

Learning with incomplete information in the Committee Machine

Urs M. Bergmann, Reimer Kühn and Ion-Olimpiu Stamatescu

October 12, 2009

Abstract

We study the problem of learning with incomplete information in a student-teacher setup for the committee machine. The learning algorithm combines unsupervised Hebbian learning of a series of associations with a delayed reinforcement step, in which the set of previously learnt associations is partly and indiscriminately unlearned, to an extent that depends on the success rate of the student on these previously learnt associations. The relevant learning parameter λ represents the strength of Hebbian learning. A coarse-grained analysis of the system yields a set of differential equations for overlaps of student and teacher weight vectors, whose solutions provide a complete description of the learning behavior. It reveals complicated dynamics showing that perfect generalization can be obtained if the learning parameter exceeds a threshold λ_c , and if the initial value of the overlap between student and teacher weights is non-zero. In case of convergence, the generalization error exhibits a power law decay as a function of the number of examples used in training, with an exponent that depends on the parameter λ . An investigation of the system flow in a subspace with broken permutation symmetry between hidden units reveals a bifurcation point λ^* above which perfect generalization does not depend on initial conditions. Finally, we demonstrate that cases of a complexity mismatch between student and teacher are optimally resolved in the sense that an over-complex student can emulate a less complex teacher rule, while an under-complex student reaches a state which realizes the minimal generalization error compatible with the complexity mismatch.

1 Introduction

Learning agents, be they natural or artificial, can be regarded as systems endowed with a cognitive apparatus which allows them to construct representations of their environment. Learning is an adaptive process in which the cognitive structure of an agent is improved to increase its ability to reach its goals, or simply to survive if the environment is competitive.

Reinforcement learning theory has been constructed as a formal language to describe learning by agents through interactions with their environment. In this context, the environmental feedback typically consists of an evaluative stimulus (a reward) obtained in response to an action of the learning agent [1, 2]. Reinforcement learning, i.e. learning via rewards which quantify the success of

specific actions, is to be contrasted with supervised learning, where the environmental feedback is instructive rather than evaluative, and consists in providing correct target outputs for the learning agent. Some neurobiological evidence supporting reinforcement learning appears to exist: signals that share characteristic features with the so-called Temporal Difference Error, the difference between the reward obtained and the expected reward that features in some of the reinforcement learning algorithms, have been found in the human brain [3, 4].

Yet there is an open issue. In general, a learning agent needs to perform a *sequence* of actions before it receives a *delayed* global reward. Therefore, it is unclear which parts of the sequence of actions were conducive and which detrimental to the actual reward received. Hence, it remains unclear in which way the cognitive apparatus ought be adapted to improve the agent’s probability of success in the future. The information provided by the reward is insufficient to decide this question, and it is in this sense *incomplete*. In what follows we will refer to the problem of learning under such conditions as “learning with incomplete information”. This problem was already recognized in the early days of learning theory and was named the (temporal) ‘credit assignment problem’ [5]. Sophisticated methods, like Dynamic Programming, Monte Carlo Methods and TD-Learning with eligibility traces, have been developed to solve this problem efficiently [1]. In nature, however, solutions of this type would require the existence of complex mechanisms already in place *before* learning can actually occur. If reinforcement learning is a fundamental learning principle underlying the behavior of most (if not all) species, it seems reasonable to postulate the existence of an elementary learning mechanism that is phylogenetically older than and thus precedes the more sophisticated methods mentioned above. Alternatively, it could constitute a basis for the emergence of more sophisticated mechanisms in the course of evolution. Either way, it would be of interest to identify this elementary mechanism.

A simple approach to the credit assignment problem is to assume non-specificity of the reward: the reward is global, is calculated from a sequence of actions, and is non-specific with regard to individual actions in the sequence [6]. For a single-layer neural network, the credit assignment problem can be solved using non-specific feedback based on an algorithm which is elementary enough to have conceivably developed under natural conditions [7, 8, 9]. The learning model proposed here consists of a simple combination of two “natural” learning steps: (i) unsupervised Hebbian learning of a series of associations combined with (ii) a reinforcement step consisting of a reward-dependent non-specific Hebbian unlearning of these associations. Referring to the two steps involved, this learning scenario has been called “Association Reinforcement” (AR) learning.

However, it is a well known fact that single-layer networks can only represent linearly separable functions [10], and are thus too restricted for many real-world applications. On the other hand, two-layer feed-forward networks of neurons with sigmoidal transfer functions have been shown to constitute universal approximators for continuous functions from \mathbb{R}^N to \mathbb{R}^M [11]. A feed-forward network consisting of a layer of N input units, a hidden layer with K units, and a single output unit implements a function that maps $\boldsymbol{\xi} \in \mathbb{R}^N$ to $s \in \mathbb{R}$ via

$$s = f \left(\frac{1}{\sqrt{K}} \sum_{k=1}^K w_k g \left(\frac{\mathbf{J}_k \cdot \boldsymbol{\xi}}{\sqrt{N}} \right) \right), \quad (1)$$

where the \mathbf{J}_k are adjustable input weight vectors of the K hidden units, and the w_k are the K adjustable hidden-to-output weights. In such systems, the transfer functions f and g are usually taken to have sigmoidal form.

In the present paper we investigate AR learning for a simple variant of a two-layer network, the committee machine, for which the hidden-to-output weights are identical and fixed at $w_k = 1$. We will choose $g(x) = \text{sgn}(x)$ and focus on the the so-called soft-committee machine [12] with $f(x) = x$. Conventional online learning in the soft-committee machine with a differentiable transfer function has been studied [13] and the convergence to perfect generalization in this framework has been proven. First results about AR learning in this system, using simulations that indicated convergence to good generalization, were reported in [14, 9]. In order to further understand AR learning for this system we will investigate it analytically using standard online learning methods to describe the learning dynamics in terms of ODEs for overlaps between the weight vectors \mathbf{J}_k and analogous weight vectors \mathbf{B}_k of a similarly structured teacher network implementing a given rule. This allows for systematic investigation of learning with incomplete information in the soft-committee machine, as described above, thereby deepening our understanding of AR learning.

Although we cannot demonstrate that AR learning is used in nature, we wish to point out that neurophysiological evidence supporting its main ingredients exists. First, replay of spike sequences has been found in various brain areas and animals. For example, they have been found in the hippocampus of rats, which recall the way through a labyrinth in reverse order immediately after spatial experience [15] and in time-compressed forward order during sleep [16]. Similar observations have been made in the rat prefrontal cortex [17]. Very recently, it has been shown that the activation of rat hippocampal assembly sequences correlates with memory and can be used for behavioral prediction [18]. In humans, hippocampal cell firing has been linked to episodic recall [19]. The observed spike sequences in these experiments might reflect the recall necessary for the reinforcement step of the proposed algorithm. Second, it has recently been found that the brain has mechanisms to reverse the polarity of plasticity processes [20]. This ability is required for the unspecific reinforcement learning step of AR learning, which has a polarity opposite to that of the unsupervised Hebbian learning step during the first phase of AR learning.

This paper is organized as follows. Section 2 explains the problem of incomplete information feedback for the committee machine. Section 2.1 presents the Association Reinforcement (AR) learning model for this system, and Section 2.2 introduces a statistical physics inspired description of AR learning for the committee machine based on the dynamical evolution of overlaps between pairs of student-student and student-teacher weight vectors [7, 8, 14, 9]. In Section 3 we explain our approach to solving the differential equations that describe the online learning dynamics. In Section 4 we present results obtained from both direct simulations and numerical computations of these differential equations. Finally, Section 5 provides a summary of the results and concluding remarks.

2 The Incomplete Information Problem for the Committee Machine

As described on the previous pages, “incomplete information” here means that the reward for a sequence of actions does not contain information concerning *each* particular action in the sequence. Here we will study AR learning for the committee machine using a teacher student scenario. That is, the student network defined by Eq. (1) will be trained to implement a function that maps $\xi \in \mathbb{R}^N$ to $t \in \mathbb{R}$, defined by a teacher network of a similar structure, viz.

$$t = \frac{1}{\sqrt{M}} \sum_{m=1}^M \operatorname{sgn} \left(\frac{\mathbf{B}_m \cdot \xi}{\sqrt{N}} \right). \quad (2)$$

Here we have already specialized to the sign function for the transfer function of the hidden units, to uniform hidden-to-output weights characteristic of committee machines, and to $f(x) = x$ for the transfer function of the output unit.

The task of the student network, Eq. (1), is to adapt its internal configuration, i.e. the weight vectors \mathbf{J}_k , to improve agreement with the output of the teacher network (2). Note that we allow for *different* relative complexities of the teacher and the student network: for $K > M$ the student is of higher complexity than the teacher, and may therefore implement the rule defined by the teacher in several ways. In the case $K = M$ both are of equal complexity, whereas for $K < M$ the student is of lower complexity than the teacher, and the rule defined by the teacher will in general not be perfectly realizable by the student.

2.1 AR Learning for the Committee Machine

In associative reinforcement learning, the student network processes whole series (‘bags’) of patterns $\{\xi^{(q,l)}; l = 1, \dots, L\}$, where $q \in \mathbb{N}$ enumerates the bags, each containing L input patterns, labeled by $l = 1, \dots, L$. In the association phase it modifies its internal weights by simple Hebbian adaptation, pairing pre- and post-synaptic activities of its hidden units using their *own classification* as computed on the basis of the *momentary values* of the synaptic weights $\mathbf{J}_k^{(q,l)} = (J_{ki}^{(q,l)})$,

$$(i): \quad J_{ki}^{(q,l+1)} = J_{ki}^{(q,l)} + \frac{\lambda}{\sqrt{N}} \operatorname{sgn} \left(x_k^{(q,l)} \right) \xi_i^{(q,l)}, \quad l = 1, \dots, L, \quad (3)$$

where we have introduced

$$x_k^{(q,l)} = \frac{\mathbf{J}_k^{(q,l)} \cdot \xi^{(q,l)}}{\sqrt{N}}, \quad (4)$$

to denote the local field of the k -th hidden unit for the input pattern $\xi^{(q,l)}$, and where λ denotes a learning rate characteristic of the association phase.

After the series of Hebbian adaptations, Eq. (3), in response to the L patterns of bag q , an “unspecific” reinforcement learning step is performed. It consists in partially, and indiscriminately *undoing* the synaptic adaptations of phase (i),

to an extent that depends on an online bag-error e_q which is a measure of the success rate of the students on the patterns in bag q ,

$$(ii) : \quad J_{ki}^{(q+1,1)} = J_{ki}^{(q,L+1)} - \frac{e_q}{\sqrt{N}} \sum_{l=1}^L \text{sgn} \left(x_k^{(q,l)} \right) \xi_i^{(q,l)} . \quad (5)$$

Note that without loss of generality we have chosen a unit learning rate for the reinforcement step in Eq. (5), such that the only parameter of the model is λ , Eq. (3).

Several possibilities exist for the error signal e_q received by the network after having completed phase (i). In the present paper we will use the so-called hidden instance (HI) error [9]

$$e_q = \frac{1}{4L^2} \left(\sum_{l=1}^L (t^{(q,l)} - s^{(q,l)}) \right)^2 , \quad (6)$$

in which $t^{(q,l)}$ and $s^{(q,l)}$ are the teacher and student-committee classifications of pattern $\xi^{(q,l)}$. The normalization in Eq. (6) leads to $0 \leq e_q \leq 1$ for binary outputs.

The present setup implements learning which deals with incomplete information on *three different* levels. First, there is incompleteness of information related to the existence of a hidden layer. Learning happens in response to the *overall* output of the committee machine, which is non-specific w.r.t. the contribution of individual hidden units to that output. Second, there is incompleteness of information due to the reinforcement signal being based on the performance of the student network on a bag containing *several* patterns, a signal which is non-specific w.r.t. the contribution of individual patterns to that error. Third, finally, the HI error measure e_q itself contains an element of incompleteness, as errors made for different patterns l in a bag can cancel. Instances of actual misclassifications remain hidden or unidentified with this error measure.

The first form of non-specificity of feedback is present already in standard supervised learning scenarios for the committee machine. The second form is a central feature of the AR learning paradigm, according to which the reinforcement signal is global for a sequence of patterns and non-specific concerning individual contributions of patterns in this sequence. The third form of non-specificity finally is a property of the HI error measure. It is used here to pose what might be called a deliberate additional challenge for the AR learning paradigm. Indeed, by demonstrating that AR learning based on the HI error signal can be successful, one can safely assume that AR learning using more informative error measures (such as the average bag error $e_q = \frac{1}{2L} \sum_{l=1}^L |t^{(q,l)} - s^{(q,l)}|$ used previously for single-layer networks [7, 9]) will be successful as well.

The unsupervised learning of phase (i) amounts to self-strengthening of the hidden units own experience *without feedback from the environment*. The reinforcement phase (ii) uses an error signal which is non-specific concerning individual patterns within a bag, and implements a Hebbian unlearning step that is non-specific *in the same way*.

The following section is devoted to deriving flow equations for a set of overlaps between student and teacher weight vectors following the standard approach to online learning [12, 13, 21]; these equations provide a comprehensive

description of the learning dynamics for this system in the thermodynamic limit $N \rightarrow \infty$.

2.2 Coarse-Grained Analysis

To derive a closed set of flow equations for the order parameters of the system [12, 13, 21], we follow [7, 9], and combine the two phases Eqs. (3) and (5) of a session of length L into a single coarse-grained step,

$$J_{ki}^{(q+1,1)} = J_{ki}^{(q,1)} + \frac{\lambda - e_q}{\sqrt{N}} \sum_{l=1}^L \text{sgn} \left(x_k^{(q,l)} \right) \xi_i^{(q,l)}. \quad (7)$$

In the following we assume the components of the input patterns to be independently identically distributed random variables with zero mean and unit variance, i.e. $\langle \xi_i^{(q,l)} \xi_j^{(q',l')} \rangle = \delta_{ij} \delta_{qq'} \delta_{ll'}$. As explained after Eq. (14) below, it then follows that in the thermodynamic limit $N \rightarrow \infty$ the learning dynamics is fully described by the evolution of a set of overlaps between student and teacher weight vectors, which we define as

$$Q_{kk'}^{(q,1)} = \frac{1}{N} \sum_{i=1}^N J_{ki}^{(q,1)} J_{k'i}^{(q,1)}, \quad (8)$$

$$R_{km}^{(q,1)} = \frac{1}{N} \sum_{i=1}^N J_{ki}^{(q,1)} B_{mi}. \quad (9)$$

The corresponding set of teacher overlaps

$$T_{mn} = \frac{1}{N} \sum_{i=1}^N B_{mi} B_{ni} \quad (10)$$

characterizes the configuration of the teacher committee that defines the rule, and is time-independent.

From Eqs. (7) – (9) one obtains

$$R_{km}^{(q+1,1)} = R_{km}^{(q,1)} + \frac{(\lambda - e_q)}{N} \sum_l^L \text{sgn} \left(x_k^{(q,l)} \right) y_m^{(q,l)}, \quad (11)$$

$$Q_{kk'}^{(q+1,1)} = Q_{kk'}^{(q,1)} + \frac{(\lambda - e_q)^2}{N} \sum_l^L \text{sgn}(x_k^{(q,l)}) \text{sgn} \left(x_{k'}^{(q,l)} \right) + \frac{\lambda - e_q}{N} \sum_l^L \left[\text{sgn}(x_k^{(q,l)}) x_{k'}^{(q,l)} + \text{sgn} \left(x_{k'}^{(q,l)} \right) x_k^{(q,l)} \right], \quad (12)$$

where, in analogy to Eq. (4), we have introduced

$$y_m^{(q,l)} = \frac{\mathbf{B}_m \cdot \boldsymbol{\xi}^{(q,l)}}{\sqrt{N}}, \quad (13)$$

to denote the local field of the m -th hidden unit of the teacher committee machine. By the Central Limit Theorem the local fields $\{x_k^{(q,l)}\}$ and $\{y_m^{(q,l)}\}$ are

jointly Gaussian with covariances which depend only on the overlaps in Eqs. (8) – (10), viz.

$$\begin{aligned}
\langle x_k^{(q,l)} x_{k'}^{(q,l')} \rangle &= \delta_{ll'} Q_{kk'}^{(q,1)} \\
\langle x_k^{(q,l)} y_m^{(q,l')} \rangle &= \delta_{ll'} R_{km}^{(q,1)} \\
\langle y_m^{(q,l)} y_n^{(q,l')} \rangle &= \delta_{ll'} T_{mn}
\end{aligned} \tag{14}$$

Note that we have neglected terms $\mathcal{O}(N^{-1})$ in Eq. (14), allowing us to restrict the discussion to the $l = l' = 1$ terms and thus to use $Q_{kk'}^{(q,1)}$ and $R_{km}^{(q,1)}$ instead of $Q_{kk'}^{(q,l)}$ and $R_{km}^{(q,l)}$, respectively. This can be seen by inserting the Hebbian weight update rule of Eq. (3) into the definition for the student field in Eq. (4) and performing the averages in Eqs. (14).

To arrive at a set of flow equations for the order parameters Q and R , we follow standard reasoning to analyze online learning [12, 13], and introduce a continuous time $\alpha = qL/N$, with $\Delta\alpha = L/N \rightarrow 0$ in the thermodynamic limit $N \rightarrow \infty$. The order parameters $R_{km}^{(q,1)} \equiv R_{km}(\alpha)$ and $Q_{kk'}^{(q,1)} \equiv Q_{kk'}(\alpha)$ then become smooth functions on the α -scale. Combining S steps of the form of Eq. (7) with $S \rightarrow \infty$ in such a way that $S/N \rightarrow 0$ in the thermodynamic limit, one obtains an autonomous set of flow equations which, by appeal to the law of large numbers, can be formulated in terms of averages over the jointly Gaussian fields, with covariances that become functions of α themselves,

$$\begin{aligned}
\frac{dR_{km}}{d\alpha} &= \left\langle \frac{\lambda - e_q}{L} \sum_l^L \operatorname{sgn}(x_k^{(q,l)}) y_m^{(q,l)} \right\rangle, \\
\frac{dQ_{kk'}}{d\alpha} &= \left\langle \frac{(\lambda - e_q)^2}{L} \sum_l^L \operatorname{sgn}(x_k^{(q,l)}) \operatorname{sgn}(x_{k'}^{(q,l)}) \right\rangle \\
&\quad + \left\langle \frac{\lambda - e_q}{L} \sum_l^L \left[\operatorname{sgn}(x_k^{(q,l)}) x_{k'}^{(q,l)} + \operatorname{sgn}(x_{k'}^{(q,l)}) x_k^{(q,l)} \right] \right\rangle.
\end{aligned} \tag{15}$$

The performance of the student is characterized in terms of its generalization error. Denoting by $\mathbf{x} = (x_k)$ and $\mathbf{y} = (y_m)$ the local fields of the student and teacher nodes for an input pattern $\boldsymbol{\xi}$, one defines the single pattern error as

$$\varepsilon(\boldsymbol{\xi}) \equiv \varepsilon(\mathbf{x}, \mathbf{y}) = \frac{1}{4} \left(\frac{1}{\sqrt{K}} \sum_{k=1}^K \operatorname{sgn}(x_k) - \frac{1}{\sqrt{M}} \sum_{m=1}^M \operatorname{sgn}(y_m) \right)^2, \tag{16}$$

giving the generalization error as its average over the distribution of inputs

$$\varepsilon_g = \langle \varepsilon(\mathbf{x}, \mathbf{y}) \rangle_{\boldsymbol{\xi}}. \tag{17}$$

The normalization factor 1/4 has been chosen to allow an interpretation as probability of disagreement between student and teacher in the case of $K = M = 1$. An analytic expression of Eq. (17) in terms of the overlap matrices \mathbf{Q} , \mathbf{R} and \mathbf{T} is derived in Appendix A. Note that the generalization error, ε_g defined by Eq. (16) and Eq. (17), is used only to assess the performance of the student, while the online bag error e_q provides the graded reinforcement signal during learning.

3 Solving of the Flow Equations

By expanding the online bag error e_q defined in Eq. (6) and expressing it in terms of the jointly Gaussian fields $\{x_k^{(q,l)}\}$ and $\{y_m^{(q,l)}\}$, the right-hand sides of Eqs. (15) can be broken down into a sum of simpler terms [14]. The resulting integrals are listed in generic form in Appendix C. There are analytic expressions for the two dimensional integrals, Eqs. (24) and (25), but unfortunately not for the four and six dimensional integrals of the form of Eqs. (26) and (27) which also appear in the expansion.

We therefore decided to use Monte Carlo sampling over the Gaussian fields $\{x_k^{(q,l)}\}$ and $\{y_m^{(q,l)}\}$ to estimate the expressions on the right-hand sides of Eqs. (15). Note that this amounts to simulating the system in the thermodynamic limit $N \rightarrow \infty$, thereby avoiding finite size effects. This approach shares some features of the Eissfeller-Opper algorithm for spin-glass dynamics [22]. The mutual correlations of the fields are given by the order parameter matrices of the system \mathbf{Q} , \mathbf{R} and \mathbf{T} , as described in Eqs. (14). Samples are easily generated using a Cholesky decomposition of the overall correlation matrix constructed from these matrices. Unless explicitly noted otherwise, we used a set of 10^6 samples to estimate the expressions. Since integrating over a single time step of Eqs. (15) requires considerable computational resources, we chose a Runge-Kutta-Fehlberg 4,5 Algorithm with an adaptive step size sensitive to a relative error of 10^{-3} to integrate these equations. Once the self-correlations of the fields became very small, the step size had to be decreased to prevent the updated self-correlations from becoming negative.

To compare the theory described in the previous section with direct stochastic simulations of the learning algorithm, we have to impose appropriate initial conditions on the weight vectors \mathbf{J}_k and \mathbf{B}_m of the simulation. We start by choosing random initial weights. Afterwards, a standard Gram-Schmidt process is used to yield a set of orthonormal basis vectors $\hat{\mathbf{J}}_k$ and $\hat{\mathbf{B}}_m$. Weight vectors with desired initial overlaps are then easily expressible as linear combinations of the orthonormal ones.

4 Results

In Fig. 1 the two different methods for analyzing the learning behavior, stochastic simulation of the learning rule defined by Eqs. (3) and (5) and Monte Carlo integration of the coarse-grained equations (15), are compared for the single-layer perceptron ($K = M = 1$). We find excellent agreement between the two methods. For the initial conditions chosen in this example, we find that for a value of $\lambda = 0.1$ the network converges to the target function, exhibiting a power law decay of the generalization error. The small deviations between the simulation and the theoretical results are mainly due to finite size effects affecting the former but not the latter. Decreasing the value of the learning rate ratio λ below a critical value λ_c changes the behavior of the system, and it gets trapped in a state with a poor generalization error. This is shown for both methods in the extreme case $\lambda = 0$ in Fig. 1. In the trapped state, the simulated system is subject to fairly large fluctuations due to the stochastic nature of the inputs; these fluctuations disappear in the thermodynamic limit described by the coarse-grained theory.

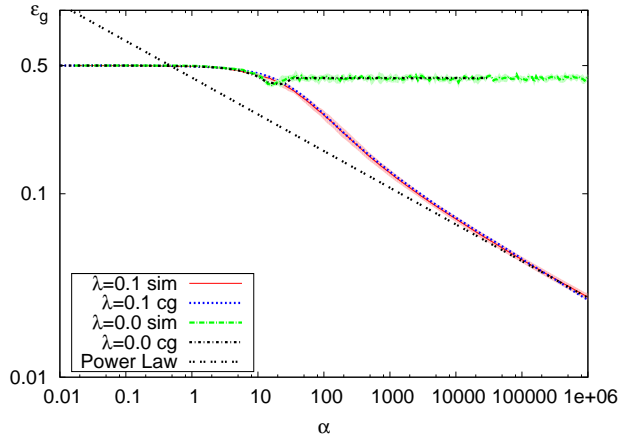


Figure 1: Comparison of simulations (sim) and coarse-grained theory (cg) for a single-layer perceptron. The size of the bags was $L = 5$ and feedback was via the hidden instance error. For the simulations, the input dimension was $N = 300$. The lighter colors in the background illustrate the standard deviations for the simulations (in this figure the standard deviations are very small and the simulation and coarse-grained data practically fall on top of each other). Initial conditions were random with $R(0) = 0$ and $Q(0) = 1.0$. The power law $\varepsilon_g(\alpha) \propto \alpha^{-\frac{1}{2L^2\lambda}}$ derived from an asymptotic analytic study [9] is in good agreement with the simulation result for $\lambda = 0.1$.

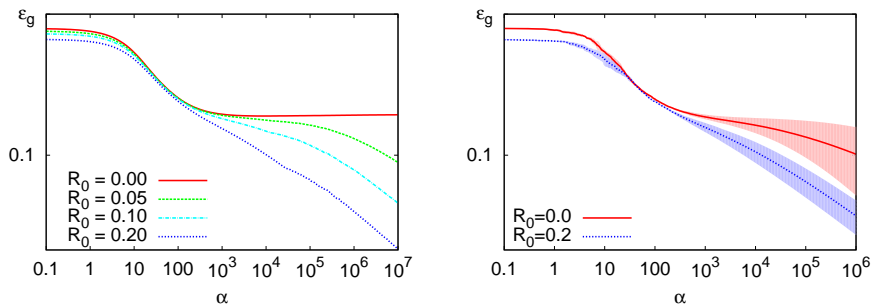


Figure 2: The evolution of the generalization error is shown for varying initial values of the teacher-student overlaps $R_{km}(0) = R_0\delta_{km}$, for a network with $K = M = 2$, $\lambda = 0.1$ and $L = 3$. The left panel shows results of the coarse-grained theory, the right panel the mean and standard deviations over 5 simulation runs for a system of $N = 300$ input nodes.

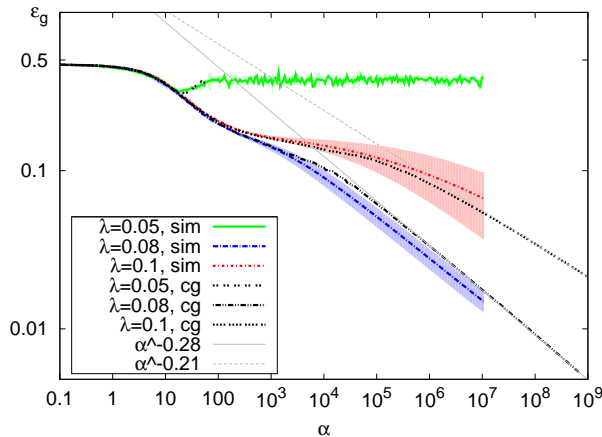


Figure 3: Simulations and coarse-grained analysis for $K = M = 2$ and bag size $L = 3$. The initial conditions were $Q_{kk'}(0) = \delta_{kk'}$ and $R_{km}(0) = 0.1\delta_{km}$. For these initial conditions, λ has to exceed a $\lambda_c \approx 0.06$ for convergence to good generalization. The input dimension was $N = 300$. The straight lines represent power law fits to the asymptotic behavior: the results from the cg-simulation with $\lambda = 0.08$ yields $\varepsilon_g = \alpha^{-0.28}$ and with $\lambda = 0.1$ yields $\varepsilon_g = \alpha^{-0.21}$.

In the case of more complex networks additional fixed points in the dynamics arise due to symmetries in the configurations of the student-teacher overlaps R_{km} [21]. As an example, consider the case of Fig. 2 for a student with $K = 2$ hidden units, trained by a teacher with $M = 2$ hidden units and $T_{mn} = \delta_{mn}$, and a bag size of $L = 3$. Initial conditions $R_{km}(0) = R_0\delta_{km}$ with different values for R_0 are chosen. In the case $R_0 = 0$, all student weights have no initial overlap with the teacher weights. This initial symmetry w.r.t. permutation of the hidden units of the student can never be broken in the thermodynamic limit (see left panel of Fig. 2). The student weights can never specialize for specific teacher weights, leading to a non-vanishing generalization error ε_g even in the $\alpha \rightarrow \infty$ limit. As the value of $R_0 \neq 0$ is increased, specialization of the student weights to teacher weights takes place earlier in the dynamics. This leads to shorter plateaus in Fig. 2 and therefore faster learning.

The right panel of Fig. 2 shows the same system for some of the initial conditions shown in the left panel, but now studied through stochastic simulations of a system with $N = 300$ input units. In finite N simulations, symmetries in initial conditions can be broken due to finite-size fluctuations, so a transition to specialization occurs even for the $R_0 = 0$ case. Analogous observations have also been made for standard supervised learning in committee machines [13, 21]. For a given R_0 , residence times in the plateau phase can vary considerably between individual runs due to large fluctuations of the time needed to achieve specialization. This is the main reason for the fairly large error bars in the ‘post-plateau’ results exhibited in the right panel of Fig. 2.

The effect of the leaning rate λ is investigated for a $K = M = 2$ committee machine in Fig. 3. As in the perceptron case, learning behavior is found to depend on λ . For a value $\lambda = 0.1 > \lambda_c \approx 0.06$, the system converges to

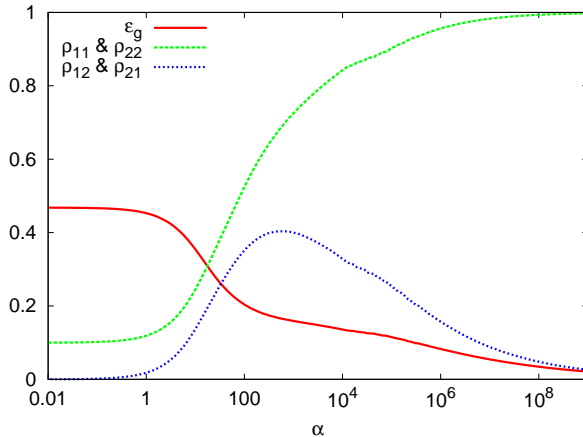


Figure 4: Plot of generalization error and the normalized order parameters $\rho_{km} = \frac{R_{km}}{\sqrt{Q_{kk}}}$ as functions of time α , for $K = M = 2, \lambda = 0.1, L = 3$. Initial conditions were $R_{km}(0) = 0.1\delta_{km}$ and $Q_{kk'} = \delta_{kk'}$, which lead to $\rho_{11} = \rho_{22}$ and $\rho_{12} = \rho_{21}$ at all values of α , as the initial symmetry is preserved in Eq. (15).

the target function with a λ -dependent power law decay of the generalization error. For $\lambda < \lambda_c$, the learner gets trapped in a state with a bad generalization error. We have found that λ_c depends on initial conditions for the student self-overlaps, for which we chose $Q_{kk'} = Q_0\delta_{kk'}$, with $Q_0 = 1$ in the present case. Theory and simulations agree to within a standard deviation of the fluctuations characteristic of the simulation results.

For an isotropic teacher with $T_{nm} = \delta_{nm}$, Figure 4 shows the normalized overlaps $\rho_{km} = \frac{R_{km}}{\sqrt{Q_{kk}}}$ for $\lambda = 0.1$; for this value of λ the system converges to a state of perfect generalization with a power law decay of the generalization error. During the transient initial phase, both the off-diagonal elements ρ_{km} and the diagonal elements ρ_{kk} increase until the hidden units of the student network specialize to emulate specific hidden units in the teacher network. If the student-teacher overlaps are initialized with $R_0 > 0$, we find that as $\alpha \rightarrow \infty$ the diagonal elements $\rho_{kk} \rightarrow 1$ while the off-diagonal elements $\rho_{km} \rightarrow 0$, indicating convergence to perfect generalization.

A global flow diagram would provide a more complete understanding of the dynamics. However, as we deal with a high-dimensional system, we have to select a reasonable subspace for visualization. We have demonstrated specialization of the student's hidden units in the case of an isotropic teacher, $T_{nm} = \delta_{nm}$. We look at a subspace with diagonal order parameter matrices: $Q_{kk'} = Q\delta_{kk'}$ and $R_{km} = R\delta_{km}$. We used $R > 0$ in order to avoid issues specifically related with symmetry between hidden units; these issues have been studied in detail (e.g. [13, 21]). In Fig. 5 we exhibit the dynamical flow in this subspace in a ε_g - \sqrt{Q} -plane, as obtained from Eqs. (15). We see that the global behavior changes with the control parameter λ . For $\lambda = 0.03$, the system has a stable fixed point at $\sqrt{Q} \approx 0.3$ and $\varepsilon_g \approx 0.4$, and a partially stable fixed point at $\sqrt{Q} \approx 0.7$ and $\varepsilon_g \approx 0.11$. The presence of these two fixed points entails that, depending on the initial conditions, learning might fail. However, we see in the second panel of

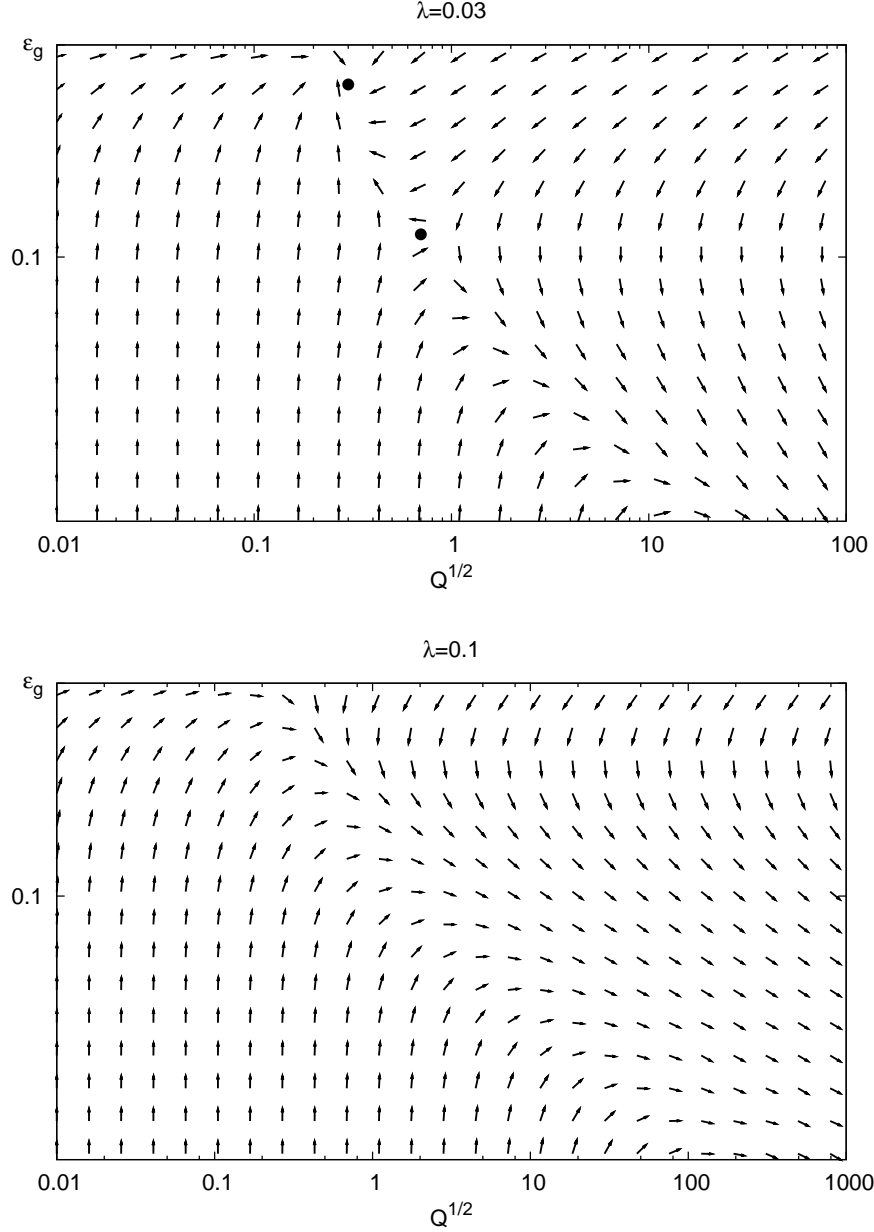


Figure 5: Flow in a subspace chosen as $Q_{kk'} = Q\delta_{kk'}$, $R_{km} = R\delta_{km}$ and $T_{mn} = \delta_{mn}$ for $L = 3$ obtained from a $K = M = 2$ system. For $\lambda = 0.03$, below the bifurcation point at $\lambda^* \approx 0.06$, a stable fixed point at $\sqrt{Q} \approx 0.3$ and $\epsilon_g \approx 0.4$ exists that may trap learning in a state with bad generalization (see Fig. 3). Above λ^* learning in this subspace always converges to perfect generalization.

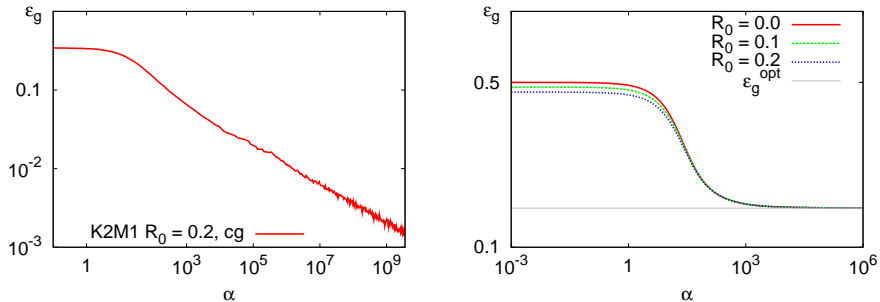


Figure 6: Over- and unrealizable learning for AR learning using a bag size $L = 3$ and $\lambda = 0.1$. The left panel shows that a student network with higher complexity ($K = 2$) than that of the target function ($M = 1$) learns perfectly, with a power law decay of the generalization error. The right panel shows that in the case of lower complexity ($K = 1$ and $M = 2$) the network finds the best possible approximation of the teacher rule.

Fig. 5 for λ above a “bifurcation point” $\lambda^* \approx 0.06$ (for the $K = M = 2$ system and $L = 3$) that the two fixed points coalesce and disappear, allowing convergence to perfect generalization *independent* of initial conditions for $\lambda > \lambda^*$ (provided that $R_0 > 0$). Similar global flow diagrams have been obtained for single-layer perceptrons [7]. Note that in general $\lambda_c \leq \lambda^*$, as for specific initial conditions learning can succeed despite of the presence of attractive fixed points with poor generalization.

In real situations, it is unlikely that a learning system, be it biological or artificial, faces the task of learning a function that has the same complexity as its inner architecture. It is therefore of both theoretical and practical interest to investigate whether the system can learn an overrealizable task ($K > M$) or whether it can approximate an unrealizable ($K < M$) task the optimal way compatible with its lower complexity. Results for these two cases are shown in Fig. 6.

Soft-committee machines of different complexity differ both in the number and range of values produced by the output unit. Thus, soft-committee machines of different complexity cannot emulate each other without adapting the normalization of their output values. There is however the possibility that by appropriately changing the normalization of the outputs a soft-committee machine with, say $K = 2$ hidden units — normally $\{-\sqrt{2}, 0, \sqrt{2}\}$ — can reproduce the outputs of a perceptron with outputs $\{-1, +1\}$, if the outputs of the former were rescaled by a factor $1/\sqrt{2}$, and the $K = 2$ -machine could be trained to avoid the output 0. Likewise, a soft-committee machine with $K = 4$ hidden units could be rescaled and trained to reproduce the outputs of a (less complex) $K = 2$ machine (e.g. by combining the hidden units of the $K = 4$ machine in pairs which do the same thing), etc. Whereas functions of lower complexity can thus sometimes be realized, the opposite is never possible: a network of lower complexity attempting to learn a target function of higher complexity will never achieve a vanishing generalization error.

Here we look at the simplest situation, a $K = 2$ machine learning a rule

defined by a perceptron with $M = 1$, and the opposite case, where a perceptron with $K = 1$ attempts to learn a rule defined by an $M = 2$ committee. In the left panel of Fig. 6 we demonstrate that the more complex network ($K = 2$) is able to approximate the simpler network ($M = 1$) with the expected power law behavior of the generalization error. For the opposite case, the right panel of Fig. 6 reveals that independently of initial overlaps of student and teacher weights, the simpler $K = 1$ perceptron is able to find the best possible approximation of an $M = 2$ committee. The optimal configuration itself is analytically characterized in Appendix B.

5 Discussion

In this paper we investigated a simple learning algorithm as a solution to the delayed reinforcement task for the committee machine. We showed that the algorithm, in addition to dealing with the non-specificity in time caused by the delayed reward, can also cope with the non-specificity in “space” presented by the hidden units in the committee machine. Good learning behavior persists over various student and teacher network complexities, extending to nonlinear decision boundaries. This flexibility can be regarded as one of the features that lends biological plausibility to the algorithm, and perhaps relevance for technical applications. The flexibility and the simplicity of the algorithm support the hypothesis that it might be phylogenetically relevant, and could have formed the basis of more advanced solutions. Indeed, its key ingredients, i.e. unsupervised Hebbian learning, global replay, and specifically the control of polarity of plasticity through neuromodulators in individual neurons, have been observed in various systems [16, 15, 20].

To analyze the system, a set of flow equations describing the deterministic evolution of the learning dynamics in the infinite size limit have been derived and were solved using Monte Carlo sampling of key integrals. The symmetries in the hidden units were shown to lead to fixed points in the dynamics, which manifest themselves as plateaus in the temporal behavior of the generalization error, much as in the conventional online learning scenario for these systems [13, 21]. In case of convergence, the power law dependence of the asymptotic generalization error over many orders of magnitude in time indicates asymptotically perfect generalization. Importantly, the power law has been found to depend in a *non universal* manner on the systems parameters. In particular, there is no perfect generalization if $\lambda < \lambda_c$, where λ_c depends on the initial conditions.

Analysis of the global flow in a subspace of symmetric systems revealed a picture similar to that of the flow in the perceptron: if the Hebbian learning rate λ characterizing the association phase of the algorithm remains below a bifurcation point λ^* , two fixed points of the learning dynamics exist, one stable and one partially stable, and convergence to perfect generalization depends on initial conditions. Increasing the value of λ moves the two fixed points closer to each other, until at λ^* they eventually coalesce and disappear, leading to asymptotic convergence independently of initial conditions in this subspace for $\lambda > \lambda^*$. Note that the fixed points originating from the permutation symmetry of the hidden units still exist, and it is left to future research to see whether the present algorithm can be modified in a manner that would speed up the breaking of this symmetry (e.g. [23, 24]).

A Generalization Error

In this section, we explicitly calculate the generalization error for the soft-committee machine with $\text{sgn}(x)$ as transfer function for the hidden units, as investigated in the present paper. We remind the reader that the term ‘‘soft’’ refers to the transfer function of the output unit. See Eq. (1) and the text below it. A similar derivation was performed in [13] for $\text{erf}(x)$ as transfer function.

As discussed in Section 2.2, in the case of independent identically distributed inputs ξ_i of zero mean and unit variance, the local fields of the hidden student and teacher units are jointly Gaussian, with covariance matrix determined by the mutual weight overlaps:

$$C = \begin{pmatrix} Q & R \\ R^T & T \end{pmatrix}.$$

We can write the multivariate Gaussian distribution of the fields as

$$P(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{(2\pi)^{M+K} |C|}} \exp\left(-\frac{1}{2}(\mathbf{x}, \mathbf{y})C^{-1}(\mathbf{x}, \mathbf{y})^T\right). \quad (18)$$

The generalization error $\langle \varepsilon(\mathbf{x}, \mathbf{y}) \rangle_{\xi}$ of Eq. (17) then becomes

$$\begin{aligned} \varepsilon_g &= \frac{1}{4K} \sum_{i,j=1}^K \langle \text{sgn}(x_i) \text{sgn}(x_j) \rangle \\ &+ \frac{1}{4M} \sum_{n,m=1}^M \langle \text{sgn}(y_n) \text{sgn}(y_m) \rangle \\ &- \frac{1}{2\sqrt{KM}} \sum_{i=1}^K \sum_{n=1}^M \langle \text{sgn}(x_i) \text{sgn}(y_n) \rangle, \end{aligned} \quad (19)$$

in which averages are over the multivariate Gaussian distribution, Eq. (18). Using marginalization we can rewrite these averages as a sum of two dimensional Gaussian integrals of the type in Eq. (25) below. Performing the appropriate substitutions we get

$$\begin{aligned} \varepsilon_g &= \frac{K}{4} - \frac{1}{2\pi K} \sum_{i,j=1}^K \arccos\left(\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}\right) \\ &+ \frac{M}{4} - \frac{1}{2\pi M} \sum_{m,n=1}^M \arccos\left(\frac{T_{mn}}{\sqrt{T_{mm}T_{nn}}}\right) \\ &- \frac{\sqrt{KM}}{2} + \frac{1}{\pi\sqrt{KM}} \sum_{i=1}^K \sum_{m=1}^M \arccos\left(\frac{R_{im}}{\sqrt{Q_{ii}T_{mm}}}\right). \end{aligned} \quad (20)$$

B Optimal ε_g for Perceptrons

In this Appendix we calculate the optimal generalization error for perceptrons approximating more complicated committee machine functions. Note that this value is independent of the specific learning algorithm and gives a bound on the performance of *any* learning algorithm.

For $K = 1$ hidden unit in the student network and M hidden units in the teacher network we find that all except the last term of Eq. (20) are independent of the student weight \mathbf{J} . Therefore minimizing the generalization error ε_g corresponds to minimizing

$$\begin{aligned}\varepsilon_g &= c + \frac{1}{\pi\sqrt{M}} \sum_{m=1}^M \arccos\left(\frac{R_{1m}}{\sqrt{Q_{11}T_{mm}}}\right) \\ &= c + \frac{1}{\pi\sqrt{M}} \sum_{m=1}^M \arccos\left(\frac{\mathbf{J} \cdot \mathbf{B}_m}{|\mathbf{J}||\mathbf{B}_m|}\right),\end{aligned}\quad (21)$$

w.r.t \mathbf{J} ; here c denotes the terms in Eq. (20) that are independent of \mathbf{J} .

As the generalization error is invariant under rescaling \mathbf{J} (or the \mathbf{B}_m), we can look for a minimum at fixed $|\mathbf{J}| = \sqrt{N}$, using a Lagrangian multiplier to enforce this constraint. For an isotropic teacher-committee with $T_{mn} = \delta_{mn}$ this gives

$$\nabla_{\mathbf{J}}(\varepsilon_g + \lambda(\mathbf{J}^2 - N)) = -\frac{1}{\pi\sqrt{M}} \sum_{m=1}^M \frac{\mathbf{B}_m}{\sqrt{1 - R_{1m}^2}} + 2\lambda\mathbf{J} = 0 \quad (22)$$

Performing a scalar product with any of the \mathbf{B}_n and exploiting isotropy, we conclude that the overlaps R_{1n} must (up to sign) be independent of n , giving

$$\mathbf{J} \propto \sum_{m=1}^M \mathbf{B}_m. \quad (23)$$

Thus the optimal weight is a linear combination of the teacher weights with equal coefficients. Inserting this into Eq. (20) for an isotropic teacher and $M = 2$ yields $\varepsilon_g^{opt} \approx 0.146$.

C Integrals

The following integrals need to be evaluated to solve the flow equations or to compute the generalization error,

$$\begin{aligned}I_1 &= \frac{1}{2\pi\sqrt{|C|}} \int_{-\infty}^{+\infty} dx_1 dx_2 \operatorname{sgn}(x_1) x_2 \exp\left(-\frac{1}{2}(x_1, x_2)C^{-1}(x_1, x_2)^T\right) \\ &= \sqrt{\frac{2}{\pi}} \frac{C_{12}}{\sqrt{C_{11}}},\end{aligned}\quad (24)$$

$$\begin{aligned}I_2 &= \frac{1}{2\pi\sqrt{|C|}} \int_{-\infty}^{+\infty} dx_1 dx_2 \operatorname{sgn}(x_1) \operatorname{sgn}(x_2) \exp\left(-\frac{1}{2}(x_1, x_2)C^{-1}(x_1, x_2)^T\right) \\ &= 1 - \frac{2}{\pi} \arccos\left(\frac{C_{12}}{\sqrt{C_{11}C_{22}}}\right),\end{aligned}\quad (25)$$

as well as their generalizations to higher dimensions

$$I_{N,1} = \frac{1}{(2\pi)^{N/2}\sqrt{|C|}} \int_{-\infty}^{+\infty} d^N x \prod_{i=1}^{N-1} \operatorname{sgn}(x_i) x_N e^{-\frac{1}{2}\mathbf{x}C^{-1}\mathbf{x}^T}, \quad (26)$$

$$I_{N,2} = \frac{1}{(2\pi)^{N/2}\sqrt{|C|}} \int_{-\infty}^{+\infty} d^N x \prod_i \operatorname{sgn}(x_i) e^{-\frac{1}{2}\mathbf{x}C^{-1}\mathbf{x}^T}. \quad (27)$$

Analytic expressions for the latter are unfortunately not in general available.

References

- [1] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 1998.
- [2] P. Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, 2001.
- [3] W. Schultz. Predictive reward signal of dopamine neurons. *J Neurophysiol*, 80:1–27, 1998.
- [4] E. T. Rolls, C. McCabe, and J. Redoute. Expected value, reward outcome, and temporal difference error representations in a probabilistic decision task. *Cereb Cortex*, 18:652–663, 2007.
- [5] M. L. Minsky. Steps toward artificial intelligence. *Proc Inst Radio Eng*, 49:8–30, 1961.
- [6] L. Mlodinov and I.-O. Stamatescu. An evolutionary procedure for machine learning. *Int J Comp Inf Sci*, 14:201–219, 1985.
- [7] R. Kühn and I.-O. Stamatescu. A two step algorithm for learning from unspecific reinforcement. *J Phys A*, 32:5749–5762, 1999.
- [8] M. Biehl, R. Kühn, and I.-O. Stamatescu. Learning structured data from unspecific reinforcement. *J Phys A*, 33:6843–6857, 2000.
- [9] R. Kühn and I.-O. Stamatescu. Learning with incomplete information and the mathematical structure behind it. *Biol Cybern*, 97:99–112, 2007.
- [10] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the theory of neural computation*. Addison Wesley, 1991.
- [11] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math Cont Sig Sys*, 2:303–314, 1989.
- [12] M. Biehl and H. Schwarze. Learning by online gradient descent. *J Phys A*, 28:643–656, 1995.
- [13] D. Saad and S. A. Solla. On-line learning in soft committee machines. *Phys Rev E*, 52:4225–4243, 1995.
- [14] U. M. Bergmann. Unspecific reinforcement learning in one and two-layered networks. Master’s thesis, University of Heidelberg, 2006.
- [15] D. J. Foster and M. A. Wilson. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440:680–683, 2006.
- [16] Z. Nádasdy, H. Hirase, A. Czurkó, J. Csicsvari, and G. Buzsáki. Replay and time compression of recurring spike sequences in the hippocampus. *J Neurosci*, 19:9497–9507, 1999.

- [17] D. R. Euston, M. Tatsuno, and B. L. McNaughton. Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *Science*, 318:1147–1150, 2007.
- [18] E. Pastalkova, V. Itskov, A. Amarasingham, and G. Buzsáki. Internally generated cell assembly sequences in the rat hippocampus. *Science*, 321:1322–1327, 2008.
- [19] H. Gelbard-Sagiv, R. Mukamel, M. Harel, R. Malach, and I. Fried. Internally generated reactivation of single neurons in human hippocampus during free recall. *Science*, 322:96–101, 2008.
- [20] G. H. Seol, J. Ziburkus, S. Huang, L. Song, I. T. Kim, K. Takamiya, R. L. Huganir, H.-K. Lee, and A. Kirkwood. Neuromodulators control the polarity of spike-timing-dependent synaptic plasticity. *Neuron*, 55:919–929, 2007.
- [21] M. Biehl, P. Riegler, and C. Wöhler. Transient dynamics of on-line learning in two-layered neural networks. *J Phys A*, 29:4769–4780, 1996.
- [22] H. Eissfeller and M. Opper. New method for studying the dynamics of disordered spin systems without finite-size effects. *Phys Rev Lett*, 68:2094–2097, 1992.
- [23] S. Bös. Matrix-update for accelerated on-line learning in multilayer networks. *J Phys A*, 31:413–417, 1998.
- [24] M. Rattray, D. Saad, and S. Amari. Natural gradient descent for on-line learning. *Phys Rev Lett*, 81:5461–5464, 1998.