J. Phys. A: Math. Gen. **32** (1999) 5749–5762. Printed in the UK PII: S0305-4470(99)02237-4

A two-step algorithm for learning from unspecific reinforcement

Reimer Kühn† and Ion-Olimpiu Stamatescu‡§

† Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 19, D-69120, Heidelberg, Germany
‡ Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 16, D-69120, Heidelberg, Germany
§ FESt, Schmeilweg 5, D-69118, Heidelberg, Germany

Received 1 March 1999, in final form 9 June 1999

Abstract. We study a simple learning model based on the Hebb rule to cope with 'delayed', unspecific reinforcement. In spite of the unspecific nature of the information-feedback, convergence to asymptotically perfect generalization is observed, with a rate depending, however, in a non-universal way on learning parameters. Asymptotic convergence can be as fast as that of Hebbian learning, but may be slower. Morever, for a certain range of parameter settings, it depends on initial conditions whether the system can reach the regime of asymptotically perfect generalization, or rather approaches a stationary state of poor generalization.

1. Introduction

Introducing biologically motivated features in models for learning usually has a double role: testing hypotheses for natural learning and finding hints for artificial learning. These problems can be stated at various sophistication levels. Here we do not take the more ambitious point of view of describing the complexity of the former or of finding optimal algorithms for the latter. On the contrary, our motivation is to investigate which are the capabilities of very elementary mechanisms.

One urgent problem with which a system, either natural or artificial, may be confronted when trying to improve its performance is to learn only from the final success/failure of a *series* of consecutive decisions. The typical situation we may consider is that of an 'agent' which let free in a complicated 'landscape' tries many 'paths' to reach a 'goal' and has to optimize its path (a local problem) knowing only the 'time' (or cost) it needs to reach the goal (global information). Here 'goal' may be a survival interest or the solution of a problem, 'path' a series of moves or of partial solution steps in a complex geographical or mathematical 'landscape' etc. The problem we want to approach here is to find out whether there are elementary features characterizing learning under such *unspecific* reinforcement conditions. From the point of view of reinforcement learning our problem may be seen under the 'class III' problems in the classification of Hertz *et al* [1]. However, we stress that our attitude is not that of finding good algorithms for tackling special problems, like movement, control or games—see, e.g., [2]. For this reason we do not consider evolved algorithms from the class of Q-learning [3], of learning using temporal differences [4], agent and critic [5] etc, but we restrict ourselves to the most primitive algorithms which we may think of having a chance to have developed under natural

0305-4470/99/315749+14$30.00

conditions. On the other hand, if such algorithms prove capable of tackling the problem they may well give further insights†.

In the case of neural network systems the usual situation lacks detailed control over the synapses and learning is achieved by confronting the 'pupil' system with the correct answer after each presentation of a pattern. For perceptrons both the unsupervised Hebb rule and the supervised perceptron algorithm are known to lead to an asymptotically perfect generalization, although with different asymptotic laws. In our problem setting, however, the pupil never knows the right answer to each question, but only the average error it makes over many tests. In previous work concerned with this problem [7] (see also [8]) we presented an analysis of a two-step algorithm based on the Hebb rule for perceptrons and used computer simulations and a rough approximation to estimate the convergence conditions. In the present work we undertake a detailed study of this learning algorithm which we call for simplicity 'association-reinforcement(AR)-Hebb rule'. This algorithm introduces two learning parameters and we find that its generalization behaviour is highly non-trivial: in the pre-asymptotic region and depending on the network parameters *fixed points* of the learning dynamics may appear. This leads either to asymptotically perfect generalization with non-universal power laws depending on the (ratio of the) learning parameters, or to stationary states of very poor generalization, according to the network parameters and initial conditions.

That this AR-Hebb algorithm may be of a more general interest is suggested by applying it to a concrete problem of optimizing paths in a landscape with obstacles and traps, in a neural network recasting of [6]; this study will be presented elsewhere (partial results have been given in [7]).

In the next section we shall introduce the problem and the algorithm, and in section 3 we shall present results from numerical simulations. In section 4 we shall study a coarse-grained approximation which is appropriate for large networks ('thermodynamic limit'). Section 5 is reserved for conclusions.

2. Learning rule for perceptrons under unspecific reinforcement

We consider perceptrons with Ising units s, $s_i = \pm 1$ and real weights (synapses) J_i:

$$s = \operatorname{sign}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} J_i s_i\right) = \operatorname{sign}\left(\frac{1}{\sqrt{N}}\boldsymbol{J}\cdot\boldsymbol{s}\right). \tag{1}$$

Here N is the number of input nodes, and we put no explicit thresholds. The network (pupil) is presented with a series of patterns $\xi_i^{(q,l)}$, $q \in \mathbb{N}$, $l = 1, \ldots, L$ to which it answers with $s^{(q,l)}$. A training period consists of the successive presentation of L patterns. The answers are compared with the corresponding answers $t^{(q,l)}$ of a teacher with pre-given weights B_i and the average error made by the pupil over one training period is calculated:

$$e_q = \frac{1}{2L}\sum_{l=1}^{L} |t^{(q,l)} - s^{(q,l)}|. \tag{2}$$

The training algorithm consists of two parts:

(a) A 'blind' Hebb-type *association* at each presentation of a pattern:

$$\boldsymbol{J}^{(q,l+1)} = \boldsymbol{J}^{(q,l)} + \frac{a_1}{\sqrt{N}} s^{(q,l)} \boldsymbol{\xi}^{(q,l)} \tag{3}$$

† An illustration of the problem was provided in an early paper [6] dealing with these questions in the simulation of a device moving on a board.

(b) An 'unspecific' but graded *reinforcement* proportional to the average error e_q introduced in (2), also Hebbian, at the end of each training period,

$$\boldsymbol{J}^{(q+1,1)} = \boldsymbol{J}^{(q,L+1)} - \frac{a_2}{\sqrt{N}} e_q \sum_{l=1}^{L} s^{(q,l)} \boldsymbol{\xi}^{(q,l)}. \tag{4}$$

Because of these two steps we call this algorithm 'AR-Hebb rule' (or 'two-Hebb rule' [7]). We are interested in the behaviour with the number of iterations q of the generalization error $\epsilon_g(q)$:

$$\epsilon_g(q) = \frac{1}{\pi} \arccos\left(\frac{\boldsymbol{J}\cdot\boldsymbol{B}}{|\boldsymbol{J}|\,|\boldsymbol{B}|}\right). \tag{5}$$

The training patterns $\boldsymbol{\xi}^{(q,l)}$ are generated randomly, and are taken to be unbiased in the present paper. The case of structured patterns is more complicated, and will be dealt with in a separate publication [9]. We shall test whether the behaviour of $\epsilon_g(q)$ follows a power law at large q:

$$\epsilon_g(q) \simeq \text{const}\; q^{-p}. \tag{6}$$

Notice the following features:

(a) During training the pupil only uses its own associations $\boldsymbol{\xi}^{(q,l)} \leftrightarrow s^{(q,l)}$ and the average error e_q which does not refer specifically to the particular steps l.
(b) Since the answers $s^{(q,l)}$ are made on the basis of the instantaneous weight values $\boldsymbol{J}^{(q,l)}$ which change at each step according to equation (3), the series of answers form a correlated sequence with each step depending on the previous one. Therefore, e_q in fact measures the performance of a 'path', an interdependent set of decisions.
(c) For $L = 1$ the algorithm reproduces the usual 'perceptron rule' (for $a_1 = 0$) or to the usual 'unsupervised Hebb rule' (for $a_2 = 2a_1$) for on-line learning, for which the corresponding asymptotic behaviour is known [10, 11].

3. Numerical results

In a preliminary analysis [7] we have tested various combinations of $L = 1, 5, 10, 15$ and $N = 50, 100, 200, 300$. We went with q up to 4×10^5. We found the convergence of the learning procedure to depend on the ratio a_1/a_2, in particular no convergence was found for L of 5 and higher if this ratio was decreased significantly below 0.2. For fixed a_1, a_2 the asymptotic behaviour with q appeared well reproduced by a power law and *the exponent was found to depend on* L. For $L = 1$ varying a_1/a_2 between 0 and $\frac{1}{2}$ interpolates between perceptron and Hebbian learning, for ratios larger than 1 new asymptotic behaviour can be expected to show up (see section 4)—we did not perform a systematic numerical analysis for $L = 1$, however.

In the present, more precise analysis we use $L = 5, 10$ and $N = 100, 300$, rising to 8×10^5 iterations. We introduce:

$$\alpha = qL/N. \tag{7}$$

We present here results for the following choices of parameters:

$$a_2 = 0.012 \tag{8}$$

$$(a) \quad a_1 = a_2/20 \tag{9}$$

$$(b) \quad a_1 = a_2/5 \tag{10}$$

$$(c) \quad \begin{aligned} &a_1 = a_2/5 \qquad \text{for} \quad \alpha < 100L \\ &a_1 = a_2/(2L) \qquad \text{for} \quad \alpha \geqslant 100L \end{aligned} \tag{11}$$

$$\begin{array}{ll} (d) & a_1 = a_2/5 \qquad \text{for} \quad \alpha < 100L \\ & a_1 = 0 \qquad\ \ \text{for} \quad \alpha \geqslant 100L. \end{array} \tag{12}$$

We use random initial conditions with the same normalization for the teacher and pupil weights, $\boldsymbol{B}^2/N = \boldsymbol{J}^2/N = 1$. The results are shown in figure 1. In agreement with the preliminary results of [7] we find no convergence in case (a) and convergence in case (b). If a certain threshold in ϵ_g is achieved, switching to a smaller ratio a_1/a_2 is seen to accelerate the asymptotic convergence—case (c)—but even then a_1 cannot be set to zero—case (d). Similar behaviour is observed for other N and $L \geqslant 5$.

This intriguing behaviour provoked us to try to obtain analytic understanding by using the coarse-grained analysis discussed in the next section.

4. Coarse-grained analysis

We combine *blind association* (3) during a learning period of L elementary steps and the graded *unspecific reinforcement* (4) at the end of each learning period into one coarse-grained step

$$\boldsymbol{J}^{(q+1,1)} = \boldsymbol{J}^{(q,1)} + \frac{1}{\sqrt{N}}(a_1 - a_2 e_q) \sum_{l=1}^{L} \operatorname{sign}(\boldsymbol{J}^{(q,l)} \cdot \boldsymbol{\xi}^{(q,l)}) \boldsymbol{\xi}^{(q,l)} \tag{13}$$

$$e_q = \frac{1}{2L} \sum_{l=1}^{L} |\operatorname{sign}(\boldsymbol{B} \cdot \boldsymbol{\xi}^{(q,l)}) - \operatorname{sign}(\boldsymbol{J}^{(q,l)} \cdot \boldsymbol{\xi}^{(q,l)})|. \tag{14}$$

We introduce the notations

$$\hat{R}(q,l) = \frac{1}{N} \boldsymbol{B} \cdot \boldsymbol{J}^{(q,l)} \qquad \hat{Q}(q,l) = \frac{1}{N} [\boldsymbol{J}^{(q,l)}]^2 \tag{15}$$

and we normalize the teacher weights to 1, i.e. $\boldsymbol{B}^2/N = 1$. In the 'thermodynamic limit' $L/N \to 0$ one can treat α as a continuous variable. We shall follow standard procedures [1, 11–13] and obtain the following expressions for the changes of $\hat{R}$ and $\hat{Q}$ over a coarse-grained step:

$$L \frac{\mathrm{d}}{\mathrm{d}\alpha} \hat{R} = \frac{1}{\sqrt{N}}(a_1 - a_2 e_q) \sum_{l=1}^{L} \operatorname{sign}(\boldsymbol{J}^{(q,l)} \cdot \boldsymbol{\xi}^{(q,l)})(\boldsymbol{B} \cdot \boldsymbol{\xi}^{(q,l)}) \tag{16}$$

$$\begin{aligned} L \frac{\mathrm{d}}{\mathrm{d}\alpha} \hat{Q} = {} & \frac{2}{\sqrt{N}}(a_1 - a_2 e_q) \sum_{l=1}^{L} \operatorname{sign}(\boldsymbol{J}^{(q,l)} \cdot \boldsymbol{\xi}^{(q,l)})(\boldsymbol{J}^{(q,l)} \cdot \boldsymbol{\xi}^{(q,l)}) \\ & + \frac{1}{N}(a_1 - a_2 e_q)^2 \left(\sum_{l=1}^{L} \operatorname{sign}(\boldsymbol{J}^{(q,l)} \cdot \boldsymbol{\xi}^{(q,l)}) \boldsymbol{\xi}^{(q,l)} \right)^2. \end{aligned} \tag{17}$$

In the following we shall consider unbiased random input-patterns with

$$\langle \xi_i^{(l,q)} \xi_j^{(k,r)} \rangle = \delta_{ij} \delta_{lk} \delta_{qr}. \tag{18}$$

The local fields:

$$h_J^{(q,l)} = \frac{1}{\sqrt{N}} \boldsymbol{J}^{(q,l)} \cdot \boldsymbol{\xi}^{(q,l)} \qquad h_B^{(q,l)} = \frac{1}{\sqrt{N}} \boldsymbol{B} \cdot \boldsymbol{\xi}^{(q,l)} \tag{19}$$

are then normally distributed with second moments

$$\langle (h_J^{(q,l)})^2 \rangle = \langle \hat{Q}(q,l) \rangle = Q \qquad \langle (h_B^{(q,l)})^2 \rangle = 1 \qquad \langle (h_J^{(q,l)} h_B^{(q,l)}) \rangle = \langle \hat{R}(q,l) \rangle = R. \tag{20}$$

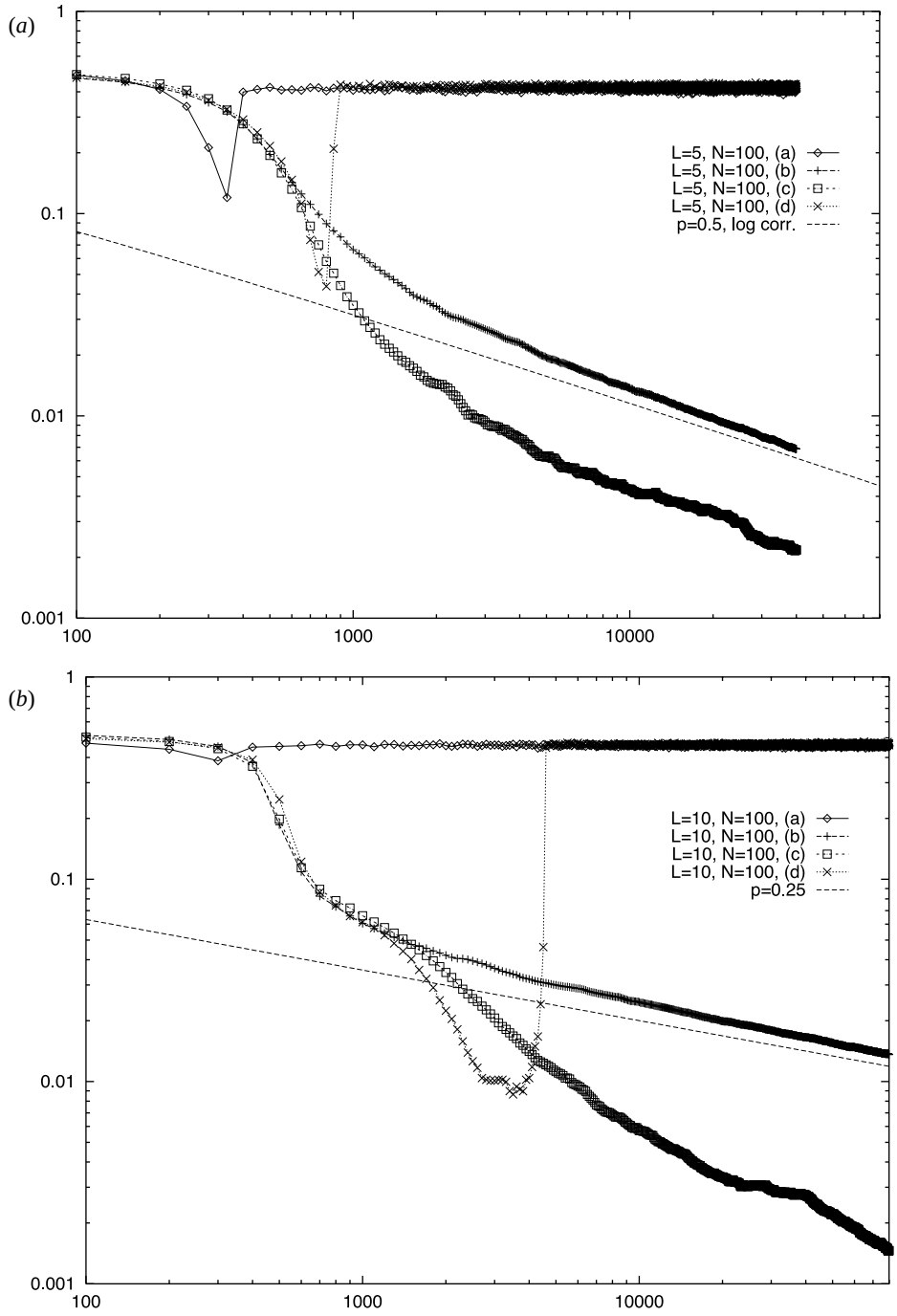
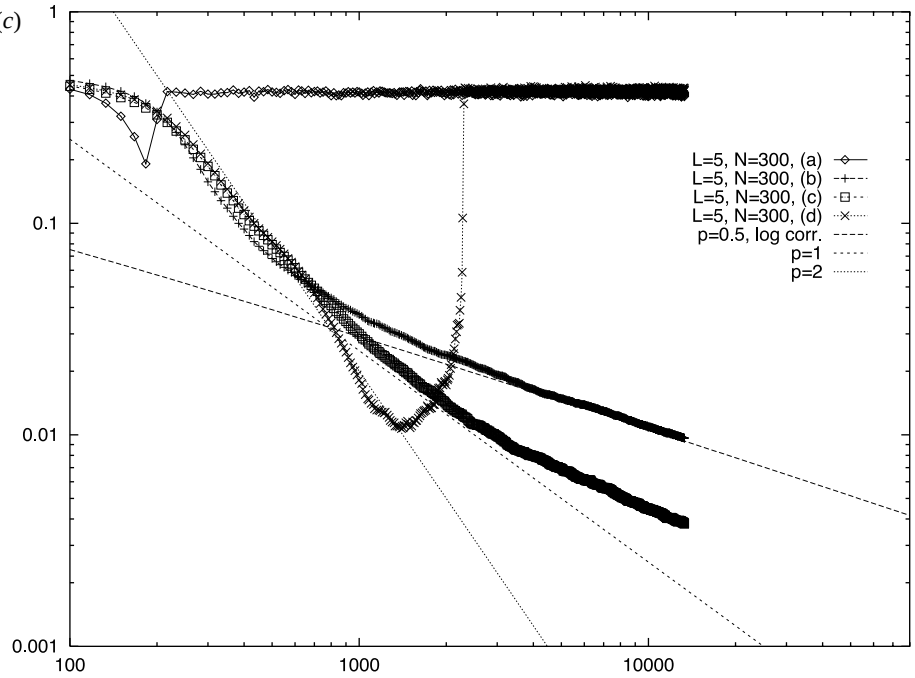


Figure 1. Generalization error ϵ_g versus α for $N = 100$, $L = 5$ and $L = 10$ (*a*), (*b*) and $N = 300$, $L = 5$ and $L = 10$ (*c*), (*d*), for the algorithms (*a*)–(*d*) of equations (9)–(12). The lines indicate the expected asymptotic behaviour as suggested by the coarse-grained approximation discussed in section 4 for the corresponding a_1/a_2 ratio, as well as two further power laws for illustration.

(*c*)



(*d*)

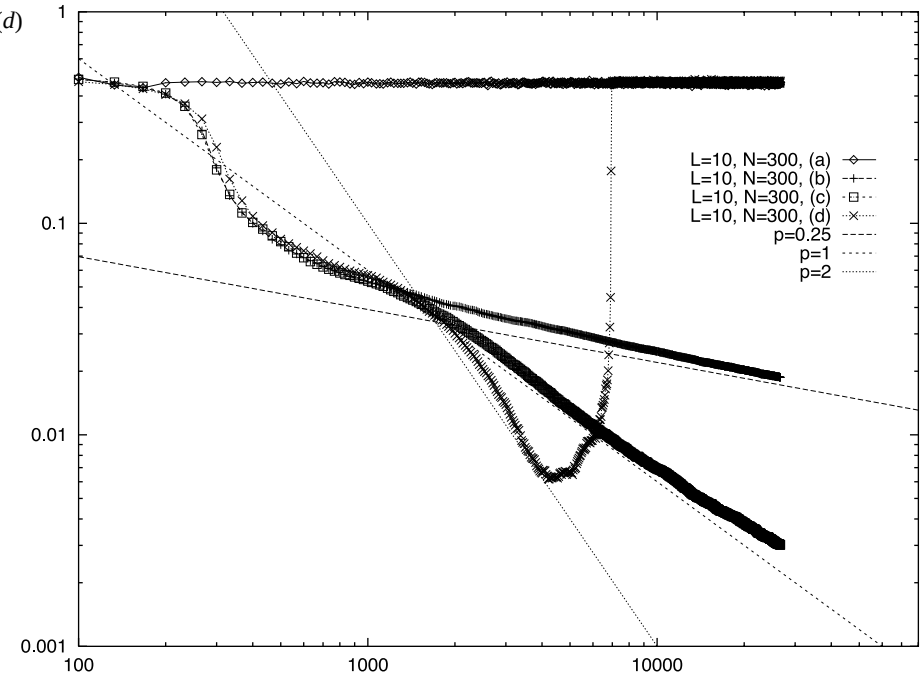


Figure 1. (Continued.)

Their joint probability density is thus given by

$$p(h_J, h_B) = \frac{1}{2\pi\sqrt{\Delta}} \exp\left(-\frac{1}{2\Delta}(Qh_B^2 - 2Rh_Jh_B + h_J^2)\right) \tag{21}$$

with

$$\Delta = Q - R^2. \tag{22}$$

In the thermodynamic limit $N \to \infty$, the self-overlap of the learner $\hat{Q}$ and its overlap $\hat{R}$ with the teacher are self-averaging, so that their evolution equations (16), (17) can be directly rewritten in terms of evolution equations for their averages. Moreover, these averages Q and R become smooth functions on the α-scale, so that we can neglect the dependence of R and Q on l in (21) when used to perform averages on the right-hand sides of (16) and (17), as it would only produce $\mathcal{O}(1/N)$ corrections to the evolution equations, which become negligible as $N \to \infty$. One thus obtains

$$\frac{\mathrm{d}R}{\mathrm{d}\alpha} = \sqrt{\frac{2}{\pi}}\left[a_1\frac{R}{\sqrt{Q}} - \frac{a_2}{2}\left(\frac{R}{\sqrt{Q}} - \frac{1}{L} - \left(1 - \frac{1}{L}\right)P\frac{R}{\sqrt{Q}}\right)\right] \tag{23}$$

$$\frac{\mathrm{d}Q}{\mathrm{d}\alpha} = 2\sqrt{\frac{2}{\pi}}\left[a_1\sqrt{Q} - \frac{a_2}{2}\left(\sqrt{Q} - \frac{R}{L} - \left(1 - \frac{1}{L}\right)P\sqrt{Q}\right)\right] + \left[a_1^2 - a_1a_2(1-P) + \frac{a_2^2}{4}\left(1 - 2P + \frac{1}{L} + \left(1 - \frac{1}{L}\right)P^2\right)\right] \tag{24}$$

where

$$P = -\frac{1}{\pi\sqrt{\Delta}}\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{sign}(x)\int_{Rx}^{\infty}\mathrm{d}y\,\mathrm{e}^{-\frac{1}{2\Delta}y^2} = 1 - \frac{2}{\pi}\arccos\left(\frac{R}{\sqrt{Q}}\right). \tag{25}$$

The generalization error is

$$\epsilon_g = \frac{1}{\pi}\arccos\left(\frac{R}{\sqrt{Q}}\right). \tag{26}$$

We may formally eliminate one of the learning parameters by rescaling our quantities by the parameter a_2:

$$R = \mathcal{R}a_2 \qquad Q = \mathcal{Q}a_2^2 \qquad \lambda = \frac{a_1}{a_2}. \tag{27}$$

We then obtain

$$\frac{\mathrm{d}\epsilon_g}{\mathrm{d}\alpha} = -\frac{1}{\sqrt{2\pi}\pi L\sqrt{\mathcal{Q}}}\sin(\pi\epsilon_g) + \frac{1}{2\pi\mathcal{Q}}\mathrm{cotg}\,(\pi\epsilon_g)\left(\lambda^2 - \left(2\lambda - \frac{1}{L}\right)\epsilon_g + \left(1 - \frac{1}{L}\right)\epsilon_g^2\right) \tag{28}$$

$$\frac{\mathrm{d}\sqrt{\mathcal{Q}}}{\mathrm{d}\alpha} = \sqrt{\frac{2}{\pi}}\left(\lambda - \left(1 - \frac{1}{L}\right)\epsilon_g - \frac{1}{2L}(1 - \cos(\pi\epsilon_g))\right) + \frac{1}{2\sqrt{\mathcal{Q}}}\left(\lambda^2 - \left(2\lambda - \frac{1}{L}\right)\epsilon_g + \left(1 - \frac{1}{L}\right)\epsilon_g^2\right). \tag{29}$$

To establish the asymptotic behaviour we look for solutions of equations (28), (29) in the limit of small ϵ_g, large $\mathcal{Q}$. To leading order (for $\lambda > 0$), these equations become

$$\frac{\mathrm{d}\epsilon_g}{\mathrm{d}\alpha} \simeq -\frac{\epsilon_g}{\sqrt{2\pi}L\sqrt{\mathcal{Q}}} + \frac{\lambda^2}{2\pi^2\mathcal{Q}\epsilon_g} \tag{30}$$

$$\frac{\mathrm{d}\sqrt{\mathcal{Q}}}{\mathrm{d}\alpha} \simeq \sqrt{\frac{2}{\pi}}\lambda \tag{31}$$

which can be solved exactly to give

$$\epsilon_g^2 \simeq \frac{\lambda}{\sqrt{2\pi}\,\pi(\frac{1}{\lambda L}-1)}\mathcal{Q}^{-1/2} + c_1\mathcal{Q}^{-\frac{1}{2\lambda L}} \qquad \text{for} \quad \lambda \neq \frac{1}{L} \tag{32}$$

$$\epsilon_g^2 \simeq \left(\frac{1}{\pi\sqrt{2\pi}L}\ln\mathcal{Q}^{1/2} + c_2\right)\mathcal{Q}^{-1/2} \qquad \text{for} \quad \lambda = \frac{1}{L} \tag{33}$$

i.e. explicitly

$$\epsilon_g^2 \simeq \frac{1}{2\pi(\frac{1}{\lambda L}-1)}\alpha^{-1} + \tilde{c}_1\alpha^{-\frac{1}{\lambda L}} \qquad \text{for} \quad \lambda \neq \frac{1}{L} \tag{34}$$

$$\epsilon_g^2 \simeq \left(\frac{1}{2\pi}\ln\alpha + \tilde{c}_2\right)\alpha^{-1} \qquad \text{for} \quad \lambda = \frac{1}{L} \tag{35}$$

$$\mathcal{Q} \simeq \frac{2}{\pi}\lambda^2\alpha^2 \tag{36}$$

asymptotically at large α.

We see that for $\lambda < \frac{1}{L}$ we obtain asymptotically perfect generalization, the dominant term exhibiting the usual power $-\frac{1}{2}$ (and, for $L = 1$, $\lambda = 0.5$, also the usual coefficient [11]), while for $\lambda > \frac{1}{L}$ the second term in (32), (34) dominates and again ensures perfect generalization but with a different power law, $-1/(2\lambda L)$. For $\lambda = \frac{1}{L}$ we obtain logarithmic corrections—see equations (33), (35). Notice that these results also hold for $L = 1$.

In the case $\lambda = 0$ one can see from (28), (29) that starting with any finite $\mathcal{Q}$ one cannot have perfect generalization for $L > 1$. For $L = 1$ one re-obtains the asymptotic behaviour found in [10].

There is, however, a non-trivial pre-asymptotic region, which turns out to be dominated by two stationarity conditions, one for the self-overlap, $\mathrm{d}\mathcal{Q}/\mathrm{d}\alpha = 0$, and one for the overlap with the teacher-configuration, $\mathrm{d}\mathcal{R}/\mathrm{d}\alpha = 0$ or, alternatively, that for the generalization error $\mathrm{d}\epsilon_g/\mathrm{d}\alpha = 0$. For suitable values of the network parameters, the two stationarity conditions may simultaneously be satisfied, leading to fixed points of the learning dynamics, one of these fully stable and with poor generalization, the other partially stable.

To this pre-asymptotic region we shall now turn our attention and thereby also obtain further specifications for the parameters. In figure 2 we show the evolution of ϵ_g and $\mathcal{Q}$ according to equations (28), (29), starting from $\epsilon_g(0) = 0.5$ and various $\mathcal{Q}(0) = \mathcal{Q}_0$†. The various trajectories are parametrized by λ. In all cases there is a critical value $\lambda_c(\mathcal{Q}_0)$ which separates flows toward a stationary state of poor generalization from flows toward perfect asymptotic generalization. The fixed point in the $\mathcal{Q}, \epsilon_g$ plane (with a location parametrized by λ) which is responsible for this behaviour has an attractive and a repulsive direction. For a given initial condition $\mathcal{Q}_0$, the critical value $\lambda_c(\mathcal{Q}_0)$ is defined as that value for which the attractive manifold connects the initial condition to the partially stable fixed point; for smaller values of λ the flow always is from the initial condition to the fully stable fixed point with poor generalization, for slightly larger values of λ the flow is towards asymptotically perfect generalization. At still larger values of λ the two fixed points eventually coalesce and disappear altogether. Then we always have asymptotically perfect generalization. Some values for $\lambda_c(\mathcal{Q}_0)$ are given in table 1.

In figure 3 we describe the flow in this plane for a given λ, this should be compared with the α-trajectories in the $\mathcal{Q}$, ϵ_g plane for various λ with different starting points $\mathcal{Q}_0$, figure 2.

† Notice that due to (27) the dependence on the initial conditions $\mathcal{Q}_0$ may be translated into a dependence on the learning rate for the initial network: for a fixed ratio λ of learning rates, and given values of the original overlaps Q and R, finer updating (smaller a_1 and a_2) is equivalent to larger values of rescaled overlaps, hence a larger value of $\mathcal{Q}_0$.

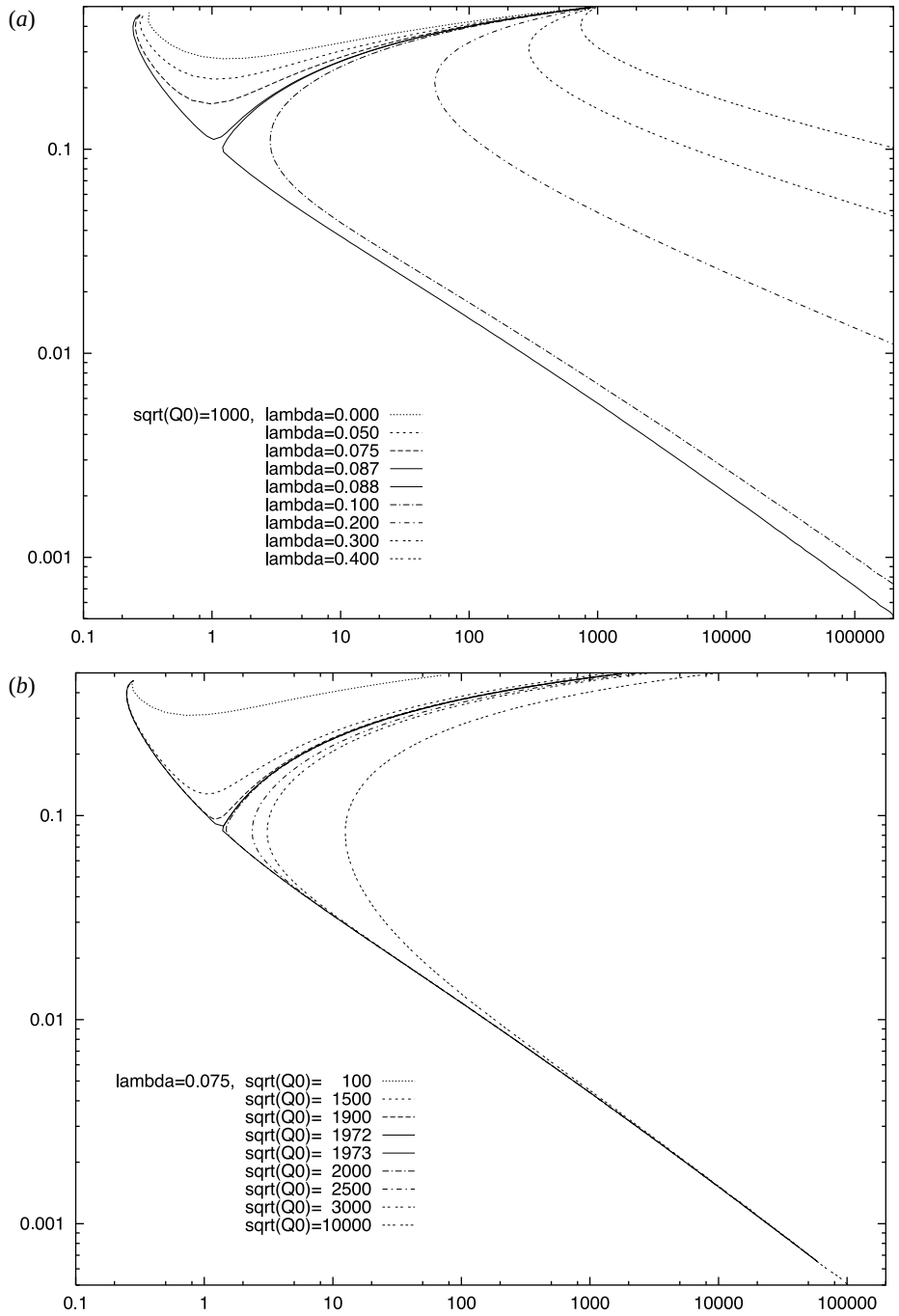


Figure 2. Evolution of the generalization error ϵ_g (vertical axis) and of $\sqrt{\mathcal{Q}}$ (horizontal axis) at $L = 10$ for various λ (*a*) and starting points $\mathcal{Q}(0)$ (*b*).

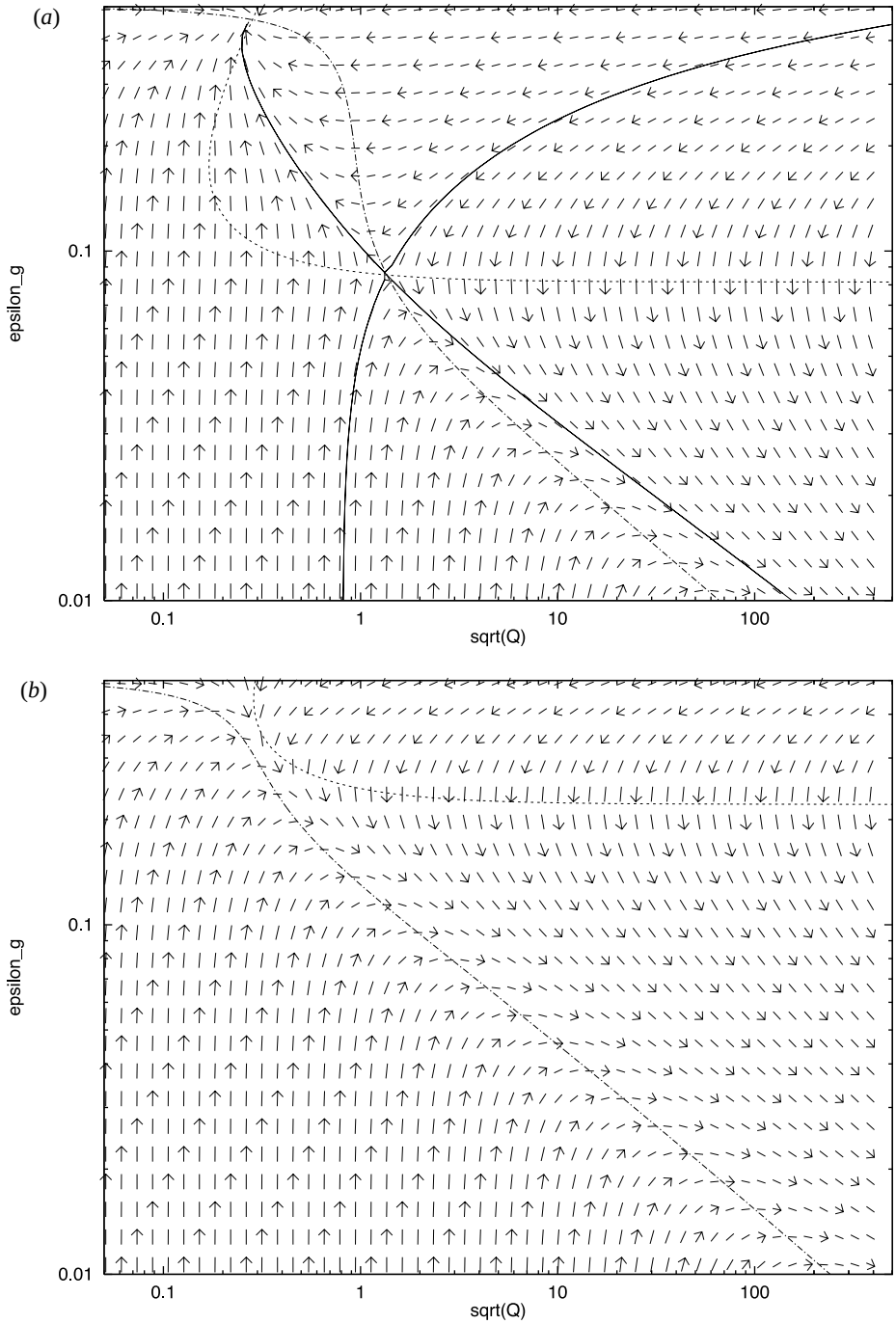


Figure 3. Flow in the plane ϵ_g, $\mathcal{Q}$. The dot-dashed curve corresponds to stationarity condition $\mathrm{d}\epsilon_g/\mathrm{d}\alpha = 0$, the dashed curve to $\mathrm{d}\mathcal{Q}/\mathrm{d}\alpha = 0$. (*a*) $L = 10$, $\lambda = 0.075$; two fixed points (one all stable, one partially stable) clearly show up, the full curves represent the stable and the unstable manifolds of the partially stable fixed point. (*b*) $L = 5$, $\lambda = 0.2$, a parameter setting for which there are no fixed points. For every starting point we have convergence to perfect generalization. (Cf also figure 2.)

Table 1. Critical value of λ for $L = 10$ and various initial conditions.

$\sqrt{\mathcal{Q}_0}$	1	10	100	1000	10 000
$\lambda_c(\mathcal{Q}_0)$	0.2545(5)	0.2185(5)	0.1385(5)	0.0875(5)	0.0485(5)

In figure 4 we directly plot $\epsilon_g(\alpha)$. As can be seen from all these figures, for $\lambda < \lambda_c$ the training leads to an initial improvement which is, however, limited and followed by a very rapid deterioration toward confusion. For $\lambda > \lambda_c$, in contrast, the learning stabilizes and leads to asymptotically perfect generalization with a λ-dependent power law in agreement with equations (34), (35).

These analytic results compare very well with the numerical results given in the previous section, both in the pre-asymptotic and in the asymptotic region (cf figure 1).

5. Summary and discussion

In the present paper we have investigated a two-phase learning algorithm for perceptrons, named the AR-Hebb algorithm. Its first phase consists of a series of Hebb-type synaptic modifications, correlating, however, input and *self-computed* output (blind association) rather than input and clamped teacher output. This first phase is followed by an unspecific but graded reinforcement-type learning step which leads to a partial reversal of the previous series of Hebb-type synaptic modifications, depending on current average success rates.

Our main motivation has been biological, attempting to honour the observation that a learner's control over its neurons and synapses might be less specific and direct than ordinary supervised learning algorithms usually presume, while basically adhering to the Hebbian learning paradigm.

Our central results can be stated as follows:

(i) Despite the fact that feedback on the learner's performance enters its learning dynamics only in an *unspecific* way in that it cannot be associated with a single identifiable correct or incorrect associations, convergence of the AR-Hebb algorithm in the sense of *asymptotically perfect generalization* is observed.

(ii) For given initial conditions, this convergence depends on the parameters of the algorithm; in particular none of these parameters can be set to zero. Alternatively, at fixed L and the ratio of the algorithm parameters convergence may depend on initial conditions.

In the details the dynamics of this algorithm was found to be unexpectedly complex. Depending on the parameters, fixed points in the dynamic flow may emerge—one stable, the other only partially stable. The attracting manifold of the latter constitutes a separatrix dividing initial states into two sets, one for which the algorithm converges, and another for which it does not in which case the flow is driven to the all-stable fixed point with poor generalization. Seen from a different point of view, a *given initial condition* (given updating speed) may be found to belong to the asymptotically converging lot, or to end up in a state of poor generalization, depending on network parameters.

On the other hand, parameter settings may be varied in such a way that the two fixed points eventually coalesce and disappear, rendering convergence of the algorithm independent of initial conditions. The pre-asymptotic regime of the learning process is then still influenced by the lines in the ϵ_g–$\mathcal{Q}$ plane along which either $\mathrm{d}\epsilon_g/\mathrm{d}\alpha$ or $\mathrm{d}\mathcal{Q}/\mathrm{d}\alpha$ (but not both) vanish.

Much to our surprise, the *convergence rate* of the algorithm was found to depend in a *non-universal manner* on the ratio of learning parameters. In spite of the non-specific nature

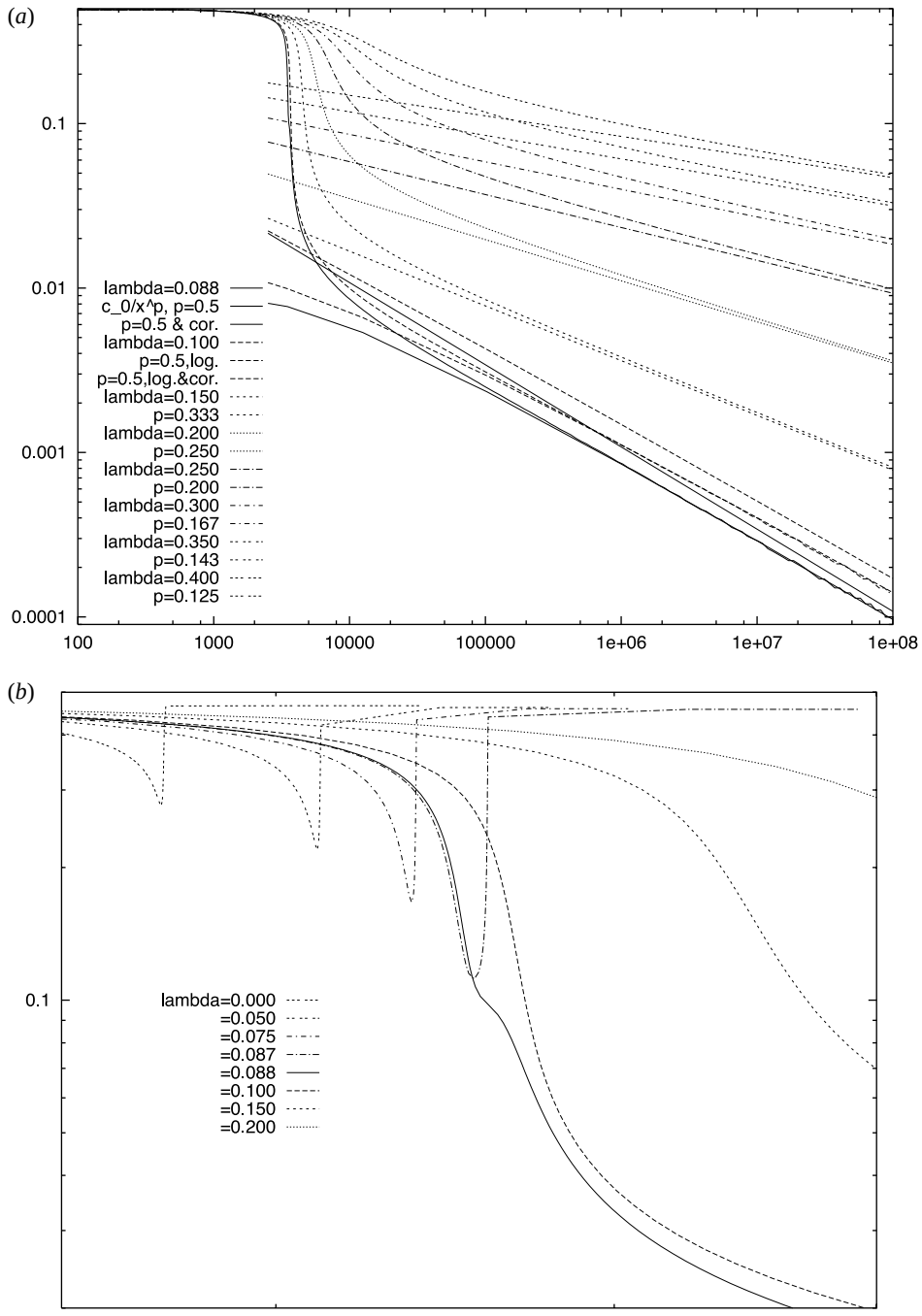


Figure 4. Generalization error ϵ_g versus α at $L = 10$ for various λ and for starting point $\mathcal{Q}(0) = 1000$. The straight lines in (*a*) show the dominant asymptotic behaviour for the corresponding λ from (34), (35) (notice that for $\lambda \leqslant 1/L$ the normalization is fixed; for $\lambda \leqslant 1/L$ we also give a fit using the subdominant terms in (34), (35)). (*b*) Amplified view at the pre-asymptotic region.

of the information feedback on the learning dynamics, convergence can be as fast as that of Hebbian learning, $\epsilon_g \sim \alpha^{-1/2}$, if $\lambda L < 1$, whereas it is slower and exhibits a non-universal parameter dependent rate, $\epsilon_g \sim \alpha^{-1/2\lambda L}$, if $\lambda L > 1$. Logarithmic corrections appear in the marginal case $\lambda L = 1$.

One may ask oneself, why there is no generalization for a perceptron-type algorithm $\lambda = 0$ (i.e., $a_1 = 0$). We can offer a simple observation which may be of heuristic value: since for $L = 1$ e_q can only be 0 or 1 $a_1 = 0$ means penalty for failure, no change for success, i.e. the usual perceptron learning rule known to converge. However, for $L > 1$ e_q can take fractional values in the interval [0, 1]. In this case $a_1 = 0$ means penalty for all answers which are short of perfect, i.e. even if the pupil is successful in far above 50% of the cases. This procedure can turn out to be destructive.

To put our findings into a broader perspective, it is perhaps appropriate to note that a similar kind of unspecific information feedback as in our setup occurs in committee-machine learning. While in our case, information feedback is unspecific in time (with respect to the pattern labels within a longer series on which the learner may have been in error), unspecificity in the committee machine refers to space, i.e., the label(s) of the node(s) which may have contributed to a wrong output upon presentation of a single pattern. In the details, though, the way in which unspecific feedback is utilized in the dynamics is different in the two setups, leading to different asymptotic laws, and to different behaviour in the pre-asymptotic regime. Although plateaus in the learning dynamics occur in both setups, this similarity is superficial. Whereas in the committee machine, the appearance of plateaus is related to a permutation symmetry of the nodes and escape therefrom to its breaking (a transition to specialization), there is strictly speaking no time-translation symmetry within a coarse-grained step, and no breaking thereof, as each coarse-grained step constitutes a whole correlated path of events during which the learner already evolves in response to the patterns presented. Quantitatively the difference manifests itself in the fact that plateaus in our setup have a much higher generalization error than those in the committee machines, and that the AR-Hebb rule may converge to a state of poor generalization even if its its initial performance is almost perfect (as can be seen in figure 3(*a*)). Still, it may be interesting to enquire whether techniques akin to those invented in order to decrease the extent of plateaus in committe-machine learning (see [14] for a recent reference) might be utilized to improve the present setup.

We have not addressed issues related to optimal parameter settings or optimal online-control of parameters (the latter issue would in some sense run against our original biologically minded starting point), nor have we investigated the performance of the algorithm in multi-layer architectures so far. Clearly these may be interesting topics to pursue in future research, as may be more detailed investigations of the algorithm as an intricate dynamical system *per se*.

Note added in proof. We would like to add the following interesting observation. A variant of the present algorithm which introduces an additional biologically motivated element of indeterminism by including patterns in the second (reinforcement) phase of a session only with probability $p < 1$ shows qualitatively the same behaviour as the algorithm studied in the present paper. A rough first quantitative characterization of this modification would be that it leads to an effective rescaling of the parameter a_2 of the algorithm by approximately a factor p, entailing a corresponding rescaling of the parameter λ and the scaled self-overlap $\mathcal{Q}$, viz. $\lambda \to \lambda/p$ and $\mathcal{Q} \to \mathcal{Q}/p^2$. This leads to a corresponding reduction of critical λ's for given initial condition Q_0 or, alternatively, to a reduction of the minimum Q_0 required for convergence at a given λ. These results are well corroborated by numerical simulations.

References

[1] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Reading, MA: Addison-Wesley)

[2] Kaelbling L P (ed) 1996 *Machine Learning* **22**

[3] Watkins C J C H 1989 Learning from delayed rewards *PhD Thesis* unpublished
[4] Sutton R S 1988 *Machine Learning* **3** 9
[5] Barto A G, Sutton R S and Anderson Ch W 1983 *IEEE Trans. Syst. Man Cybernetics* **13** 834
[6] Mlodinow L and Stamatescu I-O 1985 *Int. J. Comput. Inform. Sci.* **14** 201
[7] Stamatescu I-O 1998 Statistical features in learning (contribution to LEARNING '98, Madrid) *Preprint* cond-mat/9809135
[8] Stamatescu I-O 1996 The neural network approach *Proc. 1st Int. Conf. Philosophy of Science: Philosophy of Biology (Vigo, 1996)* ed J T Suarez (Vigo: University of Vigo) p 311
[9] Biehl M, Kühn R and Stamatescu I-O in preparation
[10] Biehl M and Riegler P 1994 *Europhys. Lett.* **28** 525
[11] Vallet F 1989 *Europhys. Lett.* **8** 747
[12] Kinouchi O and Caticha N 1992 *J. Phys. A: Math. Gen.* **25** 6243
[13] Biehl M and Schwarze H 1992 *Europhys. Lett.* **20** 733
[14] Bös S 1998 *J. Phys. A: Math. Gen.* **31** L413