Biol Cybern (2007) 97:99–112
DOI 10.1007/s00422-007-0162-4

PROSPECTS

Biological
Cybernetics

Learning with incomplete information and the mathematical structure behind it

Reimer Kühn · Ion-Olimpiu Stamatescu

Received: 4 January 2007 / Accepted: 2 May 2007 / Published online: 30 May 2007

Abstract We investigate the problem of learning with incomplete information as exemplified by learning with delayed reinforcement. We study a two phase learning scenario in which a phase of Hebbian associative learning based on momentary internal representations is supplemented by an 'unlearning' phase depending on a graded reinforcement signal. The reinforcement signal quantifies the success-rate globally for a number of learning steps in phase one, and 'unlearning' is indiscriminate with respect to associations learnt in that phase. Learning according to this model is studied via simulations and analytically within a student–teacher scenario for both single layer networks and, for a committee machine. Success and speed of learning depend on the ratio λ of the learning rates used for the associative Hebbian learning phase and for the unlearning-correction in response to the reinforcement signal, respectively. Asymptotically perfect generalization is possible only, if this ratio exceeds a critical value λ_c, in which case the generalization error exhibits a power law decay with the number of examples seen by the student, with an exponent that depends in a *non-universal* manner on the parameter λ. We find these features to be robust against a wide spectrum of modifications of microscopic modelling details. Two illustrative applications—one of a robot learning to navigate a field containing obstacles, and the problem of identifying a specific component in a collection of stimuli—are also provided.

R. Kühn (✉)
Department of Mathematics, King's College, London, UK
e-mail: reimer.kuehn@kcl.ac.uk

I.-O. Stamatescu
FESt, Heidelberg and Institut für Theoretische Physik,
Universität Heidelberg, Heidelberg, Germany
e-mail: I.O.Stamatescu@thphys.uni-heidelberg.de

1 Introduction

Learning and the ability to learn are important factors in development and evolutionary processes (Menzel 2003). Depending on the level, the complexity of learning can strongly vary. While associative learning can explain simple learning behaviour (Menzel 2003; Byrne 1999) much more sophisticated strategies seem to be involved in complex learning tasks. This is particularly evident in machine learning theory (Mitchell 1997; Sutton and Barto 2000; Vapnik 1998), but it equally shows up in trying to model natural learning behaviour (Byrne 1999).

A general setting for modelling learning processes in which statistical aspects are relevant and which is of particular interest for natural, learning by experience situations is provided by the neural network (NN) paradigm. NN learning models can incorporate elementary learning mechanisms based on neuro-physiological analogies and lead to quantitative results concerning the dynamics of the learning process (Hertz et al. 1991). However, for realistic problems, simple mechanisms often do not work and the sophistication of the algorithms rapidly increases.

Any realistic form of learning is in some sense learning from experience, since a learner interacts with an 'environment', appraises this interaction and consequently changes its 'internal structure' according to some criteria. In models of biological behaviour, as well as in the design of information processing systems, the appraisal procedure—food, pleasure, success, assessment of result—is formalized as some kind of *reinforcement*. Normal 'experience' can, however, rarely be encoded into one-to-one relations between actions and results, and a learner faces the additional task to *interpret* the environmental feedback before rewriting it as an update of its internal (cognitive) structure. An urgent problem, for example, with which an 'agent', either natural or artificial, may be confronted is to learn solely from the *final* success/failure of a series of consecutive actions, without direct information about the particular fitness of each of them, such as learning to coordinate its muscles in moving, learning a labyrinth or playing chess. Another case may be that of a feedback which only differentiates between the presence or absence of a certain kind of pattern in a mixture, such as reacting to unpalatable components in food, or realizing that some actions out of many were useful, without knowing which.

How non-trivial such problems are, can be seen from the sophisticated algorithms developed for so called 'delayed reinforcement' in the framework of machine learning theories (Sutton and Barto 2000; Wyatt 2003). In this perspective realistic, complex learning situations seem to require the availability of strategies and involved procedures *before* reinforcement can provide effective feedback mechanisms. In an evolutionary perspective of the development of learning capabilities, however, the question then arises as to how such strategies and procedures could have developed in the first place. If reinforcement were to be a general or even fundamental element involved in effecting behavioural change, one would expect reinforcement learning to act *in particular* at an intermediate level which is simple enough not to depend on involved strategies, yet powerful enough to allow complex behaviour to evolve.

A problem of *non-specific* reinforcement can be defined as follows: the reward is global, regards the cumulative result of a series of actions and the reinforcement acts non-specifically concerning these actions (Mlodinov and Stamatescu 1985). We ask whether there may exist *elementary mechanisms* that allow learning under such conditions of non-specific feedback, which may have developed also under natural conditions and which may hint at basic features of learning. For this we must not only demonstrate the existence of such mechanisms, but also uncover their structural features.

In previous papers (Kühn and Stamatescu 1999; Biehl et al. 2000; Stamatescu 2003) we introduced and studied a learning algorithm for neural networks (NN) which deals with this problem. The NN setting allows a systematic study by both numerical and analytic methods. It provides a framework to study learning with non-specific reinforcement which is transparent, and pertains to the basic machinery on which learning is believed to take place. NN can achieve complex information processing capabilities using mechanisms simple enough to have plausibly developed under natural conditions. They can model various levels of learning processes—biological, behavioural or cognitive (Hertz et al. 1991).

The purpose of the present contribution is to put the issue of learning with incomplete information into a broader perspective, paying due respect to basic underlying issues and principles, as well as generalizing previous results to a wider range of settings, including multi-layer networks and real-world applications (Kühn and Stamatescu 1999; Biehl et al. 2000; Stamatescu 2003; Bergmann et al., in preparation). Our hope is that we may thereby help understanding some basic features of learning.

Our paper is organized as follows. In Sect. 2 we describe an elementary scenario of learning with incomplete information, using a teacher-student setup for perceptron learning. Elements of the analysis are included here. In Sect. 3 we describe our results for the standard perceptron setting as well as for some variants concerning statistics of input data, the nature of the unspecific feedback signal, and more, while Sect. 4 is reserved for results on a simple two-layer network, the committee-machine. In Sect. 5, we present two 'real world' applications, viz. that of a robot learning to avoid obstacles, and that of learning to identify whether a certain key stimulus is contained in a collection of other stimuli. Section 6 provides a summary and concluding remarks. Two appendices are included to cover the more technical aspects of the analytic investigations in greater detail.

2 The learning scenario

We start our analysis by casting the problem into a classification task for a perceptron. In its *simplest version* (to which we shall stick for most of the present paper), this is a network consisting of an array of N input neurons projecting synapses onto a single output neuron. The active and inactive states of the neurons are encoded as $+1$ and -1 respectively. The 'cognitive structure' of the network is encoded in the values J_i, $i = 1, \ldots, N$, of the synaptic strengths, also called weights. The inputs to the network (the 'patterns' which must be classified) are strings of N binary values $\xi_i \in \{\pm 1\}$ loaded on the input layer. These values are weighted by the corresponding synaptic strengths and transmitted to the output neuron, where they are added to define a 'potential'

$$x = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i \xi_i, \tag{1}$$

which, according to its sign, triggers the output neuron to $s = \pm 1$, thereby attributing the pattern $\boldsymbol{\xi} = (\xi_i)$ to the class s:

$$s = \operatorname{sgn}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i \xi_i \right) = \operatorname{sgn}(x). \tag{2}$$

The standard learning problem is stated by asking a 'student' perceptron to implement a given classification rule. The rule is provided by a 'teacher' perceptron with the same architecture, whose synapses B_i are given and fixed, and who classifies the pattern $\boldsymbol{\xi}$ according to

$$t = \operatorname{sgn}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} B_i \xi_i \right) = \operatorname{sgn}(y). \tag{3}$$

The student has access to the teacher's output t which provides the rule-based classification of the inputs, but *not* to the teacher's rule represented by its weight vector $\boldsymbol{B} = (B_i)$. The student learns from a stream of inputs classified by the teacher, $\boldsymbol{\xi}^\mu \rightarrow t^\mu$, $\mu = 1, 2, \ldots$, adapting its own weights in response to these data such as to reduce the disagreement between its own classification $\boldsymbol{\xi}^\mu \rightarrow s^\mu$ and that of the teacher.

A classical, and neuro-biologically motivated learning rule is the so-called Hebb rule, where synapses change in response to coincidence of pre- and post-synaptic activity. In the present setting the appropriate formulation of Hebb's rule for the student performing the classification $\boldsymbol{\xi}^\mu \rightarrow s^\mu$ takes the form

$$J_i^\mu \rightarrow J_i^\mu + \Delta J_i^\mu, \quad \Delta J_i^\mu \propto s^\mu \xi_i^\mu. \tag{4}$$

where J_i^μ is the *momentary* value of the i-th synapse when pattern μ arrives to be classified. Within an online learning scenario, the synapse is then updated according to $J_i^{\mu+1} = J_i^\mu + \Delta J_i^\mu$.

In order to turn this simple form of associative adaptation into *learning*, it must be supplemented with some feedback concerning the 'quality' of the associations. In learning rules traditionally used in this scenario the proportionality 'constant' in Hebb's rule *depends* on the teacher answers. For instance, the supervised Hebb algorithm uses the proportionality 'constant' $at^\mu s^\mu$, giving

$$\Delta J_i^\mu = at^\mu \xi_i^\mu, \tag{5}$$

where a is a parameter setting the learning rate. In the case of the Perceptron learning algorithm the proportionality 'constant' is given by $\frac{a}{2}|s^\mu - t^\mu| t^\mu s^\mu$, resulting in

$$\Delta J_i^\mu = \frac{a}{2}|s^\mu - t^\mu| t^\mu \xi_i^\mu \tag{6}$$

instead. Note in particular that implementations of either rule must assume *immediate and direct control* over the student's synapses by clamping the desired output onto the student neuron (e.g. by providing an 'evaluating' stimulus through some other channel – see, e.g. (Menzel 2003)).

That is, in standard 'supervised online learning' the student is told after each instance what the right answer would have been (specific feedback). *In our approach, however, according to our paradigm the student can only receive a non-specific reinforcement about some global degree of correctness of its answers over several instances*. More precisely the student is presented with series ('bags') of patterns and only obtains information concerning its cumulative performance for the bag as a whole. The non-trivial learning problem is to implement this global information into a local updating rule for the synapses.

2.1 The algorithm

The learning algorithm we propose can be described as consisting of two phases. In the following each bag q is taken to contain the same number L of patterns.

In the first phase the student processes the patterns $(\xi_i^{(q,l)})$ in a bag q one by one and modifies its synapses by simple Hebbian association, using its *own classifications*

$$s^{(q,l)} = \operatorname{sgn}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i^{(q,l)} \xi_i^{(q,l)}\right) \tag{7}$$

on the basis of its *momentary* synapse values $J_i^{(q,l)}$:

$$\text{I}: \quad J_i^{(q,l+1)} = J_i^{(q,l)} + \frac{a_1}{\sqrt{N}} s^{(q,l)} \xi_i^{(q,l)}, \quad l = 1, \ldots, L. \tag{8}$$

In the second phase the student receives information about its *global performance* on a whole bag of patterns and corrects its synapses by 'reconsidering' the steps of the first phase, and by *partially undoing them in an indiscriminate, i.e. uniform way, to an extent that depends only on the global error*, independently on which steps were in fact correct and which not. This phase can be seen as Hebbian 'unlearning':

$$\text{II}: \quad J_i^{(q+1,1)} = J_i^{(q,L+1)} - e_q \frac{a_2}{\sqrt{N}} \sum_{l=1}^{L} \omega_l s^{(q,l)} \xi_i^{(q,l)}. \tag{9}$$

Here the online bag error e_q is a measure of the disagreement between student and teacher, and defines the specific problem. We have looked at two different choices for e_q, which are introduced in Eqs. (22) and (23) in Sect. 3 below. In (9), the ω_l can be 1 or 0 with probabilities ρ and $1 - \rho$, respectively, which accounts for the possibility that the replay during the second phase may be *imperfect*: the student may not recall all associations established during the first phase.

The procedure, so to say, is specific but blind association in the first phase, qualified but non-specific reinforcement in the second phase. The algorithm has therefore been called *Association-Reinforcement (AR) - Hebb* algorithm.

It is interesting to note that a kind of replay as that involved in the Phase II of our algorithm can apparently be observed in rats on track running tasks (Foster and Wilson 2006) and is seen as memory consolidation. Since experiences usually imply valuations, it is suggestive that in such replay not only neutral memories are consolidated, but memories including some measure of success (finding or not finding food at the end of the track, for instance—a *global* reward). This would mean observing here a mechanism akin to the re-weighting replay of the Phase II of our learning model. Hebbian unlearning mechanisms via replay of data either previously exposed to or triggered through random stimuli has been discussed also in other contexts (Crick and Mitchison 1983; Hopfield et al. 1983; van Hemmen 1997; van Hemmen et al. 1990). Concerning Phase I, this represents just the strengthening of the student's own association rules by using them, which is itself the meaning of the original Hebb rule (4). This suggests that the 'AR-Hebb' algorithm may be provide good modelling of natural behaviour, and that it may in fact have a neuro-physiological basis.

2.2 Analysis

At fixed $\boldsymbol{B}$, the relevant quantities which describe the progress of learning for a perceptron are the normalized scalar products (Vallet 1989; Kinzel and Rujan 1990; Opper et al. 1990; Biehl and Riegler 1994)

$$R_q = \frac{1}{N}\boldsymbol{J}^{(q,1)}\cdot\boldsymbol{B} = \frac{1}{N}\sum_{i=1}^{N} J_i^{(q,1)}\,B_i, \tag{10}$$

$$Q_q = \frac{1}{N}\boldsymbol{J}^{(q,1)}\cdot\boldsymbol{J}^{(q,1)} = \frac{1}{N}\sum_{i=1}^{N}\left(J_i^{(q,1)}\right)^2, \tag{11}$$

here evaluated at the beginning of session q. They are commonly referred to as overlaps. The overlaps are 'order parameters' in the sense that the macroscopic dynamics of learning can be fully described in terms of R_q and Q_q alone. We take $N^{-1}\boldsymbol{B}\cdot\boldsymbol{B} = 1$. Note that Q_q controls the relative learning rate (the larger Q_q, the smaller the *relative* synaptic change induced by a single learning step). We denote by

$$t^{(q,l)} = \text{sgn}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} B_i\xi_i^{(q,l)}\right). \tag{12}$$

the classification of pattern $\boldsymbol{\xi}^{(q,l)}$ provided by the teacher perceptron.

The progress of learning is analyzed by looking at the *combined* effect of phase I and phase II learning (8) and (9) on Q_q and R_q for a whole bag. In the following, we shall exploit the fact that only the ratio $\lambda = a_1/a_2$ of learning-rates in phases I and II is relevant for the analysis, which follows after a simple rescaling of the student synapses with a_2, $J_i^{(q,l)}/a_2 \to J_i^{(q,l)}$. Using the association and reinforcement rules (8) and (9), and introducing

$$x^{(q,l)} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} J_i^{(q,l)}\xi_i^{(q,l)}, \quad y^{(q,l)} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} B_i\xi_i^{(q,l)}$$

to express the updates for (10) and (11) one obtains

$$R_{q+1} = R_q + \frac{g_q}{N}\sum_{l=1}^{L}\text{sgn}\left(x^{(q,l)}\right)y^{(q,l)} \tag{13}$$

$$Q_{q+1} = Q_q + 2\frac{g_q}{N}\sum_{l=1}^{L}\text{sgn}\left(x^{(q,l)}\right)x^{(q,l)} + L\frac{g_q^2}{N}, \tag{14}$$

where we have introduced the shorthand $g_q \equiv \lambda - e_q$. We have also made use of the orthogonality of unbiased binary random patterns in the large N limit, (i.e. the fact that $N^{-1}\sum_i \xi_i^{(q,k)}\xi_i^{(q,l)} = \delta_{k,l}$ by the law of large numbers).

The remainder of the analysis follows standard online learning reasoning (Vallet 1989; Kinzel and Rujan 1990; Biehl and Riegler 1994); it consists (i) in noting that for all finite bag sizes L the single-bag increments of the order parameters in (13) and (14) are infinitesimal in the thermodynamic limit $N \to \infty$,

$$\begin{aligned}\Delta R &= R_{q+1} - R_q = \mathcal{O}(N^{-1}),\\ \Delta Q &= Q_{q+1} - Q_q = \mathcal{O}(N^{-1}),\end{aligned} \tag{15}$$

(ii) in introducing 'continuous time' $\alpha = qL/N$ with likewise infinitesimal single-bag increments $\Delta\alpha = L/N$, and writing order parameters as functions of continuous time, $R_q \to R(\alpha)$, and $Q_q \to Q(\alpha)$, (iii) in realizing that the central limit theorem entails that the fields $x^{(q,l)}$ and $y^{(q,l)}$ are zero-mean Gaussian with correlations that depend *only* on $R(\alpha)$ and $Q(\alpha)$,

$$\begin{aligned}&\left\langle x^{(q,l)}\right\rangle = 0, \quad \left\langle\left(x^{(q,l)}\right)^2\right\rangle = Q(\alpha),\\ &\left\langle y^{(q,l)}\right\rangle = 0, \quad \left\langle\left(y^{(q,l)}\right)^2\right\rangle = 1, \quad \left\langle x^{(q,l)}y^{(q,l)}\right\rangle = R(\alpha),\end{aligned} \tag{16}$$

(the independence of the $\xi_i^{(q,l)}$ is used and $\mathcal{O}(N^{-1})$ corrections are neglected to obtain these results), and finally (iv) in combining a large number M of updates (13) and (14),

$$\frac{R(\alpha + M\Delta\alpha) - R(\alpha)}{M\Delta\alpha} = \frac{1}{ML}\sum_{m=0}^{M-1} g_{q+m}\sum_{l=1}^{L}\text{sgn}\left(x^{(q+m,l)}\right)y^{(q+m,l)} \tag{17}$$

$$\frac{Q(\alpha + M\Delta\alpha) - Q(\alpha)}{M\Delta\alpha} = \frac{1}{ML}\sum_{m=0}^{M-1}\left[2g_{q+m}\sum_{l=1}^{L}\text{sgn}\left(x^{q+m,l}\right)x^{q+m,l} + Lg_{q+m}^2\right] \tag{18}$$

to obtain an autonomous pair of ODEs in the limit $M \to \infty$, $N \to \infty$, $M/N \to 0$, hence $M\Delta\alpha \to 0$, which can be formulated in terms of *averages* over these updates by the law of large numbers,

$$\frac{\mathrm{d}R}{\mathrm{d}\alpha} = \left\langle\frac{g_q}{L}\sum_{l=1}^{L}\text{sgn}\left(x^{(q,l)}\right)y^{(q,l)}\right\rangle, \tag{19}$$

$$\frac{\mathrm{d}Q}{\mathrm{d}\alpha} = 2\left\langle\frac{g_q}{L}\sum_{l=1}^{L}\text{sgn}\left(x^{(q,l)}\right)x^{(q,l)}\right\rangle + \left\langle g_q^2\right\rangle. \tag{20}$$

These equations describe the learning dynamics in the thermodynamic limit. The angled brackets denote averages over the bivariate Gaussians $x^{(q,l)}$ and $y^{(q,l)}$, which are uncorrelated for different l and have correlations given by (16), thus can be evaluated in terms of $R(\alpha)$ and $Q(\alpha)$ alone (Kühn and Stamatescu 1999). The integrals in terms of which the right hand sides of (19) and (20) can be evaluated will depend on the nature of the online bag error e_q used to quantify the reinforcement signal, which enters these equations through $g_q = \lambda - e_q$. We describe them briefly in Appendix A.

The quantity of interest is the 'generalization error' ϵ_G which measures the probability of disagreement between the teacher and the student on a random set of patterns. It can be expressed in terms of the angle between the weight vectors of student and teacher, respectively, one obtains (Vallet 1989; Kinzel and Rujan 1990; Biehl and Riegler 1994; Opper et al. 1990).

$$\epsilon_G(\alpha) = \frac{1}{\pi} \arccos(\boldsymbol{J}^{(q,1)}; \boldsymbol{B}) = \frac{1}{\pi} \arccos\left(\frac{R(\alpha)}{\sqrt{Q(\alpha)}}\right). \tag{21}$$

Note that ϵ_G refers to the classification error for any randomly chosen pattern, while e_q of phase II refers to an empirical error measured over a single bag of L patterns.

3 Results for single layer networks

We study the system by simulations, which implement the network operation (7) and the learning rules (8) and (9) at the microscopic level, and analytically via the ODE's (19) and (20), which describe network performance at a macroscopic level. They can be solved numerically at all α and sometimes also analytically for large α (Kühn and Stamatescu 1999; Biehl et al. 2000; Bergmann et al., in preparation). They reveal the fixed point structure that directs the flow as discussed in what follows.

The present section reviews results obtained for single layer networks. The main quantity of interest is the generalization error ϵ_G, and its evolution during learning. It can either be computed via (21) from the solution of (19) and (20), or *measured* in numerical simulations. In the simulations reported in the present and the following section, ϵ_G is measured by comparing the student and teacher answers on a random set of 10^4 patterns.

We have in fact looked at two variants of learning with incomplete information which are characterized by different versions of the online bag error e_q measuring the performance of the student on the patterns of bag q. The 'average error' (AE) problem (Kühn and Stamatescu 1999) is defined by using the global return (error):

$$e_q^{AE} = \frac{1}{2L} \sum_{l=1}^{L} \left| s^{(q,l)} - t^{(q,l)} \right|. \tag{22}$$

as the bag-error measure in (9), $e_q = e_q^{AE}$. It measures the fraction of inputs in the current bag, on which student and teacher disagree.

The second version, to be referred to as 'hidden instance' (HI) problem uses a global return of the form

$$e_q^{HI} = \frac{1}{4L^2} \left[\sum_{l=1}^{L} \left(s^{(q,l)} - t^{(q,l)} \right) \right]^2 \tag{23}$$

for e_q in (9) instead, which amounts to measuring the discrepancy between student and teacher concerning the *balance* between positively and negatively classified input data in the current bag. Note that $e_q^{HI} \le e_q^{AE}$ by a Schwarz inequality, and that the HI problem involves *more indeterminacy* than the AE problem. Instead of squares one can use absolute values and vice versa.

In the case to be presented first, the non-specific reinforcement uses the *average (or global) error* of the student's guesses for the whole bag. We call this the 'average error' (AE) problem—see Eq. (22). At no moment does the student know whether his particular classifications are correct or incorrect; he is only informed about the *fraction* of correct answers over the whole bag.

The exciting result of this study is that stable perfect learning can be achieved in spite of the non-specific reinforcement. The most interesting feature is the dependence of learning on λ which measures the strength of the local 'associative' compared to the global 'corrective' step.

It is found that the asymptotic decay of ϵ_G as a function of the number q of pattern-bags used for learning is described by a power-law, $\epsilon_G \sim q^{-p}$ with an exponent p that depends on λ. But even more compelling is the appearance of a threshold λ_c below which *no learning is possible*, whereas for $\lambda > \lambda_c$ perfect learning is always achieved. The exponent p is a decreasing function of λ, so that learning becomes more efficient as $\lambda \searrow \lambda_c$. The value of λ_c itself depends on L and on the initial value Q_0 of the student's self-overlap, that is, on the initial effective learning rate $(1/\sqrt{Q_0})$. *There is no such non-zero threshold for* $L = 1$, where $0 \le \lambda \le \frac{1}{2}$ just interpolates between the standard Hebb and Perceptron algorithms in (5) and (6).

Figure 1 shows typical results of numerical simulations. Below λ_c (which for the given initial conditions is located between 0.120 and 0.125) ϵ_G initially decreases with increasing number q of processed bags, then suddenly returns to a value very close to 0.5 and stays there ever after. Just above λ_c, learning is rapid but may be disrupted by finite size fluctuations, which entail that the threshold λ_c is somewhat fuzzy at finite N. Further above λ_c finite-size fluctuations cease to be effective in disrupting learning, but convergence to perfect generalization is also slower. The plot gives ϵ_G vs the normalized number of processed patterns $\alpha = qL/N$. The straight lines indicate the asymptotic behaviour expected for the corresponding λ from the analytic theory presented in Appendix A.

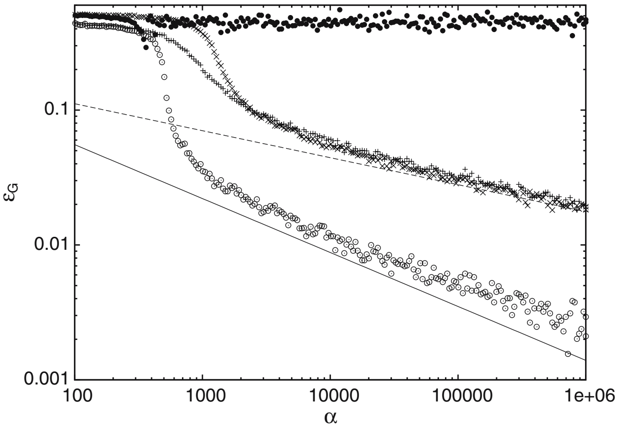

Fig. 1 Numerical simulation of the AE problem: the learner is presented with bags of patterns, feedback is given only via the *average (or total) error* made over the whole bag. We plot ϵ_G versus the normalized number of patterns $\alpha = qL/N$. Each of the bags contains $L = 10$ patterns; the size of the input layer is $N = 100$. Initial conditions are random with $R(0) = 0$ hence $\epsilon_G(0) = 0.5$, and $Q(0) = 10^4$. The replay phase is either complete ($\rho = 1$, *full circles, empty circles* and *crosses*) or partial ($\rho = 0.5$, *asterisks*). For $\lambda = 0.120$ (*full circles*) no generalization is observed, whereas for $\lambda = 0.125$ (*empty circles* and *asterisks*) and $\lambda = 0.250$ (*crosses*) perfect generalization is achievable. The *lines* represent the asymptotic behaviour expected from the analytic study $\epsilon_G(\alpha) \propto \alpha^{-\frac{1}{2L\lambda}}$

This peculiar dependence on the learning parameter λ—the increase of the learning efficiency with decreasing λ (i.e., with increasing strength of the corrective step) up to the point where the learning behaviour suddenly breaks down (no learning below λ_c) is particularly remarkable. While there is no simple *intuitive* explanation for either, the analytic theory explained in the previous section is able to uncover the *mechanism* underlying both features.

Indeed, the results observed in the simulations can be fully rationalized by the findings of the analytic investigation to which we now turn, thereby also proceeding to the second step—that of answering structural questions.

The mathematical structure uncovered in this way shows that the phase-flow (19) and (20), which describes the evolution of learning in the large system limit, is governed by a pair of fixed points, one fully stable, and the other partially stable with an attractive and a repulsive direction (see also (Kühn and Stamatescu 1999) for further details).

In Fig. 2 we present analytical results for the learning process, obtained by solving the flow equations (19) and (20) with the error definition (22).

The left panel is to be compared with the simulation results shown in Fig. 1. In the right panel, which shows the phase-flow of the learning dynamics in the $\epsilon_G, \sqrt{Q}$ −plane, we can clearly discern the existence of a separatrix connecting the starting point and a partially stable fixed point. Its repulsive manifold directs the flow either towards large Q and perfect generalization, or towards small Q and an all-attractive fixed point of poor generalization. The alternative is decided by λ which determines on which side of the separatrix the initial condition finds itself. By changing λ we actually move the fixed points around, and continuously deform the repulsive and the attractive manifolds, thereby sweeping the separatrix across the initial condition, which at λ_c exactly lies *on* the separatrix. The threshold is crisp, of course, as there are no longer any finite size fluctuations. The small difference between analytically and numerically determined values of the threshold λ_c is also a finite size effect.

The fact that *global* properties of the phase-flow are governed by the interplay of two fixed points, in particular by the stable and unstable manifolds of the partially stable fixed point, is of particular relevance. The stable manifold provides the separatrix which is responsible for the fact that the system can show either good or poor generalization, while the unstable manifold is what is actually *directing* the asymptotics of the flow in either case.

It can fairly be expected that small deformations of the right hand sides of the flow equations (19) and (20) by continuous functions of R and Q—these would correspond to modifying microscopic details of the learning dynamics—will *not* change the *global* properties of the phase flow: the two fixed points would continue to exist and maintain their stability properties, and so will the stable and unstable manifold of the partially stable fixed point, provided the deformations remain sufficiently small. One must of course expect that small perturbations will move the fixed points to (slightly) different locations in the phase plane, and that they will continuously deform the stable and the unstable manifolds. However, since the effects of such perturbations are both small and continuous, we can expect the two main features of the learning dynamics—threshold behaviour for the mere existence of good generalization, and non-universal behaviour, i.e., parameter dependence of the asymptotic rate of convergence to good generalization, to be *structurally* stable under a range of modifications of the original setup, and perhaps even be fairly general properties of our paradigm of learning with incomplete information.

A number of further studies were performed in order to substantiate this suggestion.

Incomplete replay: First, we investigated what happens if the ‘replay’ in Phase II is not perfect — accounting for the possibility that the student may not ‘remember’ all instances encountered during Phase I. We modelled this by randomly including each instance of Phase I only with probability $\rho \leq 1$ in the unlearning step of Phase II. It turned out that all features observed for complete replay are preserved; for the asymptotic domain of good generalization the modification amounts to a re-scaling of the learning parameter $\lambda \to \lambda/\rho$ (Kühn and Stamatescu 1999) – see Fig. 1.

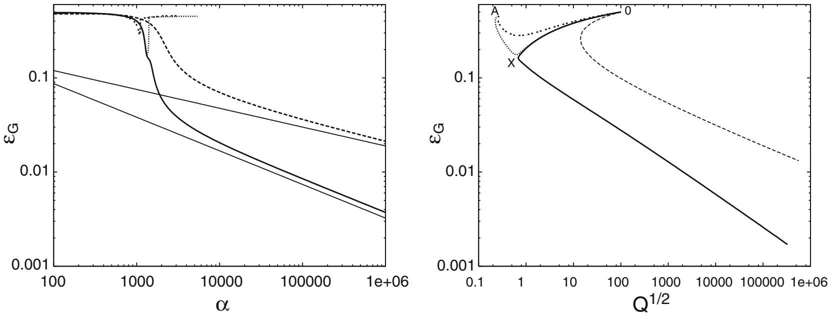

Fig. 2 Analytic study of the AE problem, $L = 10$. Shown are ϵ_G versus α (*left*), and the *flow* of the learning dynamics in the $\epsilon_G, \sqrt{Q}$ −plane (*right*) for fixed initial condition $R(0) = 0$, $Q(0) = 10^4$, and $\lambda = 0.100$ (*doubly dotted line*), 0.138 (*dotted*), 0.139 (*solid*) and 0.250 (*dashed line*). The threshold appears at $\lambda_c \simeq 0.1385$, for which the initial condition (marked with an 0) is exactly on the attractive manifold of a partially stable fixed point. The attractive fixed point (A) and the partially stable fixed point (X) are clearly identifiable. The *straight lines* on the left plot represent exact asymptotic power laws with $p = \frac{1}{2L\lambda}$

Fluctuating bag-size: Also randomly varying bag sizes L used in learning leads to results qualitatively unchanged when compared to the fixed L case. This is easily understood by noting that the law of large numbers applied to the $M \to \infty$-limit of (17) and (18) amounts to an additional sampling over the bag-size distribution $p(L)$; hence Eqs (19) and (20) remain formally unaltered, except for noting that angled brackets should be interpreted to imply an additional average over $p(L)$.

The hidden instance problem: An extension of the problem setting concerns the nature of the online bag error e_q. We investigated a case where the student is only told the *number* of patterns of, say, class +1 in a bag, which it can compare with the number perceived to belong to that class by itself – see Eq. (23). We called this the 'hidden instance problem' (HI). The information received by the student thus has an *increased degree of non-specificity* (since, e.g., $e_q = 0$ does not yet imply that the student has found the correct classifications!). Nevertheless the learning behaviour shows the same qualitative features as for the AE-problem: a threshold λ_c and a λ-dependent asymptotic power-law decay of ϵ_G above the threshold. This can be seen both in simulations, and in the corresponding analytic theory (see Fig. 3), which requires analyzing the flow equations (19) and (20) with the HI bag error definition (23), i.e., Eqs. (32) and (33) of Appendix A. Finite size effects are larger in the HI problem, which is mainly due to the larger degree of non-specificity in the feedback (its noisiness): e.g. for the parameters used in Fig. 3, numerical simulations show convergence to good generalization already for runs with $\lambda \simeq 0.014$. The system reaches the large-Q, low-ϵ_G region only with the help of strong finite size fluctuations, as the analytic study gives $\lambda_c \simeq 0.0189$ in the thermodynamic limit. However, having reached the region of good generalization, it then exhibits asymptotics fully in accordance with theoretical predictions. No systematic attempt has been made to confirm the analytic result for λ_c through simulations by measuring the fraction of runs leading to perfect (or bad) generalization, which is expected to develop a jump-discontinuity at λ_c in the $N \to \infty$ limit. Further details on this case will be published elsewhere (Bergmann et al., in preparation).

Structured inputs: A further extension looked into using structured input data in the classification task to be learnt, as a highly schematic way to model learning in a structured environment (Biehl et al. 2000). Instead of 'isotropic' patterns with $\xi_i^\mu = \pm 1$ independent and identically distributed, and with zero average $\langle \xi_i^\mu \rangle = 0$ as used in the present investigation, it is assumed that the ξ_i^μ are unit variance Gaussians, centred at $\pm m C_i$, where $\boldsymbol{C} = (C_i)$ is a fixed random vector with $\boldsymbol{C}^2 = N$ and overlap $N^{-1}\boldsymbol{C} \cdot \boldsymbol{B} = \eta$ with the coupling vector of the teacher perceptron, where m and η are parameters of the problem. Note that $\boldsymbol{C}$ and $\boldsymbol{B}$ are identical for $\eta = 1$, and the teacher perceptron is in this case particularly suitable for classifying such patterns, as the centres of the two Gaussians have maximal distance from the decision surface defined by $\boldsymbol{B}$. For other values of η the additional structure in the patterns complicates the dynamics (and its analysis) considerably. The analysis requires *three* order parameters for a full macroscopic description instead of two—namely apart from R_q and Q_q also the overlap $D_q = N^{-1}\boldsymbol{J}^{(q,1)} \cdot \boldsymbol{C}$ between the student coupling vector and the vector $\boldsymbol{C}$ characterizing the pattern-anisotropy. Also, some supplementary but simple tuning of the learning rate λ was necessary, as indeed for the corresponding *standard* supervised algorithms dealing with the same problem. Yet again we observed that perfect generalization requires λ to exceed a threshold λ_c, and find a λ dependent asymptotics. Details can be found in (Biehl et al. 2000).

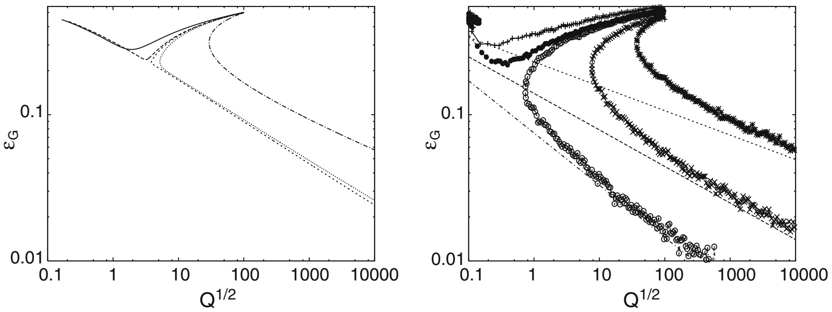

Fig. 3 Comparison of analytic and numerical results for the HI problem for $L = 10$, using (23) as online bag error e_q for feedback. Plotted is ϵ_G versus $\sqrt{Q}$. Initial conditions are random with $R(0) = 0$, hence $\epsilon_G(0) = 0.5$, and $Q(0) = 10^4$ as before. The analytic study (*left*) shows a threshold $\lambda_c \simeq 0.0189$, distinguishing between lack of generalization ($\lambda = 0.015,\ 0.0185$) and good learning ($\lambda = 0.019,\ 0.02,\ 0.03$). In the numerical simulation, the size of the input layer is $N = 200$. No generalization is observed for $\lambda = 0.01$ (*pluses*) and $\lambda = 0.012$ (*full circles*), whereas for $\lambda = 0.014$ (*empty circles*), 0.02 (*crosses*) and 0.03 (*asterisks*) perfect generalization is achievable. *Straight lines* in the simulation plot represent the large-α behaviour as expected from the asymptotic theory (36) and (37): $\epsilon_G(\alpha) \propto \alpha^{-\frac{1}{2L^2\lambda}}$, respectively $\epsilon_G(\alpha) \propto Q^{-\frac{1}{4L^2\lambda}}$

4 The committee machine

In the cases discussed above we investigated single-layer networks, for which the solvable classification tasks are limited to the class of so-called linearly separable problems. This implies that problems in a different complexity class, and presumably some of the more realistic ones, can at first sight not be solved by such networks, irrespectively of the learning algorithm used to train them.

It is now well known that the limits of linear separability can be transcended by the use of preprocessing and kernel methods (Vapnik 1995; Cristiani and Shawe-Taylor 2000; Schölkopf and Smola 2002; Shawe-Taylor and Cristiani 2004), so as to provide the capability for universal classification while adhering to the perceptron as the trainable neural element.

Nevertheless, in order to lend further credibility to the hypothesis that our model could provide a basic learning mechanism at work also *beyond the single neuron level*, and thereby plausibly contribute to the evolution of complex information processing capabilities in neural architectures, we investigate its performance on a simple two-layer network (Bergmann et al., in preparation), the so called 'committee machine'. This is a two-layer network with the neurons of the second (hidden) layer—the committee—transmitting their state via fixed synapses to the output neuron. Only the synapses from the input neurons to committee members can be modified in the learning process.

There is an important second motivation, beyond that of demonstrating the viability of the non-specific reinforcement principle for training simple multi-layer networks, capable of performing classifications outside the linearly separable class: It is related to the fact that the single output of a multilayer network is *itself non-specific* in the sense of not revealing which of possibly several states of the hidden layer was responsible for it. Specifically, in the case of a committee-machine producing a simple majority vote of the committee members, no information is revealed as to which subset of the committee was backing the majority vote. This is non-specificity with respect to contributions of hidden nodes (for simplicity referred to as 'non-specificity with respect to space'), whereas the *AR - Hebb* algorithm introduces an element of non-specificity with respect to time. By using the *AR - Hebb* algorithm to train a committee machine, we *combine* non-specificities in space *and* time, and the natural question arises whether this further reduction of the information used for feedback still permits that a rule—represented by a teacher committee of the same architecture—can be picked up on the basis of classified inputs alone.

We are able to report here recent first results about this system. The version we have looked at is a 'graded' version producing as its output the sum of the outputs of all committee members, without performing a final sign-operation on that sum. See Appendix B for details. The simulations do indeed show convergence to perfect generalization, and a threshold in the learning parameter λ, as for the perceptron. See Fig. 4. However, this turns out to be combined with an even more complex picture of the evolution of the order parameters—mutual overlaps between the coupling vectors of *all* hidden units of the student and teacher committee are needed for a full description of the dynamics. A more complicated phase flow and fixed point structure is thus to be expected. E.g. a partially stable 'symmetric' fixed point exists representing the student committee in a state where its hidden nodes have not yet specialized to which of the hidden nodes of the teacher committee they are eventually going to represent (Saad and Solla 1995). In addition there appears to be a pair of fixed points specifically associated with the unspecific

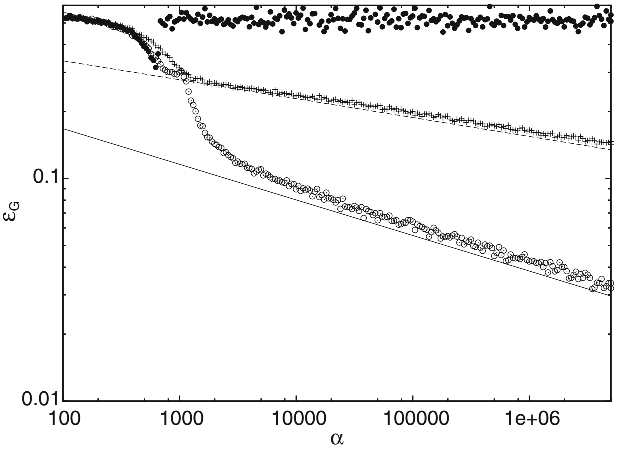

Fig. 4 Numerical simulation for a committee machine with $K = 3$ hidden units and $N = 100$ input units, AE problem with $L = 5$: generalization error ϵ_G vs α. Here $\lambda = 0.260$ (*full circles*), 0.280 (*circles*) and 0.320 (*crosses*). The *straight lines* represent power-law fits to the asymptotic behaviour with exponents $p = 0.16$ (*solid line*) and $p = 0.085$ (*dashed line*), respectively

delayed reinforcement, which we believe to be responsible for the threshold behaviour and the non-universal nature of the asymptotic approach to good generalization as in the simpler situation of the perceptron. The asymptotic approach to good generalization is generally somewhat slower than in the perceptron. We do not yet have a simple explanation for this fact. The fact that a committee machine has more adaptable parameters than a perceptron may be expected to lead to slower convergence but can by itself not fully explain the data. The more complicated phase flow is probably responsible for the fact that effects of finite size fluctuations in the simulations are stronger.

The general structure of the flow-equations for two-layer networks (with arbitrary hidden-layer-to-output function $\mathcal{F}$) is derived in Appendix B. The analytic evaluation is considerably more difficult, and ongoing; but a number of special results strengthen the findings from the numerical simulations. Further details about the investigation of this system will be reported elsewhere (Bergmann et al., in preparation).

5 Real world problems

We close with two *illustrations* to demonstrate that the learning algorithm proposed above and studied in the formal NN context can be applied to solve 'real world' problems.

5.1 Robot learning to avoid obstacles

The following simulation is intended to illustrate the above model for a 'realistic' problem: an agent moving on a board with obstacles must learn that it is good to reach the upper line and how to find its way there. The board is partitioned into a regular grid of squares and the agent takes one step at a time (up, down, left or right). It receives a positive or negative reward at the end of the journey, depending on the number of steps it took to reach the upper line (it starts at the middle of the bottom line). The 'cognitive structure' of the agent is realised as a network with 20 input neurons (storing the information *free/occupied* concerning the neighbouring cells for the last five steps) and a 'committee' of 4 neurons, each responsible for one of the 4 directions of move. The winner is chosen with probability $p_k = \mathrm{e}^{\beta h_k} / \sum_{k'=1}^{4} \mathrm{e}^{\beta h_{k'}}$, where $h_k = \sum_j J_{kj} \xi_j$ denotes the activation potential of the neuron representing direction k (the sum is over all 20 input neurons), and β is a parameter introducing a variable degree of randomness in deciding the direction. The larger β, the larger the probability p_k corresponding to the direction with the largest activation potential h_k; conversely, in the low β limit, the directions k are chosen with equal probability. The synapses (weights) from the input layer to the committee are modifiable. Learning proceeds along the lines described above: an immediate Hebbian adaptation of the weights after each step using (4), and a readjustment at the end of a path using the global information on the total number of steps, involving the usual learning parameters. Trying to run against an obstacle in direction k implies an immediate Hebb-penalty $\Delta J_{kj} = -\lambda \xi_j$ via (4), which will reduce the activation potential for this direction given the same path-history over the last five time-steps as encoded in the $\{\xi_j\}$. This problem is of the AE (average error) type above, with a journey representing a bag of decisions. It involves however a strong mutual dependence of the local decisions (since different moves may lead to different later situations).

Figure 5 illustrates simulation results for a robot moving on various rectangular (10×11) grids. The first plot in the left set of plots shows the grid, the boundary walls, the starting position on the bottom line and the goal—the upper line. The first three rows in the left set of plots also show realizations of paths taken by the robot in three different environments, 'Empty board' (first row), 'Right hand trap' (second row), and 'Open trap' (third row), respectively. Within each row the panels represent from left to right: first run, early performance, and two different realizations of late performance. The different environments were presented to the robot in three *consecutive* trials, i.e., the weights are not reinitialized before a new trial. In the simulations a maximal allowed number of steps of 100 is imposed (after which the run is stopped and maximal penalty assigned).

The fourth row represents two trials on pairs of configurations related by a mirror symmetry, 'Traps with clue' (left pair) and 'Traps without clue' (right pair). Here 120 runs are performed between switches within a pair, and the weights are not reinitialized after switches. The cutoff for the maximal number of steps is imposed at 60 for this experiment.

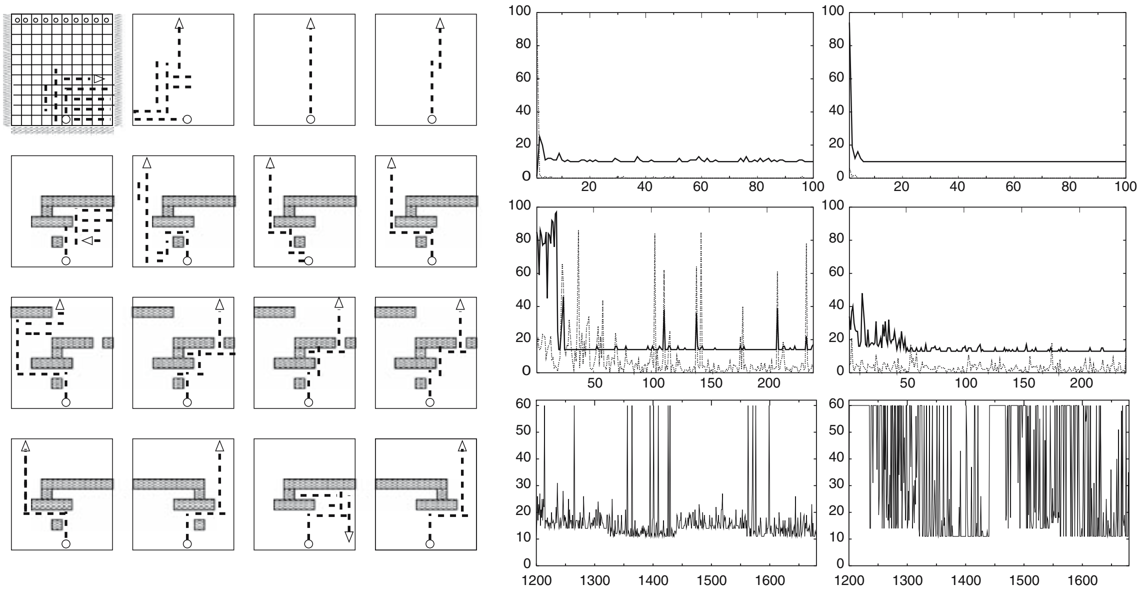

Fig. 5 Typical behaviour of a simulated robot moving on various 10 × 11 boards. See main text for explanations

The right set of plots shows the length of the path taken to the upper line of the board, plotted run by run for the trials shown in the left set of plots. First row: 'Empty board' for two settings of the randomness parameter β (small and large). The length of the best path is ten steps in this case. Second row: 'Right hand trap' (left) and 'Open trap' (right), corresponding to second and third row respectively in the left set of plots. The length of the best path is 14 and 13 steps, respectively. The lower dotted line on the plots indicates the number of steps lost running against obstacles (this has to be added to the length of the path to give the *total* number of steps).

The third row on the right shows the total number of steps on switching experiments for the symmetric pair of 'Traps with clue' (left) and the pair of 'Traps without clue' (right). The best paths have different lengths for the symmetry-related traps, viz. 13 and 11 steps for the left and right member of a trap-pair, respectively. As can be seen, the performance is much better if the situation provides a clue (left pair of traps), which means that the agent is able to 'identify' and use it.

Interesting features of the learning behaviour are:

- Good paths are found very fast, without needing any built-in strategies.
- Hierarchical learning is possible—new behavioural rules (the direction of move in a certain situation) are added without discharging old ones (if not contradictory).
- The randomness employed in choosing the direction of moves and which is controlled by β helps optimising the behaviour, or coping with changes in the situation.
- Stable behaviour is compatible with fluctuations.
- The agent 'identifies' goals, 'realises' the impenetrability of obstacles, and 'recognises' clues, such as the small obstacles in the switching experiment (four-th row in the left part of Fig. 5) which do not themselves obstruct the path, but are correlated with, and thus 'announce' the direction in which the larger obstacle encountered later on is open. These 'top-down' behavioural components are implemented 'bottom-up' using only the simple AR-Hebb rule.

There is a trade-off between the speed of learning and the stability of the behaviour on the one hand side, and flexibility in adapting to new situations on the other hand side, in which preferences can be set by varying β and the learning parameters. It turns out that there is a wide range of parameter settings that combine fast learning and stable behaviour with a reasonable degree of flexibility, so that no fine tuning is necessary. The model is described in more detail in (Stamatescu 2003).

5.2 Identifying a key stimulus

Consider the problem of identifying a certain kind of pattern which can appear (in some variants) as part of a sequence of different patterns—such as a certain kind of gene in a chromosome, say. We can only know that such a pattern is or is not there—e.g. from the expected expression of the gene in the phenotype. But there may be more gene variants with the same expression, together with those which do not match the phenotype and we must find out the class character of the interesting ones. A similar problem may be that of finding the scent signature of unpalatable components in food:

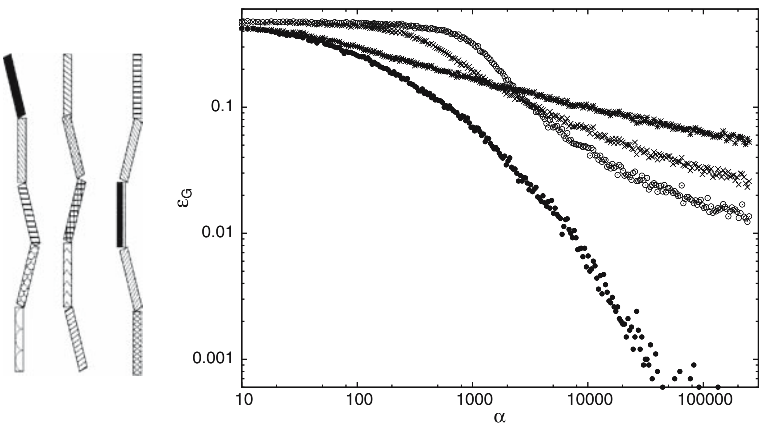

Fig. 6 Looking for strings of black signature: length of the strings $N = 20$, length of the string sequences $L = 5$. ϵ_G versus α for $Q(0) = 50,000$ (*empty circles*), 5,000 (‘x’-s) and 50 (*asterisks*), and λ near the corresponding thresholds (0.250, 0.420 and 0.560, respectively). The *black circles* represent the decay of the generalization error for the tuning $\lambda \propto \bar{e}_q$ ($\lambda(0) = 0.560$, $Q(0) = 50$)

The animal tries food in various combinations and can only judge about a combination as a whole. At the end of the day, say, after having eaten many herbs it will feel good or bad, without knowing why, but since the menu will slightly differ from day to day it may learn after some time to recognize the noxious kinds of herbs. Since both good and bad herbs may appear in various shapes or have different scents each time this amounts to a classification problem of the HI-type.

The variants of the pattern we are trying to identify thus define a class, say, +1. For definiteness we take sequences (bags) of 5 patterns, each pattern being itself a string of $N = 20$ bits of information. We assume that a ‘positive’ string may be contained at most once in each sequence (in a bag), and we allow about 10^4 variants of it (while there are about 10^6 patterns of class -1). The student is presented with sequences which contain a ‘positive’ string with 50% probability. It only knows (from the observation of the phenotype, from the effect of the food consumption) whether or not a class +1 string is contained in the bag. It therefore makes proposals on the basis of its current knowledge—the synapse strengths—and then updates the latter globally taking into account whether the prediction matches the observation for the bag. We therefore have a particular hidden instance problem with structured data (HI-SD).

In Fig. 6 we present simulation results for this system for various initial step sizes $1/\sqrt{Q(0)}$ and parameter λ chosen to be near the corresponding thresholds. We also demonstrate that it is possible to improve the algorithm by a simple tuning of λ with the running average $\bar{e}_q = \frac{1}{q}\sum_{q'=0}^{q} e_{q'}$. The beneficial effect of this tuning can in a certain sense be seen as reinforcing our claim that it is a *combination* of both mechanisms—the autonomous Hebbian association, and the synaptic response to evaluative feedback—which allows learning under the conditions of non-specific feedback to be successful. Tuning with $\bar{e}_q$ is an effective way to ensure that *the balance between both mechanisms* is maintained throughout the learning history, in particular in later stages with good generalization. Early and fast learning are thus easy to achieve: the error drops below 1% already for $\alpha < 10^4$.

6 Concluding remarks

To summarize, a simple learning algorithm based on local Hebb-type synaptic modifications can solve various non-specific reinforcement problems. Our motivation was not to design highly efficient algorithms for special AI or robotics problems, but to understand how simple mechanisms without appeal to involved strategies might be at work in non-trivial, unspecific reinforcement problems. The algorithm seems applicable to a broad range of different situations, which hints at a high level of generality. It is interesting that such a simple algorithm can cope with complex learning problems, and this feature makes it a candidate for a basic mechanism in learning. As a model for biological developments it indicates that feedback non-specificity can be dealt with at the elementary neuronal level by mechanisms which are simple enough to have plausibly developed during the early stages of evolution. A very peculiar aspect is the essential role of *both*, local Hebb potentiation and global correction via replay.

Note that the correction in phase II of the algorithm respects an important information-theoretic symmetry: the synaptic correction in response to an indiscriminate feedback is indiscriminate *in the same sense*. Any deviation from this symmetry would entail that the learning mechanism creates a hypothesis about its environment for which the environmental feedback does not provide any evidence. This is in fact in fairly close analogy with Bernoulli’s principle of insufficient reason according to which the best assumption in an information-theoretic sense about a random variable of which we have no knowledge whatsoever apart from the range of values it can take, is to assume that its distribution is *uniform* over the range of possible values.

Some features of the learning behaviour described by this model may also show up in more complex non-specific feedback situations. In a behavioural setting, for instance, we find commitment to one’s own experiences and global, critical consideration of the results as necessary prerequisites for dealing with non-specific feedback, with λ playing the role of a ‘tolerance’ parameter. In the interaction between these two factors the randomness of experiences is shaped into a knowledge landscape. A most interesting feature thereby is the increase of the learning efficiency with decreasing tolerance, on the one hand, together with the complete disruption of the learning behaviour if the tolerance is reduced below a certain threshold, on the other hand.

Behavioural suggestions aside, in our view the first merit of our model is to provide an elementary mechanism for learning from non-trivial experiences which may be relevant in an evolutionary perspective.

Acknowledgments We thank Urs Bergamann for helpful discussions, and in particular for alerting us to an error in an earlier version of Eq. (33).

Appendix

A Integrals and asymptotic behaviour

Below we collect the integrals appearing in the evaluation of (19) and (20) for the AE and the HI versions of the online bag error. Independence of patterns together with Eq. (16) imply that the joint probability density of the fields $\{x^{(q,l)}\}$ and $\{y^{(q,l)}\}$ is

$$P(\{x^{(q,l)}, y^{(q,l)}\}) = \prod_{l=1}^{L} \frac{1}{2\pi\sqrt{Q-R^2}} \exp\left[-\frac{\left(\left(x^{(q,l)}\right)^2+Q\left(y^{(q,l)}\right)^2-2Rx^{(q,l)}y^{(q,l)}\right)}{2(Q-R^2)}\right] \tag{24}$$

with $R = R(\alpha)$ and $Q = Q(\alpha)$, and marginals

$$P(x^{(q,l)}) = \frac{1}{\sqrt{2\pi Q}} \exp\left(-\frac{\left(x^{(q,l)}\right)^2}{2Q}\right), \quad P(y^{(q,l)}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(y^{(q,l)}\right)^2}{2}\right). \tag{25}$$

Writing the the AE bag error (22) as

$$e_q = e_q^{AE} = \frac{1}{2L}\sum_{l=1}^{L}\left(1-\text{sgn}\left(x^{(q,l)}\right)\text{sgn}\left(y^{(q,l)}\right)\right) = \frac{1}{2} - \frac{1}{2L}\sum_{l=1}^{L} s^{(q,l)}t^{(q,l)} \tag{26}$$

one can express (19) and (20) as

$$\frac{\mathrm{d}R}{\mathrm{d}\alpha} = \frac{\lambda-\frac{1}{2}}{L}\sum_{l=1}^{L}\left\langle s^{(q,l)}y^{(q,l)}\right\rangle + \frac{1}{2L^2}\sum_{l,l'=1}^{L}\left\langle s^{(q,l')}t^{(q,l')}s^{(q,l)}y^{(q,l)}\right\rangle \tag{27}$$

$$\frac{\mathrm{d}Q}{\mathrm{d}\alpha} = \frac{2\lambda-1}{L}\sum_{l=1}^{L}\left\langle \left|x^{(q,l)}\right|\right\rangle + \frac{1}{L^2}\sum_{l,l'=1}^{L}\left\langle s^{(q,l')}t^{(q,l')}\left|x^{(q,l)}\right|\right\rangle + \left(\lambda-\frac{1}{2}\right)^2 + \frac{\lambda-\frac{1}{2}}{L}\sum_{l=1}^{L}\left\langle s^{(q,l)}t^{(q,l)}\right\rangle + \frac{1}{4L^2}\sum_{l,l'=1}^{L}\left\langle s^{(q,l)}t^{(q,l)}s^{(q,l')}t^{(q,l')}\right\rangle. \tag{28}$$

In (27) and (28), averages involving only a single pattern (q, l) are independent of l, whereas averages involving two patterns (q, l) and (q, l') will have different values for $l = l'$ and $l \neq l'$; in the second case, independence of the fields for different l can be used to factor averages. The remaining averages needed to evaluate (27) and (28) can be expressed in terms of the following integrals,

$$\left\langle s^{(q,l)}y^{(q,l)}\right\rangle = \int \mathrm{d}x\mathrm{d}y\, P(x,y)\ \text{sgn}\,(x)y = \sqrt{\frac{2}{\pi}}\,\frac{R}{\sqrt{Q}},$$

$$\left\langle t^{(q,l)}x^{(q,l)}\right\rangle = \int \mathrm{d}x\mathrm{d}y\, P(x,y)\ \text{sgn}\,(y)x = \sqrt{\frac{2}{\pi}}\,R,$$

$$\left\langle s^{(q,l)}t^{(q,l)}\right\rangle = \int \mathrm{d}x\mathrm{d}y\, P(x,y)\ \text{sgn}\,(x)\text{sgn}\,(y) = 1-\frac{2}{\pi}\arccos\left(\frac{R}{\sqrt{Q}}\right) \equiv P,$$

$$\left\langle\left|x^{(q,l)}\right|\right\rangle = \int \mathrm{d}x\, P(x)\ |x| = \sqrt{\frac{2}{\pi}}\,\sqrt{Q},$$

$$\left\langle\left|y^{(q,l)}\right|\right\rangle = \int \mathrm{d}y\, P(y)\ |y| = \sqrt{\frac{2}{\pi}},$$

resulting in

$$\frac{\mathrm{d}R}{\mathrm{d}\alpha} = \left(\lambda-\frac{1}{2}\right)\sqrt{\frac{2}{\pi}}\,\frac{R}{\sqrt{Q}} + \frac{1}{2L}\sqrt{\frac{2}{\pi}} + \frac{L-1}{2L}\sqrt{\frac{2}{\pi}}\,\frac{R}{\sqrt{Q}}\,P \tag{29}$$

$$\frac{\mathrm{d}Q}{\mathrm{d}\alpha} = (2\lambda-1)\sqrt{\frac{2}{\pi}}\,\sqrt{Q} + \frac{1}{L}\sqrt{\frac{2}{\pi}}\,R + \frac{L-1}{L}\sqrt{\frac{2}{\pi}}\sqrt{Q}\,P + \left(\lambda-\frac{1}{2}\right)^2 + \left(\lambda-\frac{1}{2}\right)P + \frac{1}{4L} + \frac{L-1}{4L}P^2. \tag{30}$$

If the HI bag error (23)

$$e_q = e_q^{HI} = \frac{1}{4L^2}\sum_{l,l'=1}^{L}\left(s^{(q,l)}s^{(q,l')} + t^{(q,l)}t^{(q,l')} - 2s^{(q,l)}t^{(q,l')}\right) \tag{31}$$

is used instead, the algebra is a bit more involved but the results can still be expressed in terms of the integrals listed above:

$$\frac{\mathrm{d}R}{\mathrm{d}\alpha} = \left(\lambda - \frac{1}{2L}\right)\sqrt{\frac{2}{\pi}}\,\frac{R}{\sqrt{Q}} + \frac{1}{2L^2}\sqrt{\frac{2}{\pi}} + \frac{L-1}{2L^2}\sqrt{\frac{2}{\pi}}\,\frac{R}{\sqrt{Q}}\,P \tag{32}$$

$$\frac{\mathrm{d}Q}{\mathrm{d}\alpha} = \left(2\lambda - \frac{1}{L}\right)\sqrt{\frac{2}{\pi}}\,\sqrt{Q} + \frac{1}{L^2}\sqrt{\frac{2}{\pi}}\,R + \frac{L-1}{L^2}\sqrt{\frac{2}{\pi}}\sqrt{Q}\,P + \lambda^2 - \frac{\lambda}{L}(1-P) + \frac{1}{4L^3}\Big((3L-1)+(6L-4)P+3(L-1)P^2\Big). \tag{33}$$

With $\epsilon_G = \frac{1}{\pi}\arccos\left(\frac{R}{\sqrt{Q}}\right) = \frac{1}{2}(1-P)$ the evolution equations lead for large α to the asymptotic behaviour:

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

$$Q \simeq \frac{2}{\pi}\lambda^2\alpha^2 \tag{35}$$

for the AE problem, and

$$\epsilon_G^2 \simeq \frac{1}{2\pi(\frac{1}{\lambda L^2}-1)}\alpha^{-1} + c\alpha^{-\frac{1}{\lambda L^2}} \quad \text{for } \lambda \neq \frac{1}{L^2}, \tag{36}$$

$$Q \simeq \frac{2}{\pi}\lambda^2\alpha^2 \tag{37}$$

for the HI problem (for $\lambda = 1/L$, respectively, $\lambda = 1/L^2$ there are logarithmic corrections).

B Theory for two-layer networks

In this appendix we briefly describe the theory for two-layer networks. We assume a two layer structure with N input-nodes and K hidden nodes for both, student, and teacher.

The hidden node outputs of student and teacher are

$$s_k^\mu = \mathrm{sgn}\,(x_k^\mu), \quad t_k^\mu = \mathrm{sgn}\,(y_k^\mu), \; k = 1, \ldots, K, \tag{38}$$

with fields

$$x_k^\mu = \frac{1}{\sqrt{N}}\sum_i J_{ki}\xi_i^\mu, \quad y_k^\mu = \frac{1}{\sqrt{N}}\sum_i B_{ki}\xi_i^\mu, \; k=1,\ldots,K. \tag{39}$$

and J_{ki} and B_{ki} denoting synaptic weights of the hidden student and teacher nodes, respectively. The overall outputs of student and teacher are given by a fixed function $\mathcal{F}$ of the hidden node values,

$$s^\mu = \mathcal{F}(\{s_k^\mu\}), \quad t^\mu = \mathcal{F}(\{t_k^\mu\}). \tag{40}$$

The function $\mathcal{F}$ is taken to be the same for student and teacher, although a situation with *different* functions may also be contemplated. Prominent examples are the committee machine

$$\mathcal{F}(\{s_k^\mu\}) = \mathcal{F}_c(\{s_k^\mu\}) = \mathrm{sgn}\left(\frac{1}{\sqrt{K}}\sum_{k=1}^{K} s_k^\mu\right), \tag{41}$$

or its graded version, treated in Sect. 4 of the present paper, with

$$\mathcal{F}(\{s_k^\mu\}) = \mathcal{F}_g(\{s_k^\mu\}) = \frac{1}{\sqrt{K}}\sum_{k=1}^{K} s_k^\mu, \tag{42}$$

or the so-called parity machine for which

$$\mathcal{F}(\{s_k^\mu\}) = \mathcal{F}_p(\{s_k^\mu\}) = \prod_{k=1}^{K} s_k^\mu. \tag{43}$$

The generalization of the AR-Hebb rule for this set-up would be to use Hebbian adaptation on the basis of the momentary values of the students weights in phase I,

$$\mathrm{I}: \; J_{ki}^{(q,l+1)} = J_{ki}^{(q,l)} + \frac{\lambda}{\sqrt{N}} s_k^{(q,l)}\xi_i^{(q,l)}, \quad l = 1, \ldots, L, \tag{44}$$

and unlearning based on a global empirical bag-output error in phase II,

$$\mathrm{II}: \quad J_{ki}^{(q+1,1)} = J_{ki}^{(q,L+1)} - \frac{e_q}{\sqrt{N}}\sum_{l=1}^{L} s_k^{(q,l)}\xi_i^{(q,l)}\,, \tag{45}$$

with an online bag-error of the form

$$e_q = \frac{1}{2L}\sum_{l=1}^{L}\Big(\mathcal{F}(\{s_k^{(q,l)}\}) - \mathcal{F}(\{t_k^{(q,l)}\})\Big)^2 \tag{46}$$

The macroscopic analysis of the learning dynamics is analogous to that of the single layer systems. The main observation required is that the fields $x_k^{(q,l)}$ and $y_k^{(q,l)}$ are jointly Gaussian with zero-mean, uncorrelated for different l, and non-zero correlations given by

$$\left\langle y_k^{(q,l)} y_{k'}^{(q,l)}\right\rangle = P_{kk'}\,, \quad \left\langle x_k^{(q,l)} x_{k'}^{(q,l)}\right\rangle = Q_{kk'}(\alpha), \quad \left\langle k_k^{(q,l)} y_{k'}^{(q,l)}\right\rangle = R_{kk'}(\alpha) \tag{47}$$

with continuous time $\alpha = qL/N$ as before. Here

$$P_{kk'} = \frac{1}{N}\sum_i B_{ki}B_{k'i}, \quad Q_{kk'}(\alpha) = \frac{1}{N}\sum_i J_{ki}^{(q,l)} J_{k'i}^{(q,l)}, \quad R_{kk'}(\alpha) = \frac{1}{N}\sum_i J_{ki}^{(q,l)} B_{k'i}\,. \tag{48}$$

Note that the $P_{kk'}$ are given in terms of the fixed teacher weights alone, while the $Q_{kk'}$ and the $R_{kk'}$ evolve dynamically.

Following the reasoning described for single layer systems, we then get the following set of coupled flow equations for the order-parameters of the system

$$\frac{\mathrm{d}R_{kk'}}{\mathrm{d}\alpha} = \left\langle \frac{g_q}{L} \sum_{l=1}^{L} \operatorname{sgn}\left(x_k^{(q,l)}\right) y_{k'}^{(q,l)} \right\rangle, \tag{49}$$

$$\begin{aligned}\frac{\mathrm{d}Q_{kk'}}{\mathrm{d}\alpha} &= \left\langle \frac{g_q}{L} \sum_{l=1}^{L} \Big[\operatorname{sgn}\left(x_k^{(q,l)}\right) x_{k'}^{(q,l)} \right. && (50)\\ &\quad \left. + \operatorname{sgn}\left(x_{k'}^{(q,l)}\right) x_k^{(q,l)}\Big]\right\rangle \\ &\quad + \left\langle \frac{g_q^2}{L} \sum_{l=1}^{L} \operatorname{sgn}\left(x_k^{(q,l)}\right) \operatorname{sgn}\left(x_{k'}^{(q,l)}\right) \right\rangle. && (51)\end{aligned}$$

with $g_q = \lambda - e_q$ as before. The expectations on the r.h.s of these equations are evaluated using the joint Gaussian densities of the $\{x_k^{(q,l)}\}$ and the $\{y_k^{(q,l)}\}$ which can be evaluated in terms of the fixed $P_{kk'}$, the $R_{kk'}(\alpha)$ and the $Q_{kk'}(\alpha)$ alone, although, depending on $\mathcal{F}$, the details can be quite involved. However, it is clear that (49) and (51) form an autonomous closed set of flow equations which can be solved numerically to study learning dynamics in the large system limit. A feasible way out in situations where the integrals can no longer be expressed in closed form is to evaluate them by stochastic sampling over Gaussian fields $\{x_k^{(q,l)}\}$ and $\{y_k^{(q,l)}\}$, with correlations given by (47), which are easily generated using a Cholesky-decomposition of the correlation matrix (Press et al. 2002). This amounts in a certain sense to simulations of the infinite system, and is close in spirit to (though simpler than) the Eissfeller-Opper algorithm for spin-glass dynamics (Eissfeller and Opper 1992). Further details are reported in (Bergmann et al., in preparation).

References

Biehl M, Riegler P (1994) Online learning with a perceptron. Europhys Lett 28:525–530

Biehl M, Kühn R, Stamatescu I-O (2000) Learning structured data from unspecific reinforcement. J Phys A Math Gen 33:6843–6857

Byrne R (1999) The thinking ape. Oxford University Press, Oxford

Crick F, Mitchison G (1983) The function of dream sleep. Nature 304:111–114

Cristiani N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines. Cambridge University Press, Cambridge

Eissfeller H, Opper M (1992) New method for studying the dynamics of disordered spin systems without finite-size effects. Phys Rev Lett 68:2094–2097

Foster DJ, Wilson MA (2006) Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature 440: 680–683 (we thank our colleague U. Bergmann for directing our attention to this paper)

Hertz J, Krogh A, Palmer RG (1991) Introduction to the theory of neural computation. Addison-Wesley, Reading

Hopfield JJ, Feinstein DI, Palmer RG (1983) Unlearning has a stabilizing effect in collective memories. Nature 304:158–159

Kinzel W, Rujan P (1990) Improving a network generalization ability by selecting examples. Europhys Lett 13:473–477

Kühn R, Stamatescu I-O (1999) A two step algorithm for learning from unspecific reinforcement. J Phys A Math Gen 32:5749–5762

Kühn R et al (eds) (2003) Adaptivity and learning – an interdisciplinary debate. Springer, Heidelberg

Menzel R (2003) Creating presence by bridging between the past and the future: the role of learning and memory for the organization of life, in (Kühn et al. 2003), pp 59–70

Mitchell TM (1997) Machine learning. Mc Graw Hill, New York

Mlodinov L, Stamatescu I-O (1985) An evolutionary procedure for machine learning. Int J Comp Inf Sci 14:201–219

Opper M, Kinzel W, Kleinz J, Nehl R (1990) On the ability of the optimal perceptron to generalize. J Phys A 23:L581–L586

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2002) Numerical recipies in C: the art of scientific computing, 2nd edn. Cambridge University Press, Cambridge

Saad D, Solla SA (1995) Exact solution for online learning in multilayer neural networks. Phys Rev Lett 74:4337–4340

Schölkopf B, Smola A (2002) Learning with Kernels. MIT Press, Cambridge

Shawe-Taylor J, Cristiani N (2004) Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge

Stamatescu I-O (2003) A simple model for learning from nonspecific reinforcement. In (Kühn et al. 2003), pp 265–280

Sutton RS, Barto AG (2000) Reinforcement learning—an Introduction. MIT Press, Cambridge

Vallet F (1989) The Hebb rule for learning separable Boolen functions: learning and generalization. Europhys Lett 8:747–751

van Hemmen JL (1997) Hebbian learning, its correlation catastrophe, and unlearning. Network 8:V1–V17

van Hemmen JL, Ioffe LB, Kühn R, Vaas M (1990) Increasing the efficiency of a neural network through unlearning. Physica A163: 386–392

Vapnik VN (1995) The Nature of Statistical Learning Theory. Springer, Berlin

Vapnik VN (1998) Statistical learning theory. Wiley Inc, New York

Wyatt J (2003) Reinforcement learning: a brief overview, in (Kühn et al. 2003), pp 243–264