega};t}
\end{aligned}
\tag{75}
$$

The result (71) is still fairly general. The only conditions on the observables $\Omega_\mu[\boldsymbol{J}]$ needed for (71) to hold are: (i) all are of order unity for $N\to\infty$; (ii) all are of the mean-field type (69); and (iii) one of them is the squared length $\boldsymbol{J}^2$ of the student's weight vector.

The Fokker-Planck Equation 71 subsequently quantifies the properties of the ideal choice(s) for our macroscopic observables $\Omega_\mu[\boldsymbol{J}]$, if our aim is to find closed deterministic equations. Firstly:

$$\text{deterministic laws:} \quad \lim_{N\to\infty} G_{\mu\nu}[\boldsymbol{\Omega}; t] = 0 \tag{76}$$

If (76) holds, Equation 71 reduces to a Liouville equation, with solutions of the desired form $P_t(\boldsymbol{\Omega}) = \delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}^*(t)]$ in which the trajectory $\boldsymbol{\Omega}^*(t)$, in turn, is the solution of the deterministic equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Omega} = \boldsymbol{F}[\boldsymbol{\Omega}; t] \tag{77}$$

with the flow field $\boldsymbol{F}$ given either by (72) (for on-line learning) or by (74) (for batch learning). Note that condition (76) is not only sufficient to guarantee deterministic evolution, but also necessary. Secondly, we want the deterministic laws to be closed. Since any remaining explicit time-dependence in the macroscopic laws induces a dependence on the microscopic distribution $p_t(\boldsymbol{J})$, we conclude:

$$\text{closed laws:} \quad \lim_{N\to\infty} \frac{\partial}{\partial t} F_\mu[\boldsymbol{\Omega}; t] = 0 \tag{78}$$

(again this condition is sufficient and necessary). A set of mean-field observables $\Omega_\mu[\boldsymbol{J}]$ meeting the criteria (76, 78) constitutes for $N \to \infty$ an exact autonomous macroscopic level of description of the learning process, in the form of the coupled deterministic differential equations (77). However, in general, there will be no a priori guarantee that such a set of observables actually exists.

4.3. *Application to* (Q, R) *evolution*

We now apply the general results of this section to the specific duo of observables that we considered in the previous sections to describe on-line learning with complete training sets:

$$\Omega_1[\boldsymbol{J}] = Q[\boldsymbol{J}] = \boldsymbol{J}^2 \qquad \Omega_2[\boldsymbol{J}] = R[\boldsymbol{J}] = \boldsymbol{J}\cdot\boldsymbol{B} \tag{79}$$

These observables are indeed of the mean-field type (69) if all $B_i = \mathcal{O}(N^{-\frac{1}{2}})$, and are defined to be of order unity. However, the training set $\tilde{D}$ is now chosen to consist of $|\tilde{D}| = \alpha N$ randomly drawn questions $\boldsymbol{\xi}^\mu \in \{-1, 1\}^N$. We will show that Q and R obey deterministic macroscopic equations for any α. These equations, however, fail to close as soon as the training set is incomplete (for $\alpha < \infty$). In contrast, our previous results are recovered for the case of complete training sets (for $\alpha \to \infty$). In addition we will derive for the case of complete training sets the macroscopic equations for the batch version of some of the most popular learning rules.

As could have been expected, we will also need the joint input distribution

$$P(x, y) = \langle\langle \delta[x - \hat{\boldsymbol{J}}\cdot\boldsymbol{\xi}]\delta[y - \boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde{D}}\rangle_{Q,R;t} \tag{80}$$

Note that we cannot simply assume the distribution (80) to be of a Gaussian form; it will depend on α. We will now first show that the second-order moments of $P(x, y)$ remain finite for any α in the limit $N \to \infty$. For arbitrary vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ we find

$$\begin{aligned} \langle(\boldsymbol{x}\cdot\boldsymbol{\xi})(\boldsymbol{y}\cdot\boldsymbol{\xi})\rangle_{\tilde{D}} &= \boldsymbol{x}\cdot\boldsymbol{y} + L(\boldsymbol{x}, \boldsymbol{y}) \\ L(\boldsymbol{x}, \boldsymbol{y}) &= \sum_{i\neq j} x_i y_j \left[\frac{1}{\alpha N}\sum_{\mu=1}^{\alpha N} \xi_i^\mu \xi_j^\mu\right] \end{aligned}$$

The second term is bounded according to $|L(\boldsymbol{x}, \boldsymbol{y})| \leq \max_i |\lambda_i||\boldsymbol{x}||\boldsymbol{y}|$, in which the λ_i denote the (real) eigenvalues of the matrix $M_{ij} = [1 - \delta_{ij}/(\alpha N)]\sum_{\mu=1}^{\alpha N} \xi_i^\mu \xi_j^\mu$. For large N the spectrum of the matrix M can be calculated using random matrix theory (see Hertz, 1990; Hertz *et al.* 1989) and the eigenvalues will be bounded:

$$\begin{gathered} \text{eigenvalues of } M_{ij} = \frac{1 - \delta_{ij}}{\alpha N}\sum_{\mu=1}^{\alpha N} \xi_i^\mu \xi_j^\mu : \\ \begin{cases} \alpha \leq 1: & \lambda_{\min} = -1 \qquad\quad \lambda_{\max} = \frac{1}{\alpha} + \frac{2}{\sqrt{\alpha}} \\ \alpha > 1: & \lambda_{\min} = \frac{1}{\alpha} - \frac{2}{\sqrt{\alpha}} \quad \lambda_{\max} = \frac{1}{\alpha} + \frac{2}{\sqrt{\alpha}} \end{cases} \end{gathered} \tag{81}$$

From this it follows that all second-order moments (and therefore also all first-order moments) of the distribution $P(x, y)$ are finite, whatever the value of α, but also that only for $\alpha \to \infty$ we recover the familiar previous expressions (derived for complete training sets in Section 2) for the second-order moments in terms of Q and R:

$$\begin{aligned} \alpha \to \infty: &\int \mathrm{d}x\, \mathrm{d}y\; x^2 P(x, y) = \int \mathrm{d}x\, \mathrm{d}y\; y^2 P(x, y) = 1, \\ &\int \mathrm{d}x\, \mathrm{d}y\; xy\; P(x, y) = R/Q \end{aligned} \tag{82}$$

The next stage is to assess the scaling of the various diffusion terms $G_{\mu\nu}^{\star\star\star}$ in the Fokker-Planck Equation 71. These should vanish for $N \to \infty$ if our observables are to behave deterministically in the limit $N \to \infty$. For the present observables the diffusion terms (73, 75) become

$$\begin{aligned} G_{QQ}^{\mathrm{onl}}[Q, R; t] &= \frac{4\eta^2}{N} Q \int \mathrm{d}x\, \mathrm{d}y\; P(x, y) x^2 \mathscr{F}^2\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right] \\ G_{QR}^{\mathrm{onl}}[Q, R; t] &= \frac{2\eta^2}{N} Q^{\frac{1}{2}} \int \mathrm{d}x\, \mathrm{d}y\; P(x, y) xy \mathscr{F}^2\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right] \\ G_{RR}^{\mathrm{onl}}[Q, R; t] &= \frac{\eta^2}{N} \int \mathrm{d}x\, \mathrm{d}y\; P(x, y) y^2 \mathscr{F}^2\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right] \end{aligned}$$

$$G_{QQ}^{\text{bat}}[Q,R;t] = \frac{4\eta^2}{N} Q \left\{ \int \mathrm{d}x\, \mathrm{d}y\, P(x,y) x\, \mathrm{sgn}(y) \times \mathscr{F}\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right] \right\}^2$$

$$G_{QR}^{\text{bat}}[Q,R;t] = \frac{2\eta^2}{N} Q^{\frac{1}{2}} \left\{ \int \mathrm{d}x\, \mathrm{d}y\, P(x,y) x\, \mathrm{sgn}(y) \times \mathscr{F}\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right] \right\} \times \left\{ \int \mathrm{d}x\, \mathrm{d}y\, P(x,y)|y| \mathscr{F}\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right] \right\}$$

$$G_{RR}^{\text{bat}}[Q,R;t] = \frac{\eta^2}{N} \left\{ \int \mathrm{d}x\, \mathrm{d}y\, P(x,y)|y| \mathscr{F}\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right] \right\}^2$$

We conclude that all diffusion terms $G_{\star\star}^{\star\star\star}$ are of order $\mathcal{O}(1/N)$ provided $\mathscr{F}[\ldots]$ is bounded (which we assumed from the start). This implies that for $N \to \infty$ our macroscopic observables Q and R indeed evolve deterministically for any $\alpha > 0$.

The resulting deterministic equations for the duo (Q, R) for on-line and batch learning are given by combining (77) with the flow terms (72) and (74), respectively. We now work out these equations explicitly, starting with the on-line scenario. Insertion of (72) into (77) gives

$$\frac{\mathrm{d}}{\mathrm{d}t} Q = \lim_{N\to\infty} \left\{ 2\eta Q^{\frac{1}{2}} \int \mathrm{d}x\, \mathrm{d}y P(x,y) \times x\, \mathrm{sgn}(y) \mathscr{F}\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right] + \eta^2 \int \mathrm{d}x\, \mathrm{d}y P(x,y) \mathscr{F}^2\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right] \right\} \tag{83}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} R = \lim_{N\to\infty} \eta \int \mathrm{d}x\, \mathrm{d}y\, P(x,y)|y| \mathscr{F}\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right] \tag{84}$$

Note that these equations are of the same form as those derived earlier for complete training sets, i.e. (9, 10). The differences between complete and incomplete training sets are purely in the joint distribution $P(x,y)$, i.e. Equation 80.

Working out the macroscopic equations for the case of batch learning is somewhat less straightforward, although the final result will be simpler. Insertion of (74) into (77) gives, with the usual short-hand $\mathscr{F}[\ldots] = \mathscr{F}[|\boldsymbol{J}|; \boldsymbol{J}\cdot\boldsymbol{\xi}, \boldsymbol{B}\cdot\boldsymbol{\xi}]$:

$$\frac{\mathrm{d}}{\mathrm{d}t} Q = \lim_{N\to\infty} \left\{ 2\eta Q^{\frac{1}{2}} \int \mathrm{d}x\, \mathrm{d}y\, P(x,y) x\, \mathrm{sgn}(y) \mathscr{F}\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right] + \frac{\eta^2}{N} \left\langle \sum_i \langle \xi_i\, \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}) \mathscr{F}[\ldots] \rangle_{\tilde{D}}^2 \right\rangle_{\boldsymbol{\Omega};t} \right\}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} R = \lim_{N\to\infty} \eta \int \mathrm{d}x\, \mathrm{d}y\, P(x,y)|y| \mathscr{F}\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right]$$

The second term in the temporal derivative of Q can be written as the subshell average of a quantity of the form

$$\frac{1}{\alpha N^2} \sum_i \sum_{\mu\nu=1}^{\alpha N} x_\mu \xi_i^\mu \xi_i^\nu x_\nu = \frac{\boldsymbol{x}^2}{\alpha N} + K(\boldsymbol{x})$$

$$K(\boldsymbol{x}) = \frac{1}{\alpha N} \sum_{\mu\neq\nu=1}^{\alpha N} x_\mu x_\nu \left[\frac{1}{N} \sum_i \xi_i^\mu \xi_i^\nu \right]$$

in which

$$x_\mu = \frac{\eta}{\sqrt{\alpha N}}\, \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu) \mathscr{F}\left[\sqrt{Q}; \boldsymbol{J}\cdot\boldsymbol{\xi}^\mu, \boldsymbol{B}\cdot\boldsymbol{\xi}^\mu\right]$$

Clearly $\boldsymbol{x}^2 = \mathcal{O}(1)$ for $N \to \infty$, and $K(\boldsymbol{x})$ is bounded according to $|K(\boldsymbol{x})| \leq \max_i |\tilde{\lambda}_i| \boldsymbol{x}^2/\alpha N$, in which the $\tilde{\lambda}_i$ denote the (real) eigenvalues of the matrix $\tilde{M}_{\mu\nu} = [(1-\delta_{\mu\nu})/N] \sum_{i=1}^N \xi_i^\mu \xi_i^\nu$. Note that for $N \to \infty$ the eigenvalues of the matrix $\tilde{M}$ are related to those of the matrix M in (81) by simply replacing $\alpha \to 1/\alpha$ (since the relation between the two cases is interchanging αN and N). From this it follows that $\lim_{N\to\infty} K(\boldsymbol{x}) = 0$ and that in the temporal derivative of Q only the first term survives the limit $N \to \infty$. This leaves the final result:

$$\frac{\mathrm{d}}{\mathrm{d}t} Q = \lim_{N\to\infty} 2\eta Q^{\frac{1}{2}} \int \mathrm{d}x\, \mathrm{d}y\, P(x,y) x\, \mathrm{sgn}(y) \mathscr{F}\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right] \tag{85}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} R = \lim_{N\to\infty} \eta \int \mathrm{d}x\, \mathrm{d}y\, P(x,y)|y| \mathscr{F}\left[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y\right] \tag{86}$$

For any value of α, the difference between the macroscopic equations for on-line learning (83, 84) and batch learning (85, 86) (apart from a possible difference in the expressions one might find for the distribution $P(x,y)$) is simply the presence/absence of terms which are quadratic in the learning rate η.

For finite α, the case of incomplete training sets, we observe that the macroscopic equations for the pair (Q, R) (i.e. (83, 84) and (85, 86)) do not close, since the distribution $P(x,y)$ (80) need not be of a Gaussian form, and its moments need not (and almost certainly will not) be expressible in terms of the quantities Q and R.

For $\alpha \to \infty$, the case of complete training sets, we can express the second-order moments of $P(x,y)$ (80) in terms of the observables (Q, R) via (82). Moreover, we can show that the first-order moments of $P(x,y)$ are zero, since for any normalized vector $\boldsymbol{x} \in \Re^N$:

$$\langle \boldsymbol{x}\cdot\boldsymbol{\xi} \rangle_{\tilde{D}}^2 = \left[\sum_{i=1}^N x_i \left(\frac{1}{\alpha N} \sum_{\mu=1}^{\alpha N} \xi_i^\mu \right) \right]^2 \leq \sum_{i=1}^N \left(\frac{1}{\alpha N} \sum_{\mu=1}^{\alpha N} \xi_i^\mu \right)^2 \equiv \gamma(\boldsymbol{\xi})$$

in which $\gamma(\boldsymbol{\xi})$ obeys (with brackets denoting averages over the possible training sets):

$$\langle \gamma^2(\boldsymbol{\xi}) \rangle = \frac{1}{\alpha^4 N^4} \sum_{ij=1}^N \sum_{\mu\nu\rho\lambda=1}^{\alpha N} \left\langle \xi_i^\mu \xi_i^\nu \xi_j^\rho \xi_j^\lambda \right\rangle = \frac{1}{\alpha^2} + \mathcal{O}\left(\frac{1}{N}\right)$$

This shows that $\lim_{\alpha\to\infty}\gamma(\boldsymbol{\xi})=0$ and that the first-order moments of $P(x,y)$ will be zero. What cannot be demonstrated rigorously, however, is that for $\alpha\to\infty$ the distribution $P(x,y)$ is of Gaussian form. This is impossible in principle, even for $\alpha\to\infty$. We could, for instance, choose an initial state $\boldsymbol{J}(0)$ for the student weight vector of the form $J_i(0)\sim \mathrm{e}^{-i}$, in which case the Gaussian assumption would be violated for short times. If we choose our teacher vector of the form $B_i\sim \mathrm{e}^{-i}$ the situation is even worse: now the system will be forced to evolve into a macroscopic state with a non-Gaussian distribution $P(x,y)$. It will be clear that all we can hope for is that for non-pathological initial conditions $\boldsymbol{J}(0)$ and non-pathological teacher vectors $\boldsymbol{B}$ one can derive a dynamic equation for $P(x,y)$ with Gaussian solutions.

4.4. *Complete training sets: batch learning versus on-line learning*

Here we will work out the macroscopic equations for the batch versions of the Hebbian, perceptron and AdaTron learning rules, and compare the results to those of the on-line scenarios. It turns out that for these cases one can solve the macroscopic dynamical laws explicitly. We restrict ourselves to complete training sets. For $\alpha\to\infty$ and $N\to\infty$ the (exact) results of the previous subsection can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}Q = 2\eta Q^{\frac{1}{2}}\int \mathrm{d}x\,\mathrm{d}y\;P(x,y)x\,\mathrm{sgn}(y)\mathscr{F}\left[Q^{\frac{1}{2}};Q^{\frac{1}{2}}x,y\right] + \Delta\eta^2\int \mathrm{d}x\,\mathrm{d}y\;P(x,y)\mathscr{F}^2\left[Q^{\frac{1}{2}};Q^{\frac{1}{2}}x,y\right] \quad (87)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}R = \eta\int \mathrm{d}x\,\mathrm{d}y\;P(x,y)|y|\mathscr{F}\left[Q^{\frac{1}{2}};Q^{\frac{1}{2}}x,y\right] \quad (88)$$

in which $\Delta=1$ for the on-line scenario and $\Delta=0$ for the batch scenario. Of the distribution $P(x,y)$ we know, without additional assumptions:

$$P(x,y)=\lim_{N\to\infty}\left\langle\left\langle \delta[x-\hat{\boldsymbol{J}}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\right\rangle_{\tilde{D}}\right\rangle_{Q,R;t}$$

$$\langle x\rangle=\langle y\rangle=0,\quad \langle x^2\rangle=\langle y^2\rangle=1,\quad \langle xy\rangle=R/Q$$

If we now assume $\boldsymbol{J}(0)$ and $\boldsymbol{B}$ to be such that $P(x,y)$ has a Gaussian shape, the above expressions for the moments immediately dictate that for both scenarios $P(x,y)$ will be identical to (14). We now firstly recover our previous macroscopic equations (9, 10) for the case of on-line learning ($\Delta=1$), and secondly find that the macroscopic equations for the case of batch learning can, for any choice $\mathscr{F}[\ldots]$ of the details of the learning rule, be obtained from the on-line equations by simply removing from the latter all terms which are quadratic in the learning rate η. This also holds if we write the macroscopic equations in terms of the observables (E,J), since the transformation $(Q,R)\to(E,J)$ does not involve the learning rate η.

For the Hebbian rule $\mathscr{F}[|\boldsymbol{J}|;\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]=1$ we obtain the macroscopic equations describing batch learning by elimination of the η^2 terms from the on-line Equations 17 and 18, giving

$$\frac{\mathrm{d}}{\mathrm{d}t}J=\eta\cos(\pi E)\sqrt{\frac{2}{\pi}}\qquad \frac{\mathrm{d}}{\mathrm{d}t}E=-\frac{\eta\sin(\pi E)}{\pi J}\sqrt{\frac{2}{\pi}} \quad (89)$$

We can solve these equations by exploiting the existence of a conserved quantity. If we define $D(J,E)=J\sin(\pi E)$ we find, using (89), that $\frac{\mathrm{d}}{\mathrm{d}t}D=0$, which allows us to express the length $J(t)$ at any time as

$$J=J_0\frac{\sin(\pi E_0)}{\sin(\pi E)}$$

Substitution into the differential equation for the generalization error E then leads to a single non-linear differential equation involving E only:

$$\frac{\mathrm{d}}{\mathrm{d}t}E=-\sqrt{\frac{2}{\pi}}\frac{\eta\sin^2(\pi E)}{\pi J_0\sin(\pi E_0)}$$

This equation is easily solved:

$$t(E)=\frac{1}{\eta}\sqrt{\frac{\pi}{2}}J_0\sin(\pi E_0)\left[\frac{1}{\tan(\pi E)}-\frac{1}{\tan(\pi E_0)}\right] \quad (90)$$

Asymptotically this gives

$$E\sim\frac{J_0\sin(\pi E_0)}{\eta t\sqrt{2\pi}}$$

Asymptotically the gain in using the batch scenario rather than the on-line scenario is having a power law error relaxation of the form t^{-1} rather than $t^{-\frac{1}{2}}$.

For the perceptron rule $\mathscr{F}[|\boldsymbol{J}|;\ \boldsymbol{J}\cdot\boldsymbol{\xi},\ \boldsymbol{B}\cdot\boldsymbol{\xi}]=\theta[-(\boldsymbol{J}\cdot\boldsymbol{\xi})(\boldsymbol{B}\cdot\boldsymbol{\xi})]$ we obtain the macroscopic equations describing batch learning by elimination of the η^2 terms from the on-line Equations 24 and 25, giving

$$\frac{\mathrm{d}}{\mathrm{d}t}J=-\frac{\eta[1-\cos(\pi E)]}{\sqrt{2\pi}}\qquad \frac{\mathrm{d}}{\mathrm{d}t}E=-\frac{\eta\sin(\pi E)}{\pi\sqrt{2\pi}J} \quad (91)$$

Here we find that the quantity $D(J,E)=J[1+\cos(\pi E)]$ is conserved, which leads to

$$J=J_0\frac{1+\cos(\pi E_0)}{1+\cos(\pi E)}$$

Substitution into the differential equation for the generalization error E then again leads to a single non-linear differential equation involving E only:

$$\frac{\mathrm{d}}{\mathrm{d}t}E=-\frac{\eta\sin(\pi E)[1+\cos(\pi E)]}{\pi\sqrt{2\pi}J_0[1+\cos(\pi E_0)]}$$

which can be solved by writing t as an integral over $\frac{\mathrm{d}t}{\mathrm{d}E}$, and by using

$$\int\frac{\mathrm{d}x}{\sin(x)[1+\cos(x)]}=\frac{1}{2}\left\{\log\tan\left(\frac{x}{2}\right)+\frac{1}{1+\cos(x)}\right\}$$

(see Gradshteyn and Ryzhik, 1980). This results in

$$t(E) = \frac{J_0}{\eta}\sqrt{\frac{\pi}{2}}[1+\cos(\pi E_0)] \times \left[\log\tan\left(\frac{\pi E_0}{2}\right) + \frac{1}{1+\cos(\pi E_0)} - \log\tan\left(\frac{\pi E}{2}\right) - \frac{1}{1+\cos(\pi E)}\right] \tag{92}$$

Asymptotically we now find an exponential decay of the generalization error:

$$E \sim \mathrm{e}^{-\sqrt{(\frac{2}{\pi})}\frac{\eta t}{J_0(1+\cos(\pi E_0))}}$$

The gain in using the batch scenario rather than the on-line scenario for the perceptron learning rule is quite significant. The batch scenario gives an exponentially fast decay of the generalization error, compared to a power law relation of the form $t^{-1/3}$ for on-line learning.

Finally we turn to the AdaTron rule $\mathscr{F}[|\boldsymbol{J}|; \boldsymbol{J}\cdot\boldsymbol{\xi}, \boldsymbol{B}\cdot\boldsymbol{\xi}] = |\boldsymbol{J}\cdot\boldsymbol{\xi}|\theta[-(\boldsymbol{J}\cdot\boldsymbol{\xi})(\boldsymbol{B}\cdot\boldsymbol{\xi})]$. Here we obtain the macroscopic equations describing batch learning by elimination of the η^2 terms from the on-line Equations 28 and 29, giving

$$\frac{\mathrm{d}}{\mathrm{d}t}J = -\eta JE + \frac{\eta J}{\pi}\cos(\pi E)\sin(\pi E)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{\eta\sin^2(\pi E)}{\pi^2}$$

The equation for the generalization error is already decoupled from the equation giving the evolution of the length J, and can be solved directly:

$$t(E) = \frac{\pi}{\eta\tan(\pi E)} - \frac{\pi}{\eta\tan(\pi E_0)} \tag{93}$$

Asymptotically this behaves as

$$E \sim \frac{1}{\eta t}$$

For the AdaTron rule there is only little to be gained in switching from on-line learning to batch learning. Both scenarios give a power law error relaxation of the form t^{-1} (albeit with different prefactors).

We summarize the results of this section on batch learning with complete training sets in Table 2, and also illustrate the differences between the batch results and the on-line results in Fig. 14. Whereas the error evolution for the batch versions of the Hebbian and AdaTron rules is almost identical, there is clearly a remarkable difference between the perceptron learning rule on the one hand and the Hebbian and AdaTron rules on the other, in the degree to which they benefit from being executed in a batch scenario rather than an on-line scenario. Only the perceptron rule manages to significantly capitalize on the advantage of batch learning (where all question/answer pairs in the training set $\tilde{D}$ are available at each iteration step, rather than just a single question/answer pair) and realize an exponential decay of the generalization error.

5. Incomplete training sets

5.1. *The problem and our options*

We have seen in the previous section that in the case of incomplete training sets (where $|\tilde{D}| = \alpha N$) the equations for our familiar observables $Q[\boldsymbol{J}] = \boldsymbol{J}^2$ and $R[\boldsymbol{J}] = \boldsymbol{J}\cdot\boldsymbol{B}$ (or, equivalently, for $|\boldsymbol{J}|$ and the generalization error $E_{\mathrm{g}}[\boldsymbol{J}] = 1/\pi \arccos\,(R[\boldsymbol{J}]/\sqrt{Q[\boldsymbol{J}]})$) no longer close, since the distribution $P(x,y)$ (80) will no longer be Gaussian and cannot be written in such a way that its dependence on the weight vector $\boldsymbol{J}$ is only through the observables $Q[\boldsymbol{J}]$ and $R[\boldsymbol{J}]$. One can in fact show that for $\alpha < \infty$ no finite set of observables will ever obey a closed set of dynamic equations. A rigorous proof of this statement (which would be

Table 2. *Overview of exact results on batch learning rules for perceptrons with complete training sets, in the limit $N \to \infty$ (an infinite number of inputs)*

Generalization error in perceptrons with batch learning rules

Rule	Generalization error	Asymptotics
Hebbian	$t = \frac{J_0\sin(\pi E_0)}{\eta}\sqrt{\frac{\pi}{2}}\left[\frac{1}{\tan(\pi E)} - \frac{1}{\tan(\pi E_0)}\right]$	$E \sim \frac{J_0\sin(\pi E_0)}{\eta\sqrt{2\pi}}t^{-1}$
Perceptron	$t = \frac{J_0}{\eta}\sqrt{\frac{\pi}{2}}(1+\cos(\pi E_0))\left[\ln\tan\left(\frac{\pi E_0}{2}\right) + \frac{1}{1+\cos(\pi E_0)} - \ln\tan\left(\frac{\pi E}{2}\right) - \frac{1}{1+\cos(\pi E)}\right]$	$E \sim \mathrm{e}^{-\sqrt{\frac{2}{\pi}}\frac{\eta t}{J_0[1+\cos(\pi E_0)]}}$
AdaTron	$t = \frac{\pi}{\eta}\left[\frac{1}{\tan(\pi E)} - \frac{1}{\tan(\pi E_0)}\right]$	$E \sim \frac{1}{\eta}t^{-1}$

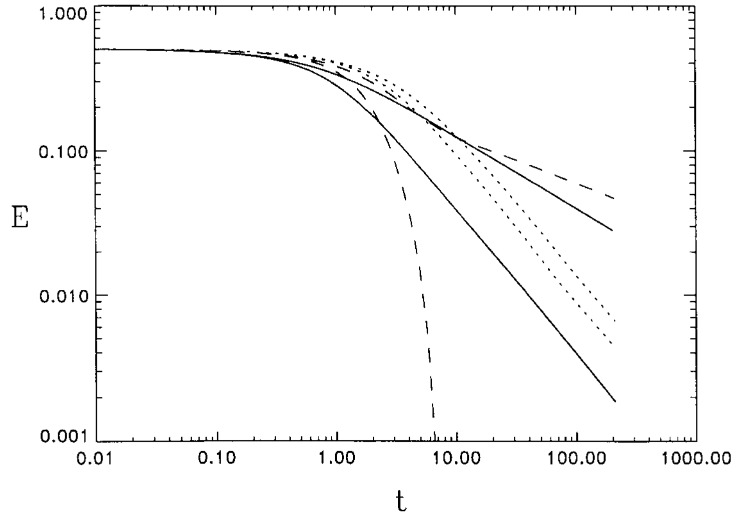

Fig. 14. *Qualitative comparison of the evolution of the error for batch (lower lines) versus on-line (upper lines) learning rules with constant learning rates $\eta = 1$. Solid lines: Hebbian rule; dashed lines: perceptron rule; dotted lines: AdaTron rule*

based on path integral techniques) is beyond the scope of this paper, but one can easily observe the generation of an infinite hierarchy of equations. To determine the temporal derivatives of $Q[\boldsymbol{J}]$ and $R[\boldsymbol{J}]$ one finds that one needs to calculate a second set of observables $U[\boldsymbol{J}]$ and $V[\boldsymbol{J}]$ (which are not expressible in terms of $Q[\mathbf{J}]$ and $R[\boldsymbol{J}]$ for $\alpha < \infty$); calculating the temporal derivatives of $U[\boldsymbol{J}]$ and $V[\boldsymbol{J}]$, in turn, generates new observables $W[\boldsymbol{J}]$ and $Y[\boldsymbol{J}]$, etc.

Closely related to this problem is the fact that our macroscopic equations always involve averages over the training set $\tilde{D}$, which for $\alpha < \infty$ will generally depend on the details of the choice made for the αN questions ξ^μ in $\tilde{D}$. Since we cannot expect to be able to solve the dynamics for any given microscopic realization $\{\xi^1, \ldots, \xi^{\alpha N}\}$ of the set $\tilde{D}$, we will be forced to restrict ourselves to calculating *averages* of observables *over all possible realizations of the training set*. In order to avoid, thereby, ending up with irrelevant statements (since we really aim to arrive at predictions for actual simulation experiments, rather than averages over many such predictions), it is of vital importance to focus on those observables which in the limit $N \to \infty$ tend towards their averages over all possible training sets anyway. Numerical simulations show that macroscopic observables such as the generalization and training errors have this property: if, for example, one chooses the questions ξ^μ in the training set $\tilde{D}$ at random from $D = \{-1, 1\}^N$, one will simply observe that for large N the curves for E_g and E_t as functions of time (such as those shown in Fig. 13) are reproducible, and depend only on the relative size α of $\tilde{D}$, not on its detailed composition $\{\xi^1, \ldots, \xi^{\alpha N}\}$:*

$$\begin{aligned}\lim_{N\to\infty} \langle E_\mathrm{t} \rangle &= \lim_{N\to\infty} \langle \langle E_\mathrm{t} \rangle \rangle_\mathrm{sets} \\ &= \lim_{N\to\infty} \left\langle \int \mathrm{d}\boldsymbol{J} p_t(\boldsymbol{J}|\xi^1, \ldots, \xi^{\alpha N}) \right. \\ &\qquad \left. \times \frac{1}{\alpha N} \sum_{\mu=1}^{\alpha N} \theta[-(\boldsymbol{J}\cdot\xi^\mu)(\boldsymbol{B}\cdot\xi^\mu)] \right\rangle_\mathrm{sets}\end{aligned} \tag{94}$$

$$\begin{aligned}\lim_{N\to\infty} \langle E_\mathrm{g} \rangle &= \lim_{N\to\infty} \langle \langle E_\mathrm{g} \rangle \rangle_\mathrm{sets} \\ &= \lim_{N\to\infty} \left\langle \int \mathrm{d}\boldsymbol{J} p_t(\boldsymbol{J}|\xi^1, \ldots, \xi^{\alpha N}) \right. \\ &\qquad \left. \times \langle \theta[-(\boldsymbol{J}\cdot\xi)(\boldsymbol{B}\cdot\xi)] \rangle_D \right\rangle_\mathrm{sets}\end{aligned} \tag{95}$$

(with the microscopic probability density $p_t(\boldsymbol{J}|\tilde{D})$ for the student weight vector, given a realization of the training set $\tilde{D}$).

In equilibrium calculations the problem is often less severe, since in many cases one at least knows the stationary microscopic probability density $p_\infty(\boldsymbol{J}|\tilde{D})$, so that one can write down the (exact) expressions for the equilibrium expectation values of the training and generalization errors (5) and their averages over the realizations of the training set (94, 95). One can then work out these expressions and obtain transparent results in the $N \to \infty$ limit upon exchanging the order of the various summations and integrations. The remaining problem is of a technical nature. In dynamical studies away from equilibrium, on the other hand, we usually do not have an expression for $p_t(\boldsymbol{J}|\tilde{D})$ at our disposal, and our problem is of a conceptual rather than a technical nature. In order to proceed we need to average over the realizations of the training sets, but we have as yet no object to average.

* This property is called 'self-averaging'.

The toolbox of non-equilibrium statistical mechanics at present offers two (in a way complementary) techniques to deal with this situation, which is a familiar one in the field of disordered magnetic systems, namely the technique of generating functionals (involving path integrals) and dynamical replica theory. Following the generating functional route one performs the average over the realizations of the training sets on an object from which one can derive all relevant observables by differentiation. In the limit $N \to \infty$ this procedure leads to exact equations for two-time correlation and response functions, which, however, are highly complicated and can be solved in practice only near equilibrium. In dynamical replica theory one derives deterministic macroscopic equations for an observable function (equivalent to an infinite number of ordinary scalar observables), which are averaged over the realizations of the training set using the so-called replica method. Here one assumes that the chosen function obeys closed deterministic equations in the $N \to \infty$ limit; the exactness of the resulting theory depends on the degree to which this assumption is correct. Solving the resulting equations numerically is feasible for transients, but as yet too CPU-intensive to allow for solution close to equilibrium.

5.2. *Route 1: generating functionals and path integrals*

This rather elegant approach, which to our knowledge has so far only been applied to learning rules with binary weights, is based on calculating a generating functional $Z[\boldsymbol{\psi}]$ which is an average over all possible 'paths' $\{\boldsymbol{J}(t)\}$ $(t \geq 0)$ of the student's weight vectors through the state space $\Re^N$, given the dynamics (64),

$$Z[\boldsymbol{\psi}] = \left\langle \mathrm{e}^{-i\sum_i \int_0^t \mathrm{d}s \psi_i(s) J_i(s)} \right\rangle \tag{96}$$

in which time is a continuous variable. As with all path integrals, averages such as (96) are understood to be defined in the following way: (i) one discretizes time in the dynamic equation (64); (ii) one calculates the desired average; and subsequently (iii) one takes the continuum limit in the resulting expression. From (96) one can calculate all relevant single and multiple time observables by functional differentiation. Averaging the generating functional over the possible realizations of the training set $\tilde{D}$ gives relations such as

$$\langle J_i(t) \rangle_{\text{sets}} = i \lim_{\boldsymbol{\psi} \to \mathbf{0}} \frac{\delta}{\delta \psi_i(t)} \langle Z[\boldsymbol{\psi}] \rangle_{\text{sets}} \tag{97}$$

$$\langle J_i(t) J_j(t') \rangle_{\text{sets}} = - \lim_{\boldsymbol{\psi} \to \mathbf{0}} \frac{\delta^2}{\delta \psi_i(t) \delta \psi_j(t')} \langle Z[\boldsymbol{\psi}] \rangle_{\text{sets}} \tag{98}$$

etc. Overall constant prefactors in $Z[\boldsymbol{\psi}]$ can always be recovered a posteriori with the identity $Z[\mathbf{0}] = 1$.

The discretized version of our Equation 64 and the corresponding discretized expression for the generating functional (96), with time-steps of duration Δ, would be

$$\begin{aligned} p_{t+\Delta}(\boldsymbol{J}) = \int \mathrm{d}\boldsymbol{J}' \{ \delta[\boldsymbol{J} - \boldsymbol{J}'] \\ + \Delta N [W[\boldsymbol{J}; \boldsymbol{J}'] - \delta[\boldsymbol{J} - \boldsymbol{J}']] \} p_t(\boldsymbol{J}') \\ (0 < \Delta \ll 1) \end{aligned} \tag{99}$$

$$Z[\boldsymbol{\psi}] = \left\langle \mathrm{e}^{-i \sum_i \sum_{t=0}^{L} \Delta \boldsymbol{\psi}_i(\ell . \Delta) J_i(\ell \cdot \Delta)} \right\rangle \tag{100}$$

At the end of our calculation the dependence of any physical observable on Δ, other than via $t = \ell \Delta$, ought to disappear. Note that, although (99) appears to be almost identical to (59) (Equation 59 can be obtained from (99) by choosing $\Delta = N^{-1}$), there is a crucial technical difference. In (99), in contrast to (59), we can control the parameter that converts time into a continuous variable (Δ) independently of the parameter that controls the fluctuations (N). This allows us to take the limit $N \to \infty$ before the limit $\Delta \to 0$. The discretized process (99) gives for the probability density $P(\boldsymbol{J}(t_0), \ldots, \boldsymbol{J}(t_\ell))$ of a temporally discretized path (with $t_n = n\Delta$):

$$\begin{aligned} & P(\boldsymbol{J}(t_0), \ldots, \boldsymbol{J}(t_\ell)) \\ & = \prod_{n=0}^{\ell-1} \{ \delta[\boldsymbol{J}(t_{n+1}) - \boldsymbol{J}(t_n)] + \Delta N [W \boldsymbol{J}(t_{n+1}); \boldsymbol{J}(t_n)] \\ & \quad - \delta[\boldsymbol{J}(t_{n+1}) - \boldsymbol{J}(t_n)]] \} P(\boldsymbol{J}(t_0)) \end{aligned}$$

so that we find for (100) after averaging over all possible training sets $\tilde{D}$:

$$\begin{aligned} \langle Z[\boldsymbol{\psi}] \rangle_{\text{sets}} = \int \cdots \int \prod_{n=0}^{t/\Delta} \left[\mathrm{d}\boldsymbol{J}(t_n)\ \mathrm{e}^{-i \sum_i \Delta \boldsymbol{\psi}_i(t_n) J_i(t_n)} \right] \\ \times P(\boldsymbol{J}(t_0)) \left\langle \prod_{n=0}^{t/\Delta - 1} \{ \delta[\boldsymbol{J}(t_{n+1}) - \boldsymbol{J}(t_n)] \right. \\ + \Delta N [W[\boldsymbol{J}(t_{n+1}); \boldsymbol{J}(t_n)] \\ \left. - \delta[\boldsymbol{J}(t_{n+1}) - \boldsymbol{J}(t_n)]] \} \right\rangle_{\text{sets}} \end{aligned} \tag{101}$$

The problem has hereby again turned into a technical one, albeit of a highly non-trivial nature. The strategy would now be to: (i) insert into (101) the recipe (60) for the learning rule to be studied; (ii) introduce appropriate δ-distributions that will isolate all occurrences of the vectors $\boldsymbol{\xi}^\mu \in D$ in (101) in such a way that the average over all training sets can be performed; (iii) take the limit $N \to \infty$ for finite Δ (this will lead to a saddle-point integral, involving integration variables with two time-arguments); (iv) take the limit $\Delta \to 0$ which restores the original dynamics and converts all integrals into path integrals; and finally (v) solve the saddle-point equations.

The saddle-point equations will describe a non-Markovian stochastic dynamical problem for an effective single weight variable; it will involve a retarded self-interaction and a stochastic noise which is not local in time (i.e. with an auto-correlation function of finite width). This causes these saddle-point equations to be extremely hard to solve, especially in the transient stages of the learning dynamics. Here we will not follow this procedure further, mainly because for the types of rules we have been considering in this review such calculations have not yet been performed (this program has so far only been carried out for learning rules involving binary weight vectors $\boldsymbol{J} \in \{-1, 1\}^N$).

5.3. *Route 2: dynamical replica theory*

The second procedure to deal with incomplete training sets is closer to the methods used so far for dealing with complete training sets than the above formalism, since it involves macroscopic differential equations for single-time observables. The ground work has already been done in Section 4, where we found that for learning rules of the usual type (6), and under certain conditions, the evolution of macroscopic observables $\boldsymbol{\Omega}[\boldsymbol{J}] = (\Omega_1[\boldsymbol{J}], \ldots, \Omega_\ell[\boldsymbol{J}])$ is in the limit $N \to \infty$ described by deterministic laws. With the short hand $\mathscr{F}[\ldots]$ for $\mathscr{F}[|\boldsymbol{J}|; \boldsymbol{J}\cdot\boldsymbol{\xi}, \boldsymbol{B}\cdot\boldsymbol{\xi}]$, and with the definition of subshell averages introduced in Section 4

$$\langle f(\boldsymbol{J})\rangle_{\boldsymbol{\Omega};t} = \frac{\int \mathrm{d}\boldsymbol{J} p_t(\boldsymbol{J}) f(\boldsymbol{J}) \delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}[\boldsymbol{J}]]}{\int \mathrm{d}\boldsymbol{J} p_t(\boldsymbol{J}) \delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}[\boldsymbol{J}]]} \tag{102}$$

these deterministic laws can be written as:

$$\text{On-Line:}\ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Omega} = \eta\left\langle\left\langle \sum_i \xi_i \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\frac{\partial \boldsymbol{\Omega}[\boldsymbol{J}]}{\partial J_i}\right\rangle_{\tilde{D}}\right\rangle_{\boldsymbol{\Omega};t} + \frac{\eta^2}{2N}\left\langle\left\langle \sum_{ij} \xi_i \xi_j \mathscr{F}^2[\ldots]\frac{\partial^2 \boldsymbol{\Omega}[\boldsymbol{J}]}{\partial J_i \partial J_j}\right\rangle_{\tilde{D}}\right\rangle_{\boldsymbol{\Omega};t} \tag{103}$$

$$\text{Batch:}\ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Omega} = \eta\left\langle \sum_i \langle \xi_i\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\rangle_{\tilde{D}} \frac{\partial \boldsymbol{\Omega}[\boldsymbol{J}]}{\partial J_i}\right\rangle_{\boldsymbol{\Omega};t} + \frac{\eta^2}{2N}\left\langle \sum_{ij} \langle \xi_i\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\rangle_{\tilde{D}} \times \langle \xi_j\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\rangle_{\tilde{D}} \frac{\partial^2 \boldsymbol{\Omega}[\boldsymbol{J}]}{\partial J_i \partial J_j}\right\rangle_{\boldsymbol{\Omega};t} \tag{104}$$

Sufficient conditions for (103, 104) to hold for $N \to \infty$ were found to be:

1. All $\Omega_\mu[\boldsymbol{J}]$ are of order unity for $N \to \infty$.
2. All $\Omega_\mu[\boldsymbol{J}]$ are mean-field observables in the sense of (69).
3. $\Omega_1[\boldsymbol{J}] = \boldsymbol{J}^2$.
4. For all $\mu, \nu \leq \ell$: $\lim_{N\to\infty} G_{\mu\nu}[\boldsymbol{\Omega}; t] = 0$.

in which the diffusion coefficients (for on-line and batch learning, respectively) are given by

$$G^{\mathrm{onl}}_{\mu\nu}[\boldsymbol{\Omega}; t] = \frac{\eta^2}{N}\left\langle\left\langle \sum_{ij} \xi_i \xi_j \mathscr{F}^2[\ldots] \times \left[\frac{\partial \Omega_\mu[\boldsymbol{J}]}{\partial J_i}\right]\left[\frac{\partial \Omega_\nu[\boldsymbol{J}]}{\partial J_j}\right]\right\rangle_{\tilde{D}}\right\rangle_{\boldsymbol{\Omega};t}$$

$$G^{\mathrm{bat}}_{\mu\nu}[\boldsymbol{\Omega}; t] = \frac{\eta^2}{N}\left\langle \sum_{ij} \langle \xi_i \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\rangle_{\tilde{D}} \times \langle \xi_j\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\rangle_{\tilde{D}} \times \left[\frac{\partial \Omega_\mu[\boldsymbol{J}]}{\partial J_i}\right]\left[\frac{\partial \Omega_\nu[\boldsymbol{J}]}{\partial J_j}\right]\right\rangle_{\boldsymbol{\Omega};t}$$

The basic idea of the formalism is to note that for those observables $\boldsymbol{\Omega}[\boldsymbol{J}]$ which obey closed deterministic dynamical laws which are self-averaging in the limit $N \to \infty$, we can use (103, 104) to fully determine these laws. If $\boldsymbol{\Omega}$ obeys closed equations we know that, at least for $N \to \infty$, the right-hand sides of (103, 104) by definition *cannot* depend on the distribution of the microscopic probabilities $p_t(\boldsymbol{J})$ within the $\boldsymbol{\Omega}$-subshells of (102). As a consequence we can simplify the evaluation of (103, 104) by making a convenient choice for $p_t(\boldsymbol{J})$: one that describes probability equipartitioning within the $\boldsymbol{\Omega}$-subshells, i.e.

$$\langle f(\boldsymbol{J})\rangle_{\boldsymbol{\Omega};t} \to \langle f(\boldsymbol{J})\rangle_{\boldsymbol{\Omega}} = \frac{\int \mathrm{d}\boldsymbol{J} f(\boldsymbol{J}) \delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}[\boldsymbol{J}]]}{\int \mathrm{d}\boldsymbol{J} \delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}[\boldsymbol{J}]]} \tag{105}$$

Combination of (105) with (103, 104) and usage of the self-averaging property, then leads to the following closed and deterministic laws:

On-Line:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Omega} = \eta \lim_{N\to\infty}\left\langle\left\langle\left\langle \sum_i \xi_i\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\frac{\partial \boldsymbol{\Omega}[\boldsymbol{J}]}{\partial J_i}\right\rangle_{\tilde{D}}\right\rangle_{\boldsymbol{\Omega}}\right\rangle_{\mathrm{sets}} + \eta^2 \lim_{N\to\infty}\left\langle \frac{1}{2N}\left\langle\left\langle \sum_{ij} \xi_i \xi_j \mathscr{F}^2[\ldots]\frac{\partial^2 \boldsymbol{\Omega}[\boldsymbol{J}]}{\partial J_i \partial J_j}\right\rangle_{\tilde{D}}\right\rangle_{\boldsymbol{\Omega}}\right\rangle_{\mathrm{sets}} \tag{106}$$

Batch:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Omega} = \eta \lim_{N\to\infty}\left\langle\left\langle \sum_i \langle \xi_i\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\rangle_{\tilde{D}} \frac{\partial \boldsymbol{\Omega}[\boldsymbol{J}]}{\partial J_i}\right\rangle_{\boldsymbol{\Omega}}\right\rangle_{\mathrm{sets}} + \eta^2 \lim_{N\to\infty}\left\langle \frac{1}{2N}\left\langle \sum_{ij} \langle \xi_i\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\rangle_{\tilde{D}} \times \langle \xi_j\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathscr{F}[\ldots]\rangle_{\tilde{D}} \frac{\partial^2 \boldsymbol{\Omega}[\boldsymbol{J}]}{\partial J_i \partial J_j}\right\rangle_{\boldsymbol{\Omega}}\right\rangle_{\mathrm{sets}} \tag{107}$$

Given the choice for the observables $\boldsymbol{\Omega}[\boldsymbol{J}]$, our problem has now again been converted into a technical one. One performs the average over all training sets using the replica identity

$$\left\langle \frac{\int \mathrm{d}\boldsymbol{J}\, f(\boldsymbol{J}|\tilde{D})W(\boldsymbol{J}|\tilde{D})}{\int \mathrm{d}\boldsymbol{J} W(\boldsymbol{J}|\tilde{D})} \right\rangle_{\text{sets}}$$
$$= \lim_{n\to 0} \left\langle \int \cdots \int \prod_{\alpha=1}^{n} [\mathrm{d}\boldsymbol{J}^{\alpha} W(\boldsymbol{J}^{\alpha}|\tilde{D})] f(\boldsymbol{J}^{1}|\tilde{D}) \right\rangle_{\text{sets}}$$

The key question that remains is how to select the observables $\boldsymbol{\Omega}[\boldsymbol{J}]$, since although the theory is guaranteed to generate the exact dynamic equations for observables which indeed obey closed, deterministic and self-averaging laws, it does not tell us which observables will have these properties beforehand. If the chosen observables $\boldsymbol{\Omega}[\boldsymbol{J}]$ do not obey closed deterministic laws, the method will generate an approximate theory in which one simply has made the closure approximation that all microscopic states $\boldsymbol{J}$ with identical values for the macroscopic observables $\boldsymbol{\Omega}[\boldsymbol{J}]$ are assumed to be equally probable. The available constraints to guide us in finding the appropriate $\boldsymbol{\Omega}[\boldsymbol{J}]$ are the four properties listed below Equation 104 and the knowledge that we will need an infinite number (i.e. $\ell \to \infty$) or, equivalently, an observable function. In addition, for those systems where the equilibrium microscopic probability density $p_{\infty}(\boldsymbol{J}|\tilde{D})$ is known and is of a Boltzmann form, i.e. $p_{\infty}(\boldsymbol{J}|\tilde{D}) \sim \mathrm{e}^{-\beta H(\boldsymbol{J}|\tilde{D})}$, one can guarantee exactness of the theory in equilibrium by choosing one of the observables to be $H(\boldsymbol{J}|\tilde{D})/N$ (or equivalently a set of observables that determine $H(\boldsymbol{J}|\tilde{D})/N$ uniquely), since in that case the equipartitioning assumption (105) is exact in equilibrium.

For the learning dynamics of the type (6), the results of Section 4.3 suggest the following choice:

$$\Omega_1[\boldsymbol{J}] = Q[\boldsymbol{J}] = \boldsymbol{J}^2 \qquad \Omega_2[\boldsymbol{J}] = R[\boldsymbol{J}] = \boldsymbol{J}\cdot\boldsymbol{B}$$
$$\Omega_{xy}[\boldsymbol{J}] = P(x,y;\boldsymbol{J}) = \langle \delta[x - \boldsymbol{J}\cdot\xi]\delta[y - \boldsymbol{B}\cdot\xi]\rangle_{\tilde{D}} \tag{108}$$

(with $x, y \in \Re$). Note that here we have defined the distribution $P(x,y:\boldsymbol{J})$ without explicit normalization of $\boldsymbol{J}$, i.e. with $x = \boldsymbol{J}\cdot\xi$ rather than $x = \hat{\boldsymbol{J}}\cdot\xi$, which will make the subsequent equations somewhat simpler. The procedure for dealing with the distribution $P(x,y;\boldsymbol{J})$ is to first represent it by a finite number of ℓ values $P(x_{\mu}, y_{\mu}; \boldsymbol{J})$ (e.g. as a histogram), and take the limit $\ell \to \infty$ after the limit $N \to \infty$ has been taken. The observables (108) satisfy the four conditions as given below Equation 104 for obeying deterministic laws in the $N \to \infty$ limit if all $B_i = \mathcal{O}(N^{-\frac{1}{2}})$. The conditions that all observables must be of order N^0 and that $\Omega_1[\boldsymbol{J}] = \boldsymbol{J}^2$ hold trivially. The mean-field nature of the observables and the vanishing of the diffusion terms, however, are easily verified for the observables $\boldsymbol{J}^2$ and $\boldsymbol{J}\cdot\boldsymbol{B}$, but involve non-trivial analysis in the case of the distribution $P(x,y;\boldsymbol{J})$ (for details see Coolen and Saad, 1998). Working out the closed Equations 106 and 107 for the observables (108) gives the following result:

$$\frac{\mathrm{d}}{\mathrm{d}t}Q = 2\eta \int \mathrm{d}x\,\mathrm{d}y\; P(x,y)x\;\mathrm{sgn}(y)\mathscr{F}\left[\sqrt{Q};x,y\right]$$
$$+ \Delta\eta^2 \int \mathrm{d}x\,\mathrm{d}y\; P(x,y)\mathscr{F}^2\left[\sqrt{Q};x,y\right] \tag{109}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}R = \eta \int \mathrm{d}x\,\mathrm{d}y\; P(x,y)|y|\mathscr{F}\left[\sqrt{Q};x,y\right] \tag{110}$$

$$\frac{\partial}{\partial t}P(x,y) = -\frac{\eta}{\alpha}\frac{\partial}{\partial x}\left[\mathrm{sgn}(y)\mathscr{F}\left[\sqrt{Q};x,y\right]P(x,y)\right]$$
$$- \left[\eta\frac{\partial}{\partial x}\int \mathrm{d}x'\,\mathrm{d}y'\;\mathrm{sgn}(y')\mathscr{F}\left[\sqrt{Q};x',y'\right]\right.$$
$$\left.\mathscr{A}[x,y;x',y']\right] + \frac{1}{2}\Delta\eta^2\frac{\partial^2}{\partial x^2}$$
$$\times \left[P(x,y)\int \mathrm{d}x'\,\mathrm{d}y'\;P(x',y')\right.$$
$$\left.\times\,\mathscr{F}^2\left[\sqrt{Q};x',y'\right]\right] \tag{111}$$

with $\Delta = 1$ for on-line learning and $\Delta = 0$ for batch learning. Again we observe that the difference between the two modes of learning is reflected only in the presence/absence of the η^2 terms in the dynamic laws. All complications are contained in the function $\mathscr{A}[x,y;x',y']$, which plays the role of a Green's function, and is given by

$$\mathscr{A}[x,y;x',y'] = \lim_{n\to 0}\lim_{N\to\infty}\left\langle \int \prod_{\alpha=1}^{n}\left\{\mathrm{d}\boldsymbol{J}^{\alpha}\delta[Q - Q[\boldsymbol{J}^{\alpha}]]\right.\right.$$
$$\left.\times\,\delta[R - R[\boldsymbol{J}^{\alpha}]]\prod_{\mu}\delta[P(x_{\mu},y_{\mu}) - P(x_{\mu},y_{\mu};\boldsymbol{J}^{\alpha})]\right\}$$
$$\times\,\langle\langle\delta[x - \boldsymbol{J}^{1}\cdot\xi]\delta[y - \boldsymbol{B}\cdot\xi](\xi\cdot\xi')\left[1 - \delta_{\xi\xi'}\right]$$
$$\left.\times\,\delta[x' - \boldsymbol{J}^{1}\cdot\xi']\delta[y' - \boldsymbol{B}\cdot\xi']\rangle_{\Omega}\rangle_{\Omega}\right\rangle_{\text{sets}} \tag{112}$$

After a number of manipulations we can perform the average over the training sets and write (112) ultimately in the form

$$\mathscr{A}[x,y;x',y'] = \int \frac{\mathrm{d}\hat{x}\,\mathrm{d}\hat{x}'\,\mathrm{d}\hat{y}\,\mathrm{d}\hat{y}'}{(2\pi)^4}\mathrm{e}^{i[x\hat{x}+x'\hat{x}'+y\hat{y}+y'\hat{y}']}$$
$$\times \lim_{n\to 0}\lim_{N\to\infty}\int \mathrm{d}\boldsymbol{q}\,\mathrm{d}\hat{\boldsymbol{q}}\,\mathrm{d}\hat{\boldsymbol{Q}}\,\mathrm{d}\hat{\boldsymbol{R}}\prod_{\alpha x'' y''}\mathrm{d}\hat{P}_{\alpha}(x'',y'')$$
$$\times\,\mathrm{e}^{N\Psi[\boldsymbol{q},\hat{\boldsymbol{q}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{R}},\{\hat{P}\}]}\mathscr{L}\left[\hat{x},\hat{y};\hat{x}',\hat{y}';\boldsymbol{q},\hat{\boldsymbol{q}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{R}},\{\hat{P}\}\right]$$

with

$$\Psi[\ldots] = i\sum_{\alpha}\hat{Q}_{\alpha}(1 - q_{\alpha\alpha}) + iR\sum_{\alpha}\hat{R}_{\alpha} + i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}q_{\alpha\beta}$$
$$+ i\sum_{\alpha}\int \mathrm{d}x''\,\mathrm{d}y''\hat{P}_{\alpha}(x'',y'')P(x'',y'')$$
$$+ \alpha\log\mathscr{D}[\boldsymbol{q},\{\hat{P}\}] + \lim_{N\to\infty}\frac{1}{N}\sum_{i}\log$$
$$\int \mathrm{d}\boldsymbol{\sigma}\;\mathrm{e}^{-i\tau_i\sqrt{Q}\sum_{\alpha}\hat{R}_{\alpha}\sigma_{\alpha} - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_{\alpha}\sigma_{\beta}} \tag{113}$$

where $\tau_i = B_i\sqrt{N}$. The functions $\mathscr{L}[\ldots]$ and $\mathscr{D}[\ldots]$ are given by complicated integrals. The term in the expression for $\mathscr{A}[\ldots]$ involving $\lim_{n\to 0}$ and $\lim_{N\to\infty}$ will be given by the intensive part $\mathscr{L}[\ldots]$ evaluated in the dominating saddle-point of Ψ, and finally we get

$$\mathscr{A}[x,y;x',y'] = \int \frac{\mathrm{d}\hat{x}\,\mathrm{d}\hat{x}'\,\mathrm{d}\hat{y}\,\mathrm{d}\hat{y}'}{(2\pi)^4}\,\mathrm{e}^{i[x\hat{x}+x'\hat{x}'+y\hat{y}+y'\hat{y}']} \times \lim_{n\to 0} \mathscr{L}\left[\hat{x},\hat{y};\hat{x}',\hat{y}';\boldsymbol{q},\hat{\boldsymbol{q}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{R}},\{\hat{P}\}\right] \tag{114}$$

in which the order parameters $\{\boldsymbol{q},\hat{\boldsymbol{q}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{R}},\{\hat{P}\}\}$ are calculated by extremization of the function $\Psi[\ldots]$ (113). The meaning of the order parameters $q_{\alpha\beta}$ in the relevant saddle point at any time t is given in terms of the (time-dependent) averaged probability distribution $\langle P_t(q)\rangle_{\text{sets}}$ for the mutual overlap between the weight vectors $\boldsymbol{J}^a$ and $\boldsymbol{J}^b$ of two independently evolving learning processes with the same realization of the training set $\tilde{D}$. One can show (for $N \to \infty$):

$$\begin{aligned}\langle P_t(q)\rangle_{\text{sets}} &= \left\langle\left\langle\left\langle \delta\left[q - \frac{\boldsymbol{J}^a\cdot\boldsymbol{J}^b}{|\boldsymbol{J}^a||\boldsymbol{J}^b|}\right]\right\rangle\right\rangle\right\rangle_{\text{sets}}\\ \langle P_t(q)\rangle_{\text{sets}} &= \lim_{n\to 0}\frac{1}{n(n-1)}\sum_{\alpha\neq\beta}\delta[q - q_{\alpha\beta}]\end{aligned} \tag{115}$$

At this stage one usually makes the so-called replica symmetric (RS) ansatz in the extremization problem, which in view of (115) is equivalent to assuming the absence of complex ergodicity breaking (simple ergodicity breaking, i.e. with only a finite number of ergodic components, is still possible via the existence of multiple solutions for the replica symmetric saddle-point equations).*This replica symmetric ansatz is usually correct in the transient stages of the dynamics. If with a modest amount of foresight we put

$$q_{\alpha\beta} = q_0\delta_{\alpha\beta} + q\left[1-\delta_{\alpha\beta}\right], \quad \hat{q}_{\alpha\beta} = \tfrac{1}{2}i\left[r - r_0\delta_{\alpha\beta}\right],$$
$$\hat{R}_\alpha = i\rho, \quad \hat{Q}_\alpha = i\phi, \quad \hat{P}_\alpha(u,v) = i\chi[u,v]$$

we end up, after a modest amount of algebra and after elimination of most of the scalar order parameters via the saddle-point equations, with an extremization problem for a quantity Ψ_{RS} involving only the function $\{\chi\}$ and the scalar q:

$$\begin{aligned}\Psi_{\text{RS}}[q,\{\chi\}] = \; & \frac{1-\alpha-R^2/Q}{2(1-q)} + \frac{1}{2}(1-\alpha)\log(1-q)\\ & - \int \mathrm{d}x'\,\mathrm{d}y' P(x,y)\chi(x,y)\\ & + \alpha\int Dy\,Dz \log\int \mathrm{d}x\;\mathrm{e}^{-\frac{1}{2}x^2/Q(1-q)+x[Ay+Bz]+\frac{1}{\alpha}\chi(x,y)}\end{aligned} \tag{116}$$

with the short-hands $A = R/(1-q)Q$ and $B = \sqrt{qQ-R^2}/(1-q)Q$ and the short-hand for the Gaussian measure $Dz = (2\pi)^{-1/2}\,\mathrm{e}^{(-1/2)z^2}\,\mathrm{d}z$ (similarly for Dy). This result is surprisingly simple, compared to similar results for other complex systems of this class (such as spin-glasses and attractor neural networks near saturation). Firstly, it involves just a small number of order parameters to be varied (just q and the function χ). Secondly, if one works out the saddle-point equations one recovers from the formalism convenient relations such as $\int \mathrm{d}x\; P(x,y) = (2\pi)^{-1/2}\mathrm{e}^{(-1/2)y^2}$ for all y (this makes sense: the distribution of $y = \boldsymbol{B}\cdot\boldsymbol{\xi}$ is Gaussian since the components B_i are statistically independent of the vectors in the training sets).

The final solution provided by dynamical replica theory thus consists of Equations 109, 110 and 111, which are to be solved numerically, in which at each infinitesimal time-step one has to solve the saddle-point problem for (116). The training and generalization errors are then at any time simply given by:

$$\langle\langle E_t\rangle\rangle_{\text{sets}} = \int \mathrm{d}x\;\mathrm{d}y\,\theta[-xy]P(x,y)$$
$$\langle\langle E_g\rangle\rangle_{\text{sets}} = \frac{1}{\pi}\arccos\left[R/\sqrt{Q}\right]$$

The need for solving a complicated saddle-point problem at each infinitesimal time-step explains why working out the predictions of the theory for very large times requires a prohibitively large amount of CPU time. However, the simple form of the present saddle-point equations hints at the possibility of analytical solution, which would lead to an *explicit* diffusion-type equation for the distribution $P(x,y)$. The experience obtained in using this formalism for other systems with a comparable complex dynamics, such as spin-glasses and recurrent neural networks near saturation, suggests that the equations resulting from this formalism will be either exact or a reliable approximation, especially in the transient stages of the learning process.

6. Bibliographical notes

The application of statistical mechanical tools to learning processes in artificial neural networks was mainly initiated by the hugely influential study by Gardner (1988). It is impossible to list even a fraction of the papers that followed. Those interested in early applications of statistical mechanics to neural network learning can find their way into the literature via the dedicated (memorial) issue of *Journal of Physics A* (1989) (mostly on statics) and the early review paper by Kinzel and Opper (1991). The approaches and styles of many subsequent statistical mechanical studies of learning dynamics were generated by the two influential papers by Krogh and Hertz (1992) and Seung *et al.* (1992). More recent reviews of the general area of the statistical mechanics of learning and generalization

* For a detailed discussion of ergodicity and (different forms of) ergodicity breaking, as well as the relation between replica symmetry breaking and complex ergodicity breaking, we refer to the textbooks by Mézard *et al.* (1987) and Fischer and Hertz (1991).

(including both statics and dynamics) are in Watkin *et al.* (1993) and Opper and Kinzel (1994).

The on-line learning algorithms studied in Section 2 were first introduced/studied by Rosenblatt (1960) (perceptron rule), Vallet (1989) (Hebbian rule) and Anlauf and Biehl (1989) (AdaTron rule), although at the time these algorithms were not yet studied with the methods described here. The convenient expression for the generalization error of binary perceptrons in terms of the inner product of the student and teacher weight vectors $E_{\mathrm{g}} = 1/\pi \arccos(\boldsymbol{J} \cdot \boldsymbol{B}/|\boldsymbol{J}|)$ appeared first in Györgyi (1990), Opper *et al.* (1990) and Kinzel and Rujan (1990). In the latter, one first finds in an embryonic form the set-up of deriving closed macroscopic equations for the observables $\boldsymbol{J}^2$ and $\boldsymbol{J} \cdot \boldsymbol{B}$. Many of the results we described on on-line learning with complete training sets in perceptrons with fixed rules and fixed learning rates can be found in Biehl and Schwarze (1992), Kinouchi and Caticha (1992), Biehl and Riegler (1994) and Sompolinsky *et al.* (1995). The general lower bound on the generalization error that can be achieved for a given number of question/answer pairs, translating into the lower bound $E_{\mathrm{g}} \sim 0.44 \ldots t^{-1}$ for on-line learning rules, was derived in Opper and Haussler (1991). Calculations involving on-line rules with time-dependent learning rates can be found in Barkai *et al.* (1995) and Sompolinsky *et al.* (1995). The systematic optimization of learning rules to achieve the fastest decay of the generalization error in perceptrons was introduced earlier in Kinouchi and Caticha (1992).

In Heskes and Kappen (1991) one first finds the method to derive exact stochastic differential equations describing learning dynamics (an application of Bedeaux *et al.* (1971)), followed by several studies aimed at extracting information from the microscopic dynamics directly.* The differences between batch learning and on-line learning appear so far to have been addressed mainly in equilibrium calculations (Kinouchi and Caticha, 1995; Opper, 1996; Van den Broeck and Reimann, 1996). There is not yet much literature on learning dynamics with incomplete training sets, apart from simple cases and linear models such as Krogh and Hertz (1992). The generating function approach to the learning dynamics for incomplete training sets was elaborated for perceptrons with binary weights in Horner (1992a,b). The version of the dynamical replica theory calculations described in Section 5 was developed in Coolen *et al.* (1996) and Laughton *et al.* (1996), originally aimed at analysing models for spin-glasses and recurrent neural networks near saturation, and is only now being applied to learning dynamics (Coolen and Saad, 1998).

Finally, even within the already confined area of statistical mechanical studies of the dynamics of learning we have concentrated on the simplest models (binary perceptrons) and the simplest types of tasks (those generated by a noise-free and realizable teacher). As a result there are many interesting areas which we had to leave out, such as the dynamics of learning in the presence of noise, for unsupervised learning rules or for non-stationary teachers (Biehl and Schwarze, 1992; Kinouchi and Caticha, 1993). The most important areas we were forced to leave out, however, are the large bodies of work done on different classes of learning rules, e.g. those involving continuous rather than binary neurons or those in the form of microscopic Fokker-Planck equations, as well as (and especially) on various families of multilayer networks (mostly so-called committee machines, in which the weights connecting the hidden layer to the output neuron(s) are fixed). Relevant recent papers in these areas are Biehl and Schwarze (1995), Copelli and Caticha (1995), Saad and Solla (1995a,b), Biehl *et al.* (1996), Kim and Sompolinsky (1996) and Simonetti and Caticha (1996).

The techniques described in this review can be applied with only minor adjustments and extensions to layered networks of graded response neurons, provided the number of neurons in the hidden layer(s) remains finite in the limit $N \to \infty$. On the other hand, as soon as we move to layered networks in which the number of hidden neurons scales proportional to N, we again face the problem of macroscopic dynamic equations which fail to close. Solving this problem, and the one of handling incomplete training sets, are the key objectives of most present-day research efforts in the research area of the statistical mechanics of learning.

Acknowledgement

It is our pleasure to thank Dr D Saad for many interesting discussions on the dynamics of learning.

References

Anlauf, J. K. and Biehl, M. (1989) *Europhysics Letters* **10**, 687.

Barkai, N., Seung, H. S. and Sompolinsky, H. (1995) *Physical Review Letters* **75**, 1415.

Bedeaux, D., Lakatos-Lindenberg, K. and Shuler, K. (1971) *Journal of Mathematical Physics* **12**, 2116.

Biehl, M. and Riegler, P. (1994) *Europhysics Letters* **28**, 525.

Biehl, M. and Schwarze, H. (1992) *Europhysics Letters* **20**, 733.

Biehl, M. and Schwarze, H. (1995) *Journal of Physics A* **28**, 643.

Biehl, M. Riegler, P. and Wöhler, C. (1996) *Journal of Physics A* **29**, 4769.

[*] This line of research, termed 'stochastic approximation theory', is sometimes presented as opposite to the approach based on deriving macroscopic dynamic equations, with only little scientific justification. The two approaches are mutually consistent and complementary; they simply concentrate on different levels of description and are (sometimes) worked out in different limits. In the present paper we used both, and switched from one to another whenever necessary.

Coolen, A. C. C. and Saad, D. (1998) submitted to Physical Review letters.

Coolen, A. C. C., Laughton, S. N. and Sherrington, D. (1996) *Physical Review B* **53**, 8184.

Copelli, M. and Caticha, N. (1995) *Journal of Physics A* **28**, 1615.

Fischer, K. H. and Hertz, J. A. (1991) *Spin Glasses*, Cambridge: Cambridge University Press.

Gardner, E. (1988) *Journal of Physics A* **21**, 257.

Gradshteyn, I. S. and Ryzhik, I. M. (1980) *Table of Integrals, Series, and Products,* San Diego, Academic Press.

Györgyi, G. (1990) *Physical Review Letters* **64**, 2957.

Hertz, J. A. (1990) *Statistical Mechanics of Neural Networks*, Berlin: Springer.

Hertz, J. A., Krogh, A. and Thorgergsson, G. I. (1989) *Journal of Physics A* **22**, 2133.

Heskes, T and Kappen, B. (1991) *Physical Review A* **44**, 2718.

Horner, H. (1992a) *Zeitschrift für physik B* **86**, 291.

Horner, H. (1992b) *Zeitschrift für physik B* **87**, 371.

Journal of Physics A (1989) **22**.

Kim, J. W. and Sompolinsky, H. (1996) *Physical Review Letters* **76**, 3021.

Kinouchi, O. and Caticha, N. (1992) *Journal of Physics A* **25**, 6243.

Kinouchi, O. and Caticha, N. (1993) *Journal of Physics A* **26**, 6161.

Kinouchi, O. and Caticha, N. (1995) *Physical Review E* **52**, 2878.

Kinzel, W. and Opper, M. (1991) in *Physics of Neural Networks* I, Berlin: Springer.

Kinzel, W. and Rujan, P. (1990) *Europhysics Letters* **13**, 473.

Krogh, A. and Hertz, J. A. (1992) *Journal of Physics A* **25**, 1135.

Laughton, S. N., Coolen, A. C. C. and Sherrington, D. (1996) *Journal of Physics A* **29**, 763.

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

Minsky, M. L. and Papert, S. A. (1969) *Perceptrons*, Cambridge, Massachusetts: MIT Press.

Opper, M. (1996) *Physical Review Letters* **77**, 4671.

Opper, M. and Haussler, D. (1991) *Physical Review Letters* **66**, 2677.

Opper, M. and Kinzel, W. (1994) in *Physics of Neural Networks* III, Berlin: Springer.

Opper, M., Kinzel, W., Kleinz, J. and Nehl, R. (1990) *Journal of Physics A* **23**, L581.

Rosenblatt, F. (1960) *Principles of Neurodynamics*, New York: Spartan.

Saad, D. and Solla, S. (1995a) *Physical Review E* **52**, 4225.

Saad, D. and Solla, S. (1995b) *Physical Review Letters* **74**, 4337.

Seung, H. S., Sompolinsky, H. and Tishby, N. (1992) *Physical Review A* **45**, 6056.

Simonetti, R. and Caticha, N. (1996) *Journal of Physics A* **29**, 4859.

Sompolinsky, H. Barkai, N. and Seung, H. S. (1995) in *Neural Networks: The Statistical Mechanics Perspective*, Singapore: World Scientific.

Vallet, F. (1989) *Europhysics Letters* **8**, 747.

Van den Broeck, C. and Reimann, P. (1996) *Physical Review Letters* **76**, 2188.

Watkin, T. L. H., Rau, A. and Biehl, M. (1993) *Review of Modern Physics* **65**, 499.

Appendix: integrals

In this appendix we give brief derivations of those integrals encountered throughout this paper that turn out to be easy, and give the appropriate reference for finding the nasty ones. All involve the following Gaussian distribution:

$$\langle f(x,y)\rangle = \int \mathrm{d}x\,\mathrm{d}y\, f(x,y)P(x,y)$$

$$P(x,y) = \frac{1}{2\pi\sqrt{1-\omega^2}}\mathrm{e}^{-\frac{1}{2}[x^2+y^2-2xy\omega]/(1-\omega^2)}$$

I: $I_1 = \langle |y| \rangle$

$$I_1 = \int \frac{\mathrm{d}y}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}y^2}|y| = \sqrt{\frac{2}{\pi}}$$

II: $I_2 = \langle x\ \mathrm{sgn}(y)\rangle$

$$\begin{aligned} I_2 &= -\int \frac{\mathrm{d}x\,\mathrm{d}y}{2\pi\sqrt{1-\omega^2}}\ \mathrm{sgn}(y)\mathrm{e}^{-\frac{1}{2}[y^2-2\omega xy]/(1-\omega^2)} \\ &\quad \times (1-\omega^2)\frac{\partial}{\partial x}\mathrm{e}^{-\frac{1}{2}x^2/(1-\omega^2)} \\ &= \int \frac{\mathrm{d}x\,\mathrm{d}y}{2\pi\sqrt{1-\omega^2}}\mathrm{e}^{-\frac{1}{2}[x^2+y^2-2\omega xy]/(1-\omega^2)}\ \mathrm{sgn}(y)\omega y \\ &= \omega\langle|y|\rangle = \omega\sqrt{\frac{2}{\pi}} \end{aligned}$$

III: $I_3 = \langle \theta[-xy]\rangle$

$$\begin{aligned} I_3 &= \int_0^\infty\int_0^\infty \frac{\mathrm{d}x\,\mathrm{d}y}{\pi\sqrt{1-\omega^2}} e^{-\frac{1}{2}[x^2+y^2+2\omega xy]/(1-\omega^2)} \\ &= \sqrt{\frac{1-\omega^2}{\pi}}\int_0^\infty\int_0^\infty \mathrm{d}x\,\mathrm{d}y\ e^{-\frac{1}{2}[x^2+y^2+2\omega xy]} \end{aligned}$$

Introduce polar coordinates $(x,y) = r(\cos\phi, \sin\phi)$:

$$\begin{aligned} I_3 &= \sqrt{\frac{1-\omega^2}{\pi}}\int_0^{\pi/2}\mathrm{d}\phi\int_0^\infty \mathrm{d}r\ \mathrm{e}^{-\frac{1}{2}r^2[1+\omega\sin(2\phi)]} \\ &= \sqrt{\frac{1-\omega^2}{2\pi}}\int_0^\pi \frac{\mathrm{d}\phi}{1+\omega\sin(\phi)} \\ &= \frac{1}{\pi}\left[\frac{\pi}{2} - \arctan\left(\frac{\omega}{\sqrt{1-\omega^2}}\right)\right] \end{aligned}$$

(the last integral can be found in Gradshteyn and Ryzhik (1980)). Finally, using $\cos[(\pi/2)-\psi] = \sin\psi$, we find

$$I_3 = \frac{1}{\pi}\arccos(\omega)$$

IV: $I_4 = \langle x\ \mathrm{sgn}(y)\theta[-xy]\rangle$

$$I_4 = -\frac{1-\omega^2}{\pi}\int_0^\infty \mathrm{d}x\, x\, \mathrm{e}^{-\frac{1}{2}x^2}\int_0^\infty \mathrm{d}y\, \mathrm{e}^{-\frac{1}{2}[y+\omega x]^2+\frac{1}{2}\omega^2 x^2}$$

$$= \frac{1}{\pi}\left[\mathrm{e}^{-\frac{1}{2}x^2}\int_{\omega x/\sqrt{1-\omega^2}}^\infty \mathrm{d}y\, \mathrm{e}^{-\frac{1}{2}y^2}\right]_0^\infty$$

$$-\frac{1}{\pi}\int_0^\infty \mathrm{d}x\, \mathrm{e}^{-\frac{1}{2}y^2}\frac{\partial}{\partial x}\int_{\omega x\sqrt{1-\omega^2}}^\infty \mathrm{d}y\, \mathrm{e}^{-\frac{1}{2}y^2}$$

$$= -\frac{1}{\sqrt{2\pi}} + \frac{\omega}{\pi\sqrt{1-\omega^2}}\int_0^\infty \mathrm{d}x\, \mathrm{e}^{-\frac{1}{2}x^2/(1-\omega^2)} = \frac{\omega-1}{\sqrt{2\pi}}$$

V: $I_5 = \langle |y|\theta[-xy]\rangle$

$$I_5 = \int_0^\infty\int_0^\infty \mathrm{d}x\,\mathrm{d}y\, y[P(x,-y)+P(-x,y)] = \frac{1-\omega}{\sqrt{2\pi}}$$

VI: $I_6 = \langle x^2\theta[-xy]\rangle$

$$I_6 = \frac{1}{\pi\sqrt{1-\omega^2}}\int_0^\infty\int_0^\infty \mathrm{d}x\,\mathrm{d}y\, x^2\, \mathrm{e}^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)}$$

$$= \frac{1}{2\pi\sqrt{1-\omega^2}}\int_0^\infty\int_0^\infty \mathrm{d}x\,\mathrm{d}y\,(x^2+y^2)\, \mathrm{e}^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)}$$

We switch to polar coordinates $(x,y) = r(\cos\theta, \sin\theta)$, and subsequently substitute $t = r^2[1+\omega\sin(2\theta)]/[1-\omega^2]$:

$$I_6 = \frac{1}{2\pi\sqrt{1-\omega^2}}\int_0^{\pi/2}\mathrm{d}\theta\int_0^\infty \mathrm{d}r\, r^3 \mathrm{e}^{-\frac{1}{2}[r^2+2\omega r^2\cos\theta\sin\theta]/(1-\omega^2)}$$

$$= \frac{(1-\omega^2)^{3/2}}{4\pi}\int_0^{\pi/2}\frac{\mathrm{d}\theta}{(1+\omega\sin(2\theta))^2}\int_0^\infty \mathrm{d}t\, t\, \mathrm{e}^{-\frac{1}{2}t}$$

$$= \frac{(1-\omega^2)^{3/2}}{2\pi}\int_0^\pi \frac{\mathrm{d}\phi}{(1+\omega\sin\phi)^2}$$

To calculate the latter integral we define

$$\tilde{I}_n = \int_0^\pi \frac{\mathrm{d}\phi}{(1+\omega\sin\phi)^n}$$

These integrals obey

$$\omega\frac{\mathrm{d}}{\mathrm{d}\omega}\tilde{I}_n - n\tilde{I}_{n+1} = -n\tilde{I}_n$$

so

$$\tilde{I}_2 = \tilde{I}_1 + \omega\frac{\mathrm{d}}{\mathrm{d}\omega}\tilde{I}_1 \qquad I_2 = \frac{2}{\sqrt{1-\omega^2}}\arccos(\omega)$$

(where we used the integral already encountered in III). We now find

$$I_6 = \frac{(1-\omega^2)^{3/2}}{2\pi}\tilde{I}_2 = \frac{(1-\omega^2)}{\pi}\arccos(\omega)$$
$$-\frac{\omega\sqrt{1-\omega^2}}{\pi} + \frac{\omega^2}{\pi}\arccos(\omega)$$

VII: $I_7 = \langle |x||y|\theta[-xy]\rangle$

$$I_7 = \int_0^\infty\int_0^\infty \frac{\mathrm{d}x\,\mathrm{d}y}{\pi\sqrt{1-\omega^2}} xy\, \mathrm{e}^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)}$$

We use the relation

$$x\mathrm{e}^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)}$$
$$= -(1-\omega^2)\frac{\partial}{\partial x}\mathrm{e}^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)}$$
$$-\omega y\mathrm{e}^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)}$$

to give us, using VI:

$$I_7 = \frac{\sqrt{1-\omega^2}}{\pi}\int_0^\infty \mathrm{d}y\, y\, \mathrm{e}^{-\frac{1}{2}\frac{y^2}{(1-\omega^2)}} - \omega\langle y^2\theta[-xy]\rangle$$

$$= \frac{(1-\omega^2)^{3/2}}{\pi} - \frac{\omega(1-\omega^2)}{\pi}\arccos(\omega)$$
$$+\frac{\omega^2\sqrt{1-\omega^2}}{\pi} - \frac{\omega^3}{\pi}\arccos(\omega)$$

VIII: $I_8(x) = \int \mathrm{d}y\theta[y]P(x,y)$

$$I_8(x) = \int_0^\infty \frac{\mathrm{d}y}{2\pi\sqrt{1-\omega^2}}\mathrm{e}^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)}$$

$$= \frac{\mathrm{e}^{-\frac{1}{2}x^2}}{2\pi\sqrt{1-\omega^2}}\int_0^\infty \mathrm{d}y\, \mathrm{e}^{-\frac{1}{2}[y-\omega x]^2/(1-\omega^2)}$$

$$= \frac{\mathrm{e}^{-\frac{1}{2}x^2}}{2\sqrt{2\pi}}\left[1+\mathrm{erf}\left[\frac{\omega x}{\sqrt{2}\sqrt{1-\omega^2}}\right]\right]$$

IX: $I_9(x) = \int \mathrm{d}y\theta[y](y-\omega x)P(x,y)$

$$I_9(x) = -\frac{\sqrt{1-\omega^2}}{2\pi}\int_0^\infty \mathrm{d}y\frac{\partial}{\partial y}\mathrm{e}^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)}$$

$$= \frac{\sqrt{1-\omega^2}}{2\pi}\mathrm{e}^{-\frac{1}{2}x^2/(1-\omega^2)}$$