On-line learning and generalization in coupled perceptrons

2002 J. Phys. A: Math. Gen. 35 2093

(http://iopscience.iop.org/0305-4470/35/9/302)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **35** (2002) 2093–2109 PII: S0305-4470(02)31340-4

On-line learning and generalization in coupled perceptrons

D Bollé and P Kozłowski

Instituut voor Theoretische Fysica, KU Leuven, B-3001 Leuven, Belgium

E-mail: desire.bolle@fys.kuleuven.ac.be and piotr.kozlowski@fys.kuleuven.ac.be

Received 27 November 2001
Published 22 February 2002
Online at stacks.iop.org/JPhysA/35/2093

Abstract

We study supervised learning and generalization in coupled perceptrons trained on-line using two learning scenarios. In the first scenario the teacher and the student are independent networks and both are represented by an Ashkin–Teller perceptron. In the second scenario the student and the teacher are simple perceptrons but are coupled by an Ashkin–Teller-type four-neuron interaction term. Expressions for the generalization error and the learning curves are derived for various learning algorithms. The analytical results find excellent confirmation in numerical simulations.

PACS numbers: 87.18.Sn, 05.20.−y, 87.10.+e

1. Introduction

One of the more interesting properties of neural networks is their ability to learn from examples. In on-line learning processes a student network updates its couplings after the presentation of each example in order to make its outputs agree with those of the teacher. In the standard situation the student knows only the inputs and the corresponding outputs of the teacher and has no further knowledge of the rule used by the latter. Furthermore, in the course of learning the student is also able to correctly classify new examples, which he has never seen before. The latter property is called generalization.

Various aspects of learning and generalization in neural networks have been intensively studied in many different contexts. For about a decade now statistical mechanical methods have been used successfully in these studies (for recent reviews see, e.g., [1–4]).

Much theoretical research has been concentrated on the simplest models, such as the binary perceptron. Simultaneously with the progress in these investigations, new more realistic models have been considered, e.g. models with multi-state neurons [5], multi-neuron interactions [6, 7] and many layers (see, e.g., [8–10]).

In this paper we study on-line learning and generalization in a recently introduced model, allowing two different types of binary neurons at each site, possibly having different

functions [11, 12]. More specifically, this so-called Ashkin–Teller (AT) perceptron contains, besides two-neuron interaction terms, a four-neuron interaction term. For the underlying biological motivation for the introduction of different types of neurons, we refer to [13]. Here, we recall that the maximal capacity of the AT perceptron model II introduced in [11, 12] can be larger than that of the standard binary perceptron [12] and that the corresponding recurrent network model can be a more efficient associative memory than the sum of two Hopfield models [13]. A natural question is then how this AT perceptron performs in on-line learning and generalization tasks.

Two learning scenarios turn out to be of interest. In the first scenario, where the student and the teacher are independent AT perceptrons, we show that the resulting learning curves do not differ very much from those already known for perceptrons with multi-state neurons. For some particular values of the network parameters, we precisely reproduce the learning curve of the four-state Potts perceptron [5].

In the second scenario both the student and the teacher are represented by a simple perceptron but they are coupled by an AT-type four-neuron interaction term. Hence, contrary to the standard set-up, they are not independent. This can be considered as a sort of 'hardware' coupling. As a result, the teacher mapping also changes in the process of learning. We obtain a set of learning curves which qualitatively differ from those found in the independent set-up. We also find different asymptotic behaviour when the number of examples increases to infinity. For certain values of the network parameters, such a coupling describes the realistic situation that the rule used by the teacher is partially shared by the student.

The rest of the paper is organized as follows. In section 2 the model and the learning scenarios are introduced. The formulae for the generalization error are derived in section 3. The differential equations for the evolution of the order parameters are obtained in section 4. Their solutions, compared with numerical simulations, can be found in section 5. In section 6 some concluding remarks are presented. Finally, the two appendices contain some technical details of the derivations.

2. The model and the learning scenarios

The AT perceptron is defined as a mapping of the binary (± 1) inputs $\{s_i, \sigma_i\}$, $i = 1, \ldots, N$ into two binary (± 1) outputs s and σ:

$$s = \mathrm{sgn}(h_1) + \theta(\gamma_3|h_3| - \gamma_1|h_1|)\theta(\gamma_2|h_2| - \gamma_1|h_1|)(\mathrm{sgn}(h_2h_3) - \mathrm{sgn}(h_1)) \tag{1}$$

$$\sigma = \mathrm{sgn}(h_2) + \theta(\gamma_3|h_3| - \gamma_2|h_2|)\theta(\gamma_1|h_1| - \gamma_2|h_2|)(\mathrm{sgn}(h_1h_3) - \mathrm{sgn}(h_2)) \tag{2}$$

where θ is the Heaviside step function and $\gamma_r \geqslant 0$, $r = 1, 2, 3$, denote the strength of the local fields h_r which are defined as follows:

$$\begin{aligned} h_1 &= \tfrac{1}{n_1}\textstyle\sum_i J_i^{(1)} s_i & h_2 &= \tfrac{1}{n_2}\textstyle\sum_i J_i^{(2)}\sigma_i \\ h_3 &= \tfrac{1}{n_3}\textstyle\sum_i J_i^{(3)} s_i\sigma_i & n_r^2 &= \textstyle\sum_i \left(J_i^{(r)}\right)^2 . \end{aligned} \tag{3}$$

The mapping (1)–(2) can be equivalently represented by the set of three equations (cf model I in [12])

$$s = \mathrm{sgn}(\gamma_1 h_1 + \sigma\gamma_3 h_3) \tag{4}$$

$$\sigma = \mathrm{sgn}(\gamma_2 h_2 + s\gamma_3 h_3) \tag{5}$$

$$s\sigma = \mathrm{sgn}(\sigma\gamma_1 h_1 + s\gamma_2 h_2). \tag{6}$$

For $\gamma_3 = 0$ the outputs s and σ are completely independent and defined as in the simple perceptron

$$s = \mathrm{sgn}(h_1) \tag{7}$$
$$\sigma = \mathrm{sgn}(h_2). \tag{8}$$

2.1. Learning scenario I

First, we consider the standard situation where the student and the teacher are two completely independent networks. In our case they are represented by AT perceptrons which means that the outputs of the teacher $\{s_T, \sigma_T\}$ and the student $\{s_S, \sigma_S\}$ are both determined by the mapping (1)–(2) but with different couplings: $\boldsymbol{J}_r^T$ and $\boldsymbol{J}_r^S$ respectively, with $\boldsymbol{J}_r = \left\{J_i^{(r)}\right\}$. Initially, the student and the teacher couplings are not correlated. At each time step t, an example is presented to the student. The student network then updates its couplings according to the following learning rule F:

$$\boldsymbol{J}_1^S(t+1) = \boldsymbol{J}_1^S(t) + \frac{1}{N} F s_T(t) \boldsymbol{s}(t) \tag{9}$$

$$\boldsymbol{J}_2^S(t+1) = \boldsymbol{J}_2^S(t) + \frac{1}{N} F \sigma_T(t) \boldsymbol{\sigma}(t) \tag{10}$$

$$\boldsymbol{J}_3^S(t+1) = \boldsymbol{J}_3^S(t) + \frac{1}{N} F s_T(t) \sigma_T(t) \boldsymbol{\psi}(t) \tag{11}$$

where

$$\boldsymbol{s} = \{s_i\} \qquad \boldsymbol{\sigma} = \{\sigma_i\} \qquad \boldsymbol{\psi} = \{s_i \sigma_i\}. \tag{12}$$

In this scenario we consider only Hebbian learning for which $F = 1$. Furthermore, examples are chosen randomly with equal probability out of the complete set of examples.

2.2. Learning scenario II

Alternatively, the AT perceptron can also be seen as two coupled perceptrons with outputs s and σ. In the second scenario we precisely analyse learning between such coupled perceptrons (or branches of the AT perceptron). The outputs of the student s and the teacher σ are defined by equations (1) and (2), respectively.

When $h_3 > 0$, the teacher and the student use two different mixtures of two perceptron mappings defined by the couplings $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$. It implies that s and σ are always equal to $\mathrm{sgn}(h_1)$ or $\mathrm{sgn}(h_2)$ and sometimes, depending on the relation between $\gamma_1 h_1$, $\gamma_2 h_2$ and $\gamma_3 h_3$, $s = \sigma$. In the limit $\gamma_3 \to \infty$, the student and the teacher networks become so strongly coupled that one always has $s = \sigma$ and the mapping (1)–(2) can be simplified to

$$s = \sigma = \mathrm{sgn}(h) \qquad h = \{h_x : |h_x| > |h_y|;\ x, y = 1, 2\}. \tag{13}$$

For $h_3 < 0$, the situation is quite different. Even with $\boldsymbol{J}_1 = \boldsymbol{J}_2$, there is always a non-zero fraction of disagreements between the student and the teacher, as long as $\gamma_3 > 0$. In the limit $\gamma_3 \to \infty$, the student always disagrees with the teacher, and the mapping (1)–(2) can be written in the form

$$s = \begin{cases} -\sigma = \mathrm{sgn}(h_1) & \text{if} \quad |h_1| > |h_2| \\ -\sigma = -\mathrm{sgn}(h_2) & \text{if} \quad |h_1| < |h_2| \end{cases}. \tag{14}$$

For any value of the coupling field h_3 and $\gamma_3 = 0$, the student and the teacher are independent and they use the mappings defined by only one coupling vector (cf (7)–(8)).

In what follows, we take $\boldsymbol{s} = \boldsymbol{\sigma}$ because the student and the teacher must have the same inputs. We remark that this implies that $h_3 = \sum_i J_i^{(3)} / n_3$ (cf (3)). Again, at each time step t, an example is presented to the student network and its coupling vector $\boldsymbol{J}_1$ is updated as follows:

$$\boldsymbol{J}_1(t+1) = \boldsymbol{J}_1(t) + \frac{1}{N} F(\gamma_1 h_1, \gamma_3 h_3, s, \sigma)\, \sigma(t) \boldsymbol{s}(t). \tag{15}$$

Furthermore, at each time step a new coupling vector $\boldsymbol{J}_3$ is generated thus making the coupling between the perceptrons random. The coupling vector of the teacher, $\boldsymbol{J}_2$, is not changed in the process of learning, but later on we average over all possible teachers. In this scenario we consider three learning rules F:

Hebbian $F(\gamma_1 h_1, \gamma_3 h_3, s, \sigma) = 1$

perceptron $F(\gamma_1 h_1, \gamma_3 h_3, s, \sigma) = \theta(-s\sigma)$

Adatron $F(\gamma_1 h_1, \gamma_3 h_3, s, \sigma) = -(\sigma \gamma_1 h_1 + \gamma_3 h_3)\theta(-s\sigma)$.

3. Generalization error

A topic of interest in what follows is the generalization error. It is defined as the probability that the student and the teacher disagree, i.e. that their outputs are different. When the teacher and the student are simple independent perceptrons, the generalization error $\varepsilon_g = \arccos(\rho)/\pi$ is a simple function of the overlap $\rho = \boldsymbol{J}^T \cdot \boldsymbol{J}^S / (n^S n^T)$ between the student and the teacher couplings, which in this case plays the role of an order parameter. Unfortunately, for more complicated models, this relation takes a much more involved form (see, e.g., [5]).

3.1. Scenario I

In the first scenario the definition of the generalization error reads

$$\varepsilon_g(\rho_1, \rho_2, \rho_3) = \left\langle 1 - \tfrac{1}{4}(1 + s_T s_S)(1 + \sigma_T \sigma_S) \right\rangle_{\mathrm{I}} \tag{16}$$

with the overlaps ρ_r defined by

$$\rho_r = \frac{\boldsymbol{J}_r^S \cdot \boldsymbol{J}_r^T}{n_r^T n_r^S} \tag{17}$$

and with $\langle \cdots \rangle_{\mathrm{I}} = \int \mathrm{d}\boldsymbol{h}^{\boldsymbol{T}}\, \mathrm{d}\boldsymbol{h}^{\boldsymbol{S}} \cdots P_{\mathrm{I}}(\boldsymbol{h}^{\boldsymbol{T}}, \boldsymbol{h}^{\boldsymbol{S}})$ denoting the average over the teacher field, $\boldsymbol{h}^{\boldsymbol{T}} = \left\{h_1^T, h_2^T, h_3^T\right\}$, and the student field, $\boldsymbol{h}^{\boldsymbol{S}} = \left\{h_1^S, h_2^S, h_3^S\right\}$, which have a joint probability distribution $P_{\mathrm{I}}(\boldsymbol{h}^{\boldsymbol{T}}, \boldsymbol{h}^{\boldsymbol{S}})$. The averages over these fields are double averages: one over the examples and the other over the couplings. This arises because the couplings and the examples enter the mapping (1)–(2) and the learning rules only through the local fields. We assume that the examples are taken randomly with equal probability out of the full training set. Then, in the thermodynamic limit, the local fields become correlated Gaussian variables and the joint probability distribution $P_{\mathrm{I}}(\boldsymbol{h}^{\boldsymbol{T}}, \boldsymbol{h}^{\boldsymbol{S}})$ can be written in the form

$$P_{\mathrm{I}}(\boldsymbol{h}^{\boldsymbol{T}}, \boldsymbol{h}^{\boldsymbol{S}}) = \left(\left(1-\rho_1^2\right)\left(1-\rho_2^2\right)\left(1-\rho_3^2\right)\right)^{-1/2} \frac{1}{2\pi^3} \exp\left\{ \frac{\rho_1 h_1^S h_1^T}{1-\rho_1^2} + \frac{\rho_2 h_2^S h_2^T}{1-\rho_2^2} + \frac{\rho_3 h_3^S h_3^T}{1-\rho_3^2} - \frac{1}{2}\left[\frac{\left(h_1^S\right)^2 + \left(h_1^T\right)^2}{1-\rho_1^2} + \frac{\left(h_2^S\right)^2 + \left(h_2^T\right)^2}{1-\rho_2^2} + \frac{\left(h_3^S\right)^2 + \left(h_3^T\right)^2}{1-\rho_3^2} \right] \right\}. \tag{18}$$

Performing the averages in (16) explicitly leads to the expression

$$\varepsilon_g(\rho_1, \rho_2, \rho_3) = \frac{3}{4} - \sum_{r=1}^{3} I_r \tag{19}$$

with

$$I_r = \frac{1}{2}\int_0^\infty \mathrm{D}h_r^T \,\mathrm{erf}\left(\frac{\rho_r h_r^T}{\sqrt{2\left(1-\rho_r^2\right)}}\right)\left[1-2\left(1-\mathrm{erf}\left(\frac{\gamma_r h_r^T}{\gamma_{r'}\sqrt{2}}\right)\right)\left(1-\mathrm{erf}\left(\frac{\gamma_r h_r^T}{\gamma_{r''}\sqrt{2}}\right)\right)\right]$$
$$+\frac{1}{4}\int \mathrm{D}\left(h_r^T, h_r^S\right)\left[(a_{rr'}^{+}-a_{rr'}^{-})(a_{rr''}^{+}-a_{rr''}^{-})\right.$$
$$\left.+(a_{rr'}^{+}+a_{rr'}^{-})(a_{rr''}^{+}+a_{rr''}^{-})\mathrm{sgn}\left(h_r^T h_r^S\right)\right] \tag{20}$$

$$a_{rr'}^{\pm} = \frac{1}{2}\left(1-\mathrm{erf}\left(\frac{\gamma_r\left|h_r^T\right|}{\gamma_{r'}\sqrt{2}}\right)\right) - \int_{-\infty}^{-\frac{\gamma_r\left|h_r^T\right|}{\gamma_{r'}}} \mathrm{D}h_{r'}^T \,\mathrm{erf}\left(\frac{\gamma_r\left|h_r^S\right| \pm \gamma_{r'}\rho_{r'}h_{r'}^T}{\gamma_{r'}\sqrt{2\left(1-\rho_v^2\right)}}\right) \tag{21}$$

where $\mathrm{D}z = \mathrm{d}z\exp(-z^2/2)/\sqrt{2\pi}$ is the Gaussian measure, $r', r'' = 1, 2, 3$ $(r \neq r' \neq r'' \neq r)$ and

$$\mathrm{D}\left(h_r^T, h_r^S\right) = \frac{\mathrm{d}h_r^T\,\mathrm{d}h_r^S}{2\pi\sqrt{1-\rho_r^2}}\exp\left\{-\frac{1}{2}\frac{\left(h_r^T\right)^2+\left(h_r^S\right)^2-2\rho_r h_r^S h_r^T}{1-\rho_r^2}\right\} \tag{22}$$

is a correlated Gaussian.

3.2. Scenario II

In the second scenario the generalization error is given by

$$\varepsilon_g(\rho) = \langle 1-\tfrac{1}{2}(1+s\sigma)\rangle_{\mathrm{II}} = \int \mathrm{d}\boldsymbol{h}\,P_{\mathrm{II}}(\boldsymbol{h})(1-\tfrac{1}{2}(1+s\sigma)) \tag{23}$$

with the overlap ρ defined by

$$\rho = \frac{\boldsymbol{J}_1\cdot\boldsymbol{J}_2}{n_1 n_2}. \tag{24}$$

Here again, as in the first scenario, the average over the examples and the couplings is done through averaging over the local fields. The examples are chosen randomly with equal probability out of the full set of examples. In the thermodynamic limit, this leads to a Gaussian distribution of the local fields. Since the behaviour of the system strongly depends on the sign of the coupling field h_3, we consider three different field distributions P_{II}

$$P_{\pm}(\boldsymbol{h}) = ((2\pi)^3(1-\rho^2))^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{h_1^2+h_2^2-2h_1h_2\rho}{1-\rho^2}+h_3^2\right)\right\} \tag{25}$$

$$P_{+}(\boldsymbol{h}) = 2P_{\pm}(\boldsymbol{h})\theta(h_3) \tag{26}$$

$$P_{-}(\boldsymbol{h}) = 2P_{\pm}(\boldsymbol{h})\theta(-h_3). \tag{27}$$

In the case of the distribution $P_{\pm}$, the components of the vector $\boldsymbol{J}_3$ are taken randomly (with equal probability) from some interval $(-a, a)$, with a being a positive real number. In the case of the distributions P_{+} and P_{-}, these components are chosen in the same way but those

values which lead to negative respectively positive values of the field h_3 are omitted. The generalization error in these three situations reads, with obvious notation,

$$\varepsilon_g^c(\rho) = \frac{1}{\pi}\arccos(\rho) + I_c \qquad c = \pm, +, - \tag{28}$$

where

$$I_\pm = \tfrac{1}{2}(u_{12}^- - u_{12}^+ + u_{21}^- - u_{21}^+) \qquad I_+ = -u_{12}^+ - u_{21}^+ \qquad I_- = u_{12}^- + u_{21}^- \tag{29}$$

and

$$u_{rr'}^\pm = \int_{-\infty}^0 \mathrm{D}h_2 \left(1 + \mathrm{erf}\left(\frac{\gamma_r h_2}{\gamma_3\sqrt{2}}\right)\right)\left(1 + \mathrm{erf}\left(\frac{h_2(\gamma_r/\gamma_r' \pm \rho)}{\sqrt{2(1-\rho^2)}}\right)\right). \tag{30}$$

It is easy to realize that only for positive h_3 (i.e. for $P_{\mathrm{II}} = P_+$), does the generalization error $\varepsilon_g^+(\rho)$ go to zero as ρ goes to 1. It is also equal to zero for any ρ when $P_{\mathrm{II}} = P_+$ and $\gamma_3 = \infty$.

4. Order parameters and their evolution

As can be seen from the formulae in the last section, the generalization error is a function of the overlaps ρ or ρ_r, which play the role of order parameters in the learning process. Their evolution is coupled with the evolution of the norms of the couplings n_r, and in the thermodynamic limit $N \to \infty$ it can be described by ordinary differential equations [14].

In the first scenario a standard calculation (for a review see, e.g., [2]) leads to the following result for Hebbian learning:

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}n_r = \left\langle \Psi_r^T h_r^S \right\rangle_{\mathrm{I}} + \frac{1}{2n_r} \qquad \frac{\mathrm{d}}{\mathrm{d}\alpha}\rho_r = \frac{1}{n_r}\left\langle \Psi_r^T \left(h_r^T - \rho_r h_r^S\right)\right\rangle_{\mathrm{I}} - \frac{\rho_r}{2n_r^2} \qquad r = 1, 2, 3 \tag{31}$$

where $\Psi_1^T = s_T$, $\Psi_2^T = \sigma_T$, $\Psi_3^T = s_T\sigma_T$ and $\alpha = t/N$ is the number of examples scaled with the size of the system. It becomes continuous in the thermodynamic limit. After performing the averages, we arrive at

$$\frac{\mathrm{d}n_r}{\mathrm{d}\alpha} = \rho_r b_r + \frac{1}{2n_r} \qquad \frac{\mathrm{d}\rho_r}{\mathrm{d}\alpha} = \frac{1-\rho^2}{n_r}b_r - \frac{\rho_r}{2n_r^2} \tag{32}$$

with the quantity b_r given by

$$b_r = \sqrt{\frac{2}{\pi}}\left\{\frac{1}{\sqrt{c_{r'r}}}\left[1 - 2\int_0^\infty \mathrm{D}h\,\mathrm{erf}\left(\frac{h\gamma_{r'}}{\gamma_{r''}\sqrt{2c_{r'r}}}\right)\right] + \frac{1}{\sqrt{c_{r''r}}}\left[1 - 2\int_0^\infty \mathrm{D}h\,\mathrm{erf}\left(\frac{h\gamma_{r''}}{\gamma_{r'}\sqrt{2c_{r''r}}}\right)\right]\right\} \tag{33}$$

$$c_{rr'} = 1 + \left(\frac{\gamma_r}{\gamma_{r'}}\right)^2. \tag{34}$$

For $\gamma_1 = \gamma_2 = \gamma_3$, this quantity simplifies to

$$b_r = \frac{2}{\sqrt{\pi}}\left(1 - \frac{2}{\pi}\arctan\left(\frac{1}{\sqrt{2}}\right)\right) \cong 0.6864. \tag{35}$$

We remark that the differential equations (32) for a given r have the same form as those found for the simple perceptron with Hebbian learning [2]. More specifically, they differ only by the value of the coefficient b_r, which for the simple perceptron is equal to $\sqrt{2/\pi} \approx 0.798$.

For the Hebbian learning we are considering, it is possible to construct a simple expression for ρ_r as a function of α. Following Opper and Kinzel [1], we slightly modify the update rule (9)–(11) (substituting $1/N$ by $1/\sqrt{N}$) and easily arrive at

$$\rho_r = \sqrt{\frac{\alpha a_r^2}{\alpha a_r^2 + \pi}} \tag{36}$$

where we have taken $\rho(0) = 0$ as an initial condition and

$$a_r = 2\sqrt{\pi} \int_0^\infty \mathrm{D}b\, b \left[1 - \left(1 - \mathrm{erf}\left(\frac{\gamma_r b}{\gamma_{r'}\sqrt{2}} \right) \right) \left(1 - \mathrm{erf}\left(\frac{\gamma_r b}{\gamma_{r''}\sqrt{2}} \right) \right) \right]. \tag{37}$$

This expression differs from the solution of (32) only for small values of α and has the advantage of having a simple form. The evolution of ρ in the case of simple perceptrons is described by the single equation (36), but with a coefficient $a_r = \sqrt{2}$. Since these results are very similar to those obtained for the simple perceptron, we do not test other algorithms in this scenario because we expect that in those cases also a strong resemblance to the simple perceptron occurs.

In the second scenario with the learning rule F defined in subsection 2.2, we have to solve the following set of differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} n_1 = \langle h_1 \sigma F(\gamma_1 h_1, \gamma_3 h_3, s, \sigma) \rangle_{\mathrm{II}} + \frac{1}{2n_1} \left\langle F^2(\gamma_1 h_1, \gamma_3 h_3, s, \sigma) \right\rangle_{\mathrm{II}} \tag{38}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \rho = \frac{1}{n_1} \langle \sigma F(\gamma_1 h_1, \gamma_3 h_3, s, \sigma)(h_2 - \rho h_1) \rangle_{\mathrm{II}} - \frac{\rho}{2n_1^2} \left\langle F^2(\gamma_1 h_1, \gamma_3 h_3, s, \sigma) \right\rangle_{\mathrm{II}}. \tag{39}$$

Performing the averages leads to much more complicated expressions than those obtained in the first scenario. The explicit form of these expressions obtained for Hebbian, perceptron and Adatron learning with the distributions $P_\pm$ and P_+ can be found in appendix A.

5. Results

In this section we discuss the numerical solutions of the differential equations (31), (38) and (39) and compare them with the results of simulations. Because only the ratios of the strength parameters γ_1, γ_2 and γ_3 are important, we take $\gamma_1 = \gamma_2 = 1$ and vary only γ_3.

5.1. Scenario I

The learning curves for small values of the number of examples α obtained in the first scenario using formula (36) are presented in figure 1. All curves start with an initial generalization error $\varepsilon_g = 0.75$ corresponding to random guessing in four-state models. For $\gamma_3 = 0$, learning between two independent perceptrons is described. For $\gamma_3 = 1$ the learning curve is identical with that of the four-state Potts perceptron [5] (cf [11, 12]). In the limit $\alpha \to \infty$, ε_g decays as $\alpha^{-\frac{1}{2}}$ for all values of γ_3, precisely as in the case of learning between simple perceptrons.

5.2. Scenario II

A careful analysis of expression (28) leads to the conclusion that in the second scenario the generalization error can be non-zero even when the normalized angle between the student and the teacher couplings, $\phi = \arccos(\rho)/\pi$, is equal to zero. This happens when we allow the field h_3 to take negative values. Therefore, we follow the evolution of two dynamical variables in the following: the generalization error ε_g and the normalized angle between the student

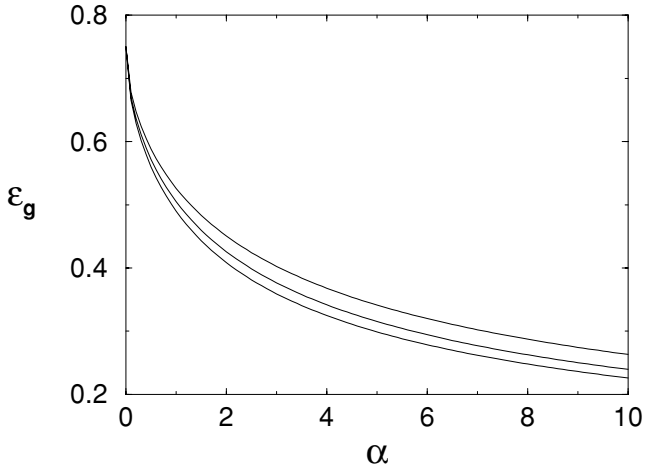

Figure 1. Learning scenario I: the generalization error ε_g as a function of the number of examples α with $\gamma_1 = \gamma_2 = 1$ and $\gamma_3 = \infty, 1, 0$ from top to bottom.

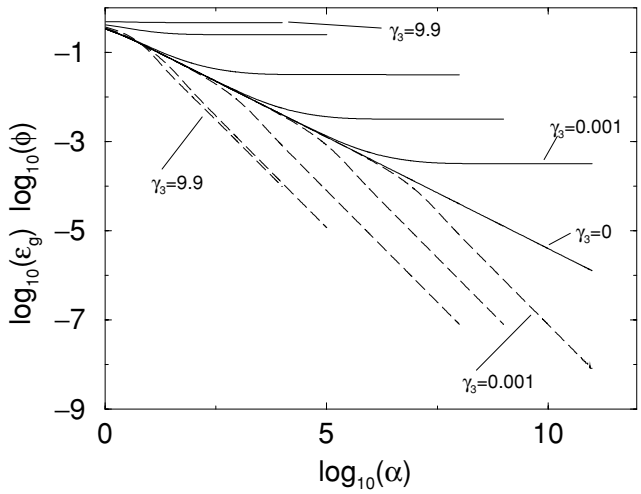

Figure 2. Learning scenario II: log–log plot of the generalization error ε_g (solid lines) and the normalized angle between the teacher and the student ϕ (broken lines) for $P = P_\pm$ and Hebbian learning as a function of the number of examples α. Intermediate curves not marked on the figure are for $\gamma_3 = 0.01, 0.1, 1$.

and the teacher ϕ. For all the learning algorithms and distributions of the fields that we have considered, we observe an abrupt change in the asymptotic behaviour in α when γ_3 changes from 0 to some non-zero value. Logarithmic plots of the learning curves for two distributions of the fields, $P_\pm$ and P_+, are presented in figures 2–7. The learning curves for the distribution P_- are qualitatively very similar to the curves obtained for $P_\pm$.

5.2.1. $P_{\mathrm{II}} = P_\pm$. Let us first analyse the results obtained for the distribution $P_\pm$ in more detail. For $\gamma_3 \neq 0$, the generalization error saturates at some non-zero value. For Hebbian and perceptron learning, the angle ϕ between the student and the teacher is asymptotically

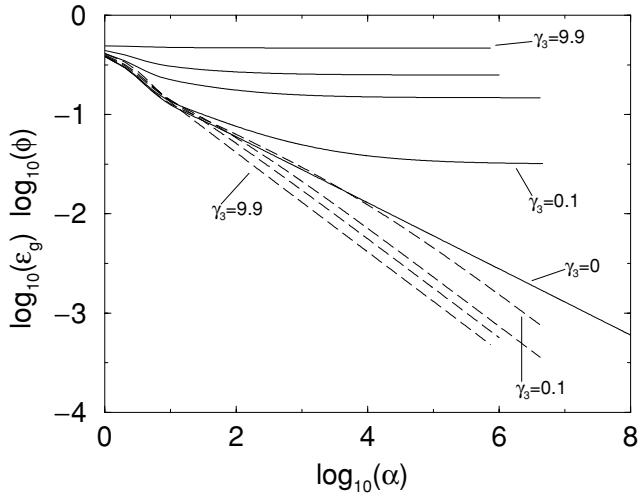

Figure 3. As in figure 2 but for the perceptron algorithm. Intermediate curves not marked on the figure are for $\gamma_3 = 0.5, 1$.

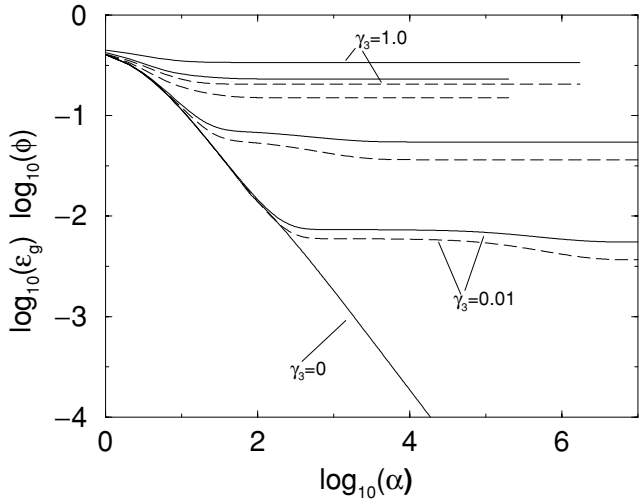

Figure 4. As in figure 2 but for the Adatron algorithm. Intermediate curves not marked on the figure are for $\gamma_3 = 0.1, 0.5$.

decreasing to zero at a rate higher than that in the decoupled case $\gamma_3 = 0$. For Hebbian learning we find that in the limit $\alpha \to \infty$, $\phi \sim \alpha^{-1}$, versus $\phi \sim \alpha^{-\frac{1}{2}}$ for $\gamma_3 = 0$, whereas in the case of the perceptron algorithm $\phi \sim \alpha^{-\frac{1}{2}}$, versus $\phi \sim \alpha^{-\frac{1}{3}}$ for $\gamma_3 = 0$. For the Adatron algorithm ϕ and ε_g, both saturate at some non-zero value. In spite of the fact that the generalization error never vanishes, the student is able to learn the couplings of the teacher using the Hebbian or perceptron algorithm.

5.2.2. $P_{\mathrm{II}} = P_+$. We observe that for all algorithms the generalization error goes asymptotically to zero. For Hebbian and perceptron learning it decreases faster than in

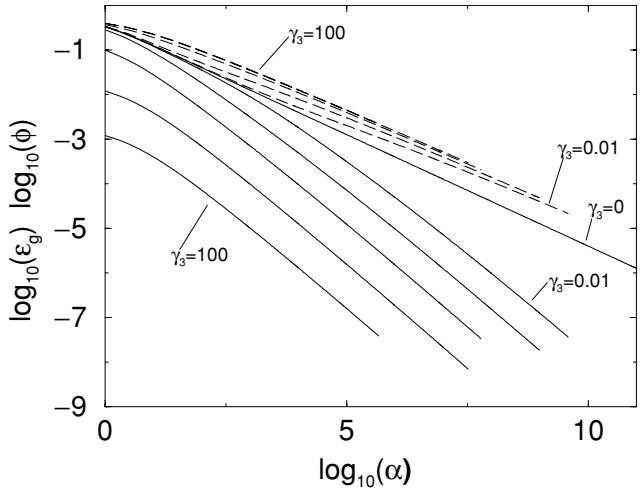

Figure 5. Learning scenario II: log–log plot of ε_g (solid lines) and ϕ (broken lines), for $P = P_+$ and Hebbian learning as a function of α. Intermediate curves not marked on the figure are for $\gamma_3 = 0.1, 1.0, 9.9$.

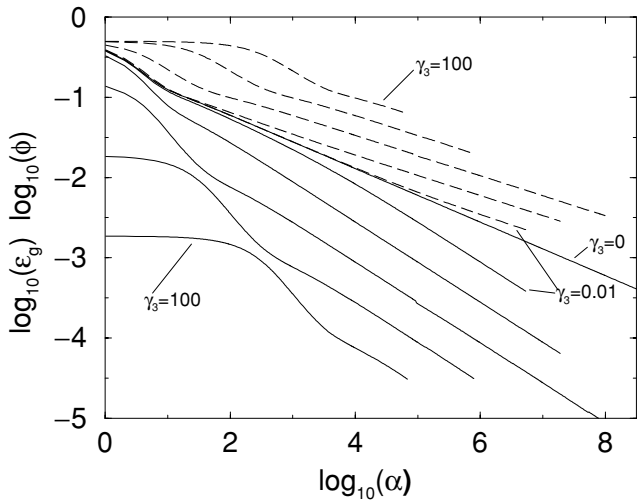

Figure 6. As in figure 5 but for the perceptron algorithm.

the decoupled case. In the limit $\alpha \to \infty$, we get $\varepsilon_g \sim \alpha^{-1}$ for Hebbian learning while $\varepsilon_g \sim \alpha^{-\frac{1}{2}}$ for perceptron learning. For Adatron learning we obtain the same decay exponent as in the decoupled case. Surprisingly, for the perceptron and Adatron algorithms the decay of the angle between the student and the teacher, ϕ, is slower than in the decoupled case in the limit $\alpha \to \infty$. For the perceptron we have $\phi \sim \alpha^{-\frac{1}{4}}$ and for the Adatron we find $\phi \sim \alpha^{-\frac{1}{2}}$. In contrast, for Hebbian learning $\phi \sim \alpha^{-\frac{1}{2}}$ as for the decoupled case.

Since an analytical analysis of the differential equations (see appendix A) is rather involved, the asymptotic exponents discussed above have been determined numerically. Only in the case of Hebbian learning with the field distribution P_+ was the numerical analysis not

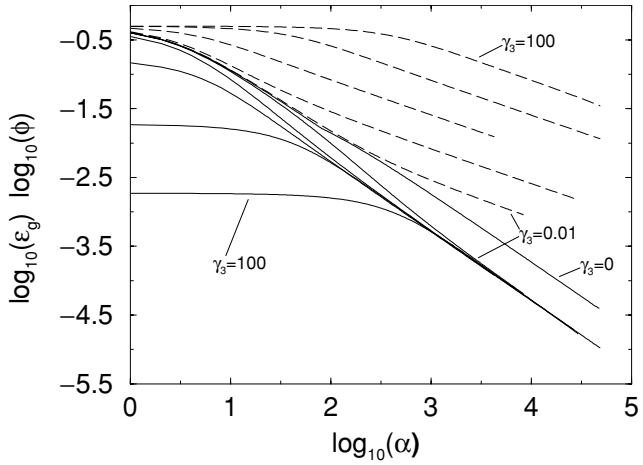

Figure 7. As in figure 5 but for the Adatron algorithm.

entirely unambiguous. Therefore, we have derived the corresponding exponents analytically. The details can be found in appendix B.

The initial generalization error is a function of the strength parameter γ_3, which measures the strength of the coupling between the two perceptrons. The larger the γ_3, the greater the common knowledge between the student and the teacher, and hence the smaller the initial error. For $\gamma_3 \to \infty$, the student and the teacher use precisely the same rule (13) in order to determine their outputs.

Finally, the numerical solution of equations (38)–(39) suggests that there is a simple relation between the decay exponents of ϕ and ε_g, denoted by y_ϕ and y_g, respectively,

$$y_g = 2y_\phi. \tag{40}$$

This relation can also be derived analytically (see appendix B). For $\gamma_1 = \gamma_2$ we find in the limit $\alpha \to \infty$ (and $\phi \to 0$) that

$$\varepsilon_g^+ \sim \frac{\pi^2}{4\sqrt{2}\gamma_3}\phi^2 \tag{41}$$

which confirms the observation (40).

5.3. Computer simulations

To check the analytical results described above, we have performed numerical simulations. The system sizes have been varied between $N = 100$ and $N = 999$ neurons. An excellent agreement has been found for both scenarios and all learning algorithms, even for relatively small N. As a representative example, we present a comparison between simulations and analytical results obtained in the second scenario with the Adatron algorithm for $\gamma_3 = 0.1$ and $P_{\mathrm{II}} = P_{\pm}$. For the sake of clarity, we show the results obtained for small and large α separately. The analytical results for small α are compared with simulations for a system with $N = 999$ neurons (figure 8). For larger α we have made simulations for smaller systems ($N = 100$), which are displayed in figure 9. In both cases only the results obtained for one sample are shown.

For small α the simulations are smoothly aligned along the theoretical curves. This points to the self-averaging property of the learning process. For larger values of α, very strong

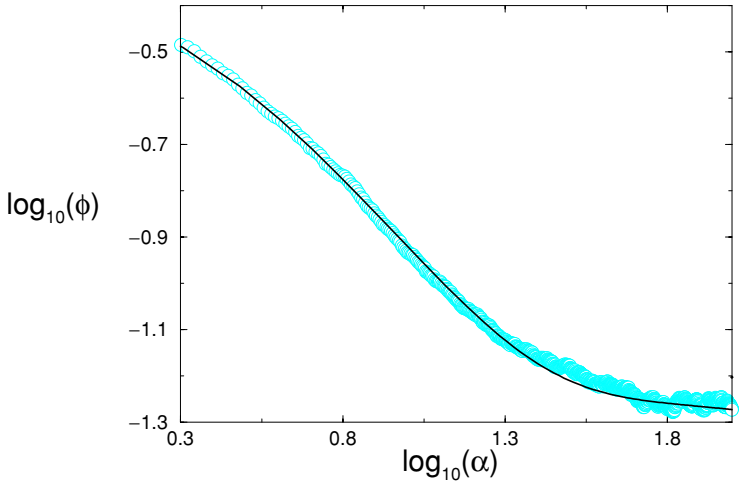

Figure 8. Second learning scenario with Adatron learning and $\gamma_3 = 1$. Simulations (grey circles) with $N = 999$ versus theoretical results (solid black line) for ϕ as a function of α.

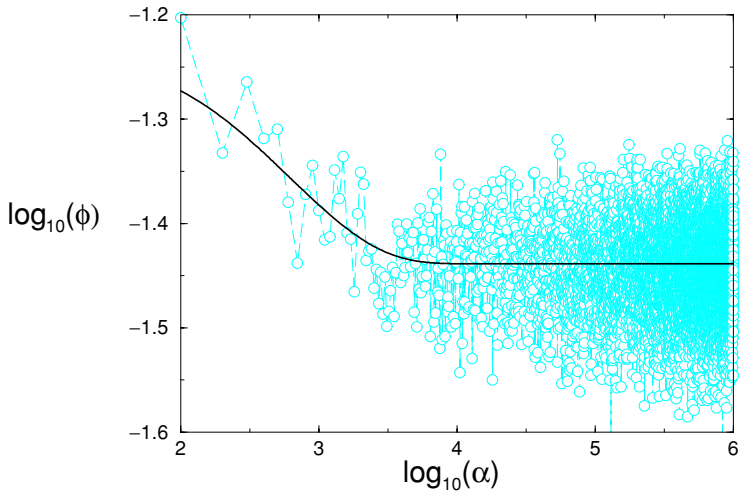

Figure 9. As in figure 8 with $N = 100$.

fluctuations occur around the theoretical result. This happens only for the Adatron algorithm and $P_{\rm II} = P_{\pm}$ and, hence, cannot be explained entirely by the relatively small size of the system. Indeed, as has been noted in section 5, in this case there is always a non-zero fraction of disagreement between the student and the teacher. So, a strategy used by the Adatron algorithm, which updates the couplings proportional to the error made by the student, must lead to rather large random changes. Nevertheless, the simulation points in figure 9 are evenly distributed on both sides of the theoretical curve.

6. Conclusions

In this paper we have studied on-line learning and generalization using the AT perceptron. Two learning scenarios have been considered. The results obtained in the first scenario, where the student and the teacher are represented by independent AT perceptrons, are very similar to those obtained for the simpler models [2]. For a particular choice of the network parameters, the learning curve precisely reproduces that found for the four-state Potts perceptron [5].

In the second scenario the student and the teacher are taken to be simple perceptrons coupled by a four-neuron interaction term. The particular results depend crucially on the distribution of the couplings $\boldsymbol{J}_3$.

For the field distribution $P_{\mathrm{II}} = P_{\pm}$, the generalization error always saturates at some non-zero value. This is not surprising since this distribution allows the field h_3 to take negative values, which inevitably leads to a non-vanishing fraction of disagreements between the student and the teacher even when $\boldsymbol{J}_1 = \boldsymbol{J}_2$ (cf (1) and (2)). In spite of this, for Hebbian and perceptron learnings, the student manages to learn the couplings of the teacher perfectly (in the limit $\alpha \to \infty$). This does not happen, however, for the Adatron algorithm, which in the standard (decoupled) situation proved to be the fastest [2]. The reason is that this algorithm changes the couplings of the student proportionally to the error made by the latter. Since this error is non-zero even for $\boldsymbol{J}_1 = \boldsymbol{J}_2$, this cannot be a good strategy. Hence, the more 'blind' updates (Hebbian and perceptron) appear to be more effective.

For $P_{\mathrm{II}} = P_{+}$ we have obtained quite different results. In this case the generalization error goes to zero when ρ goes to 1. For Hebbian and perceptron learnings, we observe a faster decay of ε_g than in the decoupled case. For Adatron learning the decay exponent of ε_g is the same as that for $\gamma_3 = 0$. Surprisingly, for all algorithms we find the same or slower decay of ϕ compared with the decoupled case.

The best asymptotic decay of the generalization error has been obtained for $P_{\mathrm{II}} = P_{+}$ with the Adatron rule: $\varepsilon_g \sim 0.618\alpha^{-1}$. Comparing with the case of independent perceptrons, we see that it is better than the lower bound for on-line learning [2] $\left(\varepsilon_g \sim 0.88\alpha^{-1}\right)$ and worse than the Bayesian lower bound [15] $\left(\varepsilon_g \sim 0.44\alpha^{-1}\right)$.

We remark that in the course of a learning process in the second scenario also the teacher mapping is changed but not the teacher couplings. This can be interpreted as a kind of effective mutual learning caused by the ('hardware') coupling of the two perceptrons. This is different from the mutual learning process analysed in [16, 17], the only other learning process of this type known to us. There, in contrast to our set-up, the teacher explicitly learns from the student. In our model the decay exponent of ε_g is not influenced by a particular value of the strength parameter γ_3 as long as it is non-zero.

The model analysed in the second scenario with $P_{\mathrm{II}} = P_{+}$, where a part of the learning rule is shared by the teacher and the student, can be compared to a real life situation in which both of them, for example, have the same cultural background, followed the same education, etc. One can expect that in such a situation the learning process is much more efficient since the student and the teacher speak, in a sense, the same language. It corresponds to a faster asymptotic decay of the generalization error in our model. It would be interesting to see, for example, whether an optimization of the learning process [14] would further improve these results.

Acknowledgment

This study was supported in part by the Fund for Scientific Research, Flanders (Belgium).

Appendix A. The evolution of the order parameters in the second learning scenario

The set of differential equations, (38) and (39), for the order parameters in the second learning scenario can be written in the following form:

$$\frac{\mathrm{d}n_1}{\mathrm{d}\alpha} = f_1(\rho, \gamma_{13}, \gamma_{23}) + \frac{1}{2n_1} f_2(\rho, \gamma_{13}, \gamma_{23})$$
$$\frac{\mathrm{d}\rho}{\mathrm{d}\alpha} = \frac{1}{n_1} f_3(\rho, \gamma_{13}, \gamma_{23}) - \frac{\rho}{2n_1^2} f_2(\rho, \gamma_{13}, \gamma_{23})$$

with $\gamma_{rr'} = \gamma_r/\gamma_{r'}$ and where the explicit form of $f_1(\rho, \gamma_{13}, \gamma_{23})$, $f_2(\rho, \gamma_{13}, \gamma_{23})$ and $f_3(\rho, \gamma_{13}, \gamma_{23})$ depends on the algorithm used and the distribution of the fields.

In the case of the distribution $P_{\mathrm{II}} = P_{\pm}$, we have for
Hebbian learning

$$f_1(\rho, \gamma_{13}, \gamma_{23}) = \rho f_{21} + g_{21}$$
$$f_2(\rho, \gamma_{13}, \gamma_{23}) = 1$$
$$f_3(\rho, \gamma_{13}, \gamma_{23}) = f_{21}(1-\rho^2) - \rho g_{21}$$

Perceptron learning

$$f_1(\rho, \gamma_{13}, \gamma_{23}) = \tfrac{1}{2}(\rho f_{21} - f_{12} + g_{21})$$
$$f_2(\rho, \gamma_{13}, \gamma_{23}) = \tfrac{1}{\pi}\arccos(\rho) + I_{\pm}$$
$$f_3(\rho, \gamma_{13}, \gamma_{23}) = \tfrac{1}{2}\big(f_{21}(1-\rho^2) - g_{12} - \rho g_{21}\big)$$

Adatron learning

$$f_1(\rho, \gamma_{13}, \gamma_{23}) = -\gamma_1\left(f_a - f_{12}^{+} + f_{12}^{-} + \frac{1}{2}\right) - \gamma_3\left(t_{12} - \rho t_{21} + \frac{\sqrt{1-\rho^2}}{2\pi}\left(\frac{1}{\sqrt{c_{21}^{-}}} + \frac{1}{\sqrt{c_{21}^{+}}}\right)\right)$$

$$f_2(\rho, \gamma_{13}, \gamma_{23}) = \gamma_1^2\left(f_a - f_{12}^{+} + f_{12}^{-} + \frac{1}{2}\right) - \gamma_3^2\left(\frac{1}{\pi}\arcsin(\rho) - I_{\pm} - \frac{1}{2}\right)$$
$$+ \gamma_1\gamma_3\left(t_{12} - 2\rho t_{21} + \frac{\sqrt{1-\rho^2}}{\pi}\left(\frac{1}{\sqrt{c_{21}^{-}}} + \frac{1}{\sqrt{c_{21}^{+}}}\right)\right) - \gamma_2\gamma_3 t_{21}$$

$$f_3(\rho, \gamma_{13}, \gamma_{23}) = \gamma_1\left(\frac{1}{\pi}\left(\sqrt{1-\rho^2} + \rho\arcsin(\rho)\right) + g_{21}^{a} + g_{12}^{a} + \rho(f_a + f_{21}^{+} - f_{21}^{-})\right)$$
$$+ \gamma_3\left(t_{21}(1-\rho^2) + \frac{\sqrt{1-\rho^2}}{2\pi}\left(\frac{1}{\sqrt{c_{12}^{-}}} + \frac{1}{\sqrt{c_{12}^{+}}} + \frac{\rho}{\sqrt{c_{21}^{-}}} + \frac{\rho}{\sqrt{c_{21}^{+}}}\right)\right)$$

with

$$f_{rr'} = \sqrt{\frac{2}{\pi}} - \int_0^{\infty} \mathrm{D}h_r\, h_r \left(1 - \mathrm{erf}\left(\frac{\gamma_{r3}h_r}{\sqrt{2}}\right)\right)\left[2 - \mathrm{erf}\left(\frac{h_r(\gamma_{rr'}+\rho)}{\sqrt{2(1-\rho^2)}}\right)\right.$$
$$\left. - \mathrm{erf}\left(\frac{h_r(\gamma_{rr'}-\rho)}{\sqrt{2(1-\rho^2)}}\right)\right]$$

$$g_{rr'} = \frac{1}{b_{rr'}}\sqrt{\frac{1-\rho^2}{2\pi}}\left(1 - \frac{2}{\pi}\arctan\left(\frac{\gamma_{r3}}{b_{rr'}}\right)\right) - \frac{1}{a_{rr'}}\sqrt{\frac{1-\rho^2}{2\pi}}\left(1 - \frac{2}{\pi}\arctan\left(\frac{\gamma_{r3}}{a_{rr'}}\right)\right)$$

$$g^a_{rr'} = \frac{\sqrt{1-\rho^2}}{2\pi}\left\{\frac{1}{a^2_{rr'}}\left[\left(1+\gamma^2_{3r}a^2_{rr'}\right)^{-\frac{1}{2}}-1\right]+\frac{1}{b^2_{rr'}}\left[\left(1+\gamma^2_{3r}b^2_{rr'}\right)^{-\frac{1}{2}}-1\right]\right\}$$

$$f_a = \int_{-\infty}^{0} \mathrm{D}h_1\, h_1^2\,\mathrm{erf}\left(\frac{h_1\rho}{\sqrt{2(1-\rho^2)}}\right) + 2(1-\rho^2)^{\frac{3}{2}}\int_0^{\infty}\mathrm{D}h_2\left(1-\mathrm{erf}\left(\frac{\gamma_{23}h_2}{\sqrt{2}}\sqrt{1-\rho^2}\right)\right)$$
$$\times\int_{\gamma_{21}h_2}^{\infty}\mathrm{D}h_1\, h_1^2\sinh(\rho h_1 h_2)$$

$$f^{\pm}_{rr'} = \frac{1}{2}\int_{-\infty}^{0}\mathrm{D}h_r\, h_r^2\left(1+\mathrm{erf}\left(\frac{\gamma_{r3}h_r}{\sqrt{2}}\right)\right)\mathrm{erf}\left(\frac{h_r(\gamma_{rr'}\pm\rho)}{\sqrt{2(1-\rho^2)}}\right)$$

$$t^{\pm}_{rr'} = -\frac{1}{2\pi c_{r3}}\left(\frac{c_{r3}(1-\rho^2)}{(\gamma_{rr'}\pm\rho)^2}+1\right)^{-\frac{1}{2}}+\frac{1}{c_{r3}2\pi}\qquad t_{rr'}=t^{+}_{rr'}-t^{-}_{rr'}$$

$$c^{\pm}_{rr'} = 1+(\gamma_{r3})^2+\frac{(\gamma_{rr'}\pm\rho)^2}{1-\rho^2}\qquad a_{rr'}=\sqrt{\frac{1+(\gamma_{rr'})^2-2\gamma_{rr'}\rho}{1-\rho^2}}$$

$$b_{rr'} = \sqrt{\frac{1+(\gamma_{rr'})^2+2\gamma_{rr'}\rho}{1-\rho^2}}$$

where $I_{\pm}$ is given by expression (29) and c_{r3} is defined in expression (34).

In the case of the distribution $P_{\mathrm{II}} = P_{+}$, we have for

Hebbian learning

$$f_1(\rho,\gamma_{13},\gamma_{23}) = \rho f^{'+}_{21} + g^{+}_{21}$$
$$f_2(\rho,\gamma_{13},\gamma_{23}) = 1$$
$$f_3(\rho,\gamma_{13},\gamma_{23}) = (1-\rho^2) f^{'+}_{21} - \rho g^{+}_{21}$$

Perceptron learning

$$f_1(\rho,\gamma_{13},\gamma_{23}) = \tfrac{1}{2}(\rho f^{'+}_{21} - f^{'+}_{12} + g^{+}_{21})$$
$$f_2(\rho,\gamma_{13},\gamma_{23}) = \frac{1}{\pi}\arccos(\rho) + I_{+}$$
$$f_3(\rho,\gamma_{13},\gamma_{23}) = \tfrac{1}{2}(f^{'+}_{21}(1-\rho^2) - g^{+}_{12} - \rho g^{+}_{21})$$

Adatron learning

$$f_1(\rho,\gamma_{13},\gamma_{23}) = -\gamma_1\left[f^{+}_a - 2f^{+}_{12} - g_a + \frac{1}{2}\right] + \gamma_3\left[\frac{1}{\pi}(1-\rho) - 2t^{+}_{12} + 2\rho t^{+}_{21} - \frac{1}{\pi}\sqrt{\frac{1-\rho^2}{c^{+}_{21}}}\right]$$

$$f_2(\rho,\gamma_{13},\gamma_{23}) = \gamma_1^2\left[f^{+}_a - 2f^{+}_{12} - g_a + \frac{1}{2}\right] - \gamma_3^2\left[\frac{1}{\pi}\arcsin(\rho) - I_{+} - \frac{1}{2}\right]$$
$$+\gamma_1\gamma_3\left[-\frac{2}{\pi}(1-\rho) + 2t^{+}_{12} - 4\rho t^{+}_{21} + \frac{2}{\pi}\sqrt{\frac{1-\rho^2}{c^{+}_{21}}}\right] - 2\gamma_2\gamma_3 t^{+}_{21}$$

$$f_3(\rho,\gamma_{13},\gamma_{23}) = \gamma_1\left[\frac{1}{\pi}\left(\sqrt{1-\rho^2}+\rho\arcsin(\rho)\right)+\rho f^{+}_a + \rho g_a + 2\left(g^{b}_{12}+g^{b}_{21}+\rho f^{+}_{21}\right)\right]$$
$$+\gamma_3\left[2t^{+}_{21}(1-\rho^2) - \frac{1}{\pi}\left(1-\rho^2-\sqrt{\frac{1-\rho^2}{c^{+}_{12}}}-\rho\sqrt{\frac{1-\rho^2}{c^{+}_{21}}}\right)\right]$$

with

$$f_a^+ = \int_{-\infty}^{0} \mathrm{D}h_1\, h_1^2 \operatorname{erf}\left(\frac{\rho h_1}{\sqrt{2(1-\rho^2)}}\right) - 2(1-\rho^2)^{\frac{3}{2}} \int_{-\infty}^{0} \mathrm{D}h_2 \left(1+\operatorname{erf}\left(\frac{\gamma_{23}h_2}{\sqrt{2}}\sqrt{1-\rho^2}\right)\right)$$
$$\times \int_{-\infty}^{-\gamma_{21}|h_2|} \mathrm{D}h_1\, h_1^2 \exp(-\rho h_1 h_2)$$

$$f_{rr'}^{\prime +} = \sqrt{\frac{2}{\pi}} + 2\int_{-\infty}^{0} \mathrm{D}h\, h \left(1+\operatorname{erf}\left(\frac{\gamma_{r3}h}{\sqrt{2}}\right)\right)\left[1+\operatorname{erf}\left(\frac{h(\gamma_{rr'}+\rho)}{\sqrt{2(1-\rho^2)}}\right)\right]$$

$$g_a = \int_{-\infty}^{0} \mathrm{D}h_1\, h_1^2 \left[1+\operatorname{erf}\left(\frac{\gamma_{13}h_1}{\sqrt{2}}\right)\right] \qquad g_{rr'}^{+} = \frac{2}{b_{rr'}}\sqrt{\frac{1-\rho^2}{2\pi}}\left(1-\frac{2}{\pi}\arctan\left(\frac{\gamma_{r3}}{b_{rr'}}\right)\right)$$

$$g_{rr'}^{b} = \frac{\sqrt{1-\rho^2}}{2\pi}\frac{1}{b_{rr'}^2}\left[\left(1+\gamma_{3r}^2 b_{rr'}^2\right)^{-\frac{1}{2}} - 1\right]$$

and where I_+ is given by expression (29).

Appendix B. The asymptotic form of the solution in the second scenario for Hebbian learning with $P_{\mathrm{II}} = P_+$

Because the dependence of the generalization error ε_g^+ on the overlap ρ is rather complicated (see (28)), we derive the asymptotic form for ε_g^+ in two steps. First, we find the asymptotic relation between ε_g^+ and ϕ and then determine the behaviour of ϕ as a function of α in the limit $\alpha \to \infty$.

B.1 Asymptotic relation between ε_g^+ and ϕ

The generalization error ε_g^+ is defined as (see (28))

$$\varepsilon_g^+ = \frac{1}{\pi}\arccos\rho - u_{12}^+ - u_{21}^+ = \phi - u_{12}^+ - u_{21}^+$$

with the integrals u_{12}^+, u_{21}^+ given by (30). We now expand these integrals as a function of ϕ for small values of ϕ. First, we change the variables to get

$$u_{rr'}^+ = \phi \int_{-\infty}^{0} \frac{\mathrm{e}^{-\frac{1}{2}x^2\phi^2}}{\sqrt{2\pi}}\,\mathrm{d}x(1+\operatorname{erf}(a\phi x))(1+\operatorname{erf}(cx)) \equiv \phi\int_{-\infty}^{0}\mathrm{d}x f(\phi,x)$$

where

$$a = \frac{\gamma_r}{\gamma_3\sqrt{2}} \qquad c = \frac{\gamma_{rr'}+1}{\pi\sqrt{2}}.$$

Expanding $f(\phi,x)$ with respect to ϕ and taking $\gamma_1 = \gamma_2 = 1$, we get

$$u_{rr'}^+ = \phi\frac{1}{\sqrt{2\pi}c} - \phi^2\frac{\sqrt{2}a}{4c^2\pi} - o(\phi^3)$$

which leads to

$$\varepsilon_g^+ = \frac{\pi^2}{4\sqrt{2}\gamma_3}\phi^2 + o(\phi^3). \tag{B.1}$$

B.2 Asymptotic relation between ϕ and α

The differential equations (38) and (39) can be written in terms of the variables n_1 and ϕ. For Hebbian learning and $P_{\mathrm{II}} = P_+$, this gives

$$\frac{\mathrm{d}n_1}{\mathrm{d}\alpha} = f_1(\cos(\pi\phi), \gamma_{13}, \gamma_{23}) + \frac{1}{2n_1} f_2(\cos(\pi\phi), \gamma_{13}, \gamma_{23}) \tag{B.2}$$

$$\frac{\mathrm{d}\phi}{\mathrm{d}\alpha} = -\frac{f_3(\cos(\pi\phi), \gamma_{13}, \gamma_{23})}{n_1 \pi \sin(\pi\phi)} + \frac{\cos(\pi\phi)}{2\pi n_1^2 \sin(\pi\phi)} f_2(\cos(\pi\phi), \gamma_{13}, \gamma_{23}). \tag{B.3}$$

The functions $f_1(\cos(\pi\phi), \gamma_{13}, \gamma_{23})$, $f_2(\cos(\pi\phi), \gamma_{13}, \gamma_{23})$ and $f_3(\cos(\pi\phi), \gamma_{13}, \gamma_{23})$ are defined in appendix A. By expanding the rhs of the differential equations (B.2) and (B.3) around $\phi = 0$ up to the first non-vanishing term, we can easily find that for $\gamma_1 = \gamma_2$

$$\phi = \sqrt{\frac{\sqrt{2}}{2\pi(\sqrt{2}-1)}} \alpha^{-\frac{1}{2}}.$$

By combining this result with (B.1), we obtain the asymptotic formula for the generalization error:

$$\varepsilon_g^+ = \frac{\pi}{8\gamma_3(\sqrt{2}-1)} \alpha^{-1} + o\left(\alpha^{-\frac{3}{2}}\right).$$

References

[1] Opper M and Kinzel W 1996 *Models of Neural Networks III* ed E Domany *et al* (Berlin: Springer) p 151
[2] Mace C W H and Coolen A C C 1998 *Stat. Comput.* **8** 55
[3] Saad D (ed) 1998 *On-line Learning in Neural Networks* (Cambridge: Cambridge University Press)
[4] Engel A and Van den Broeck C 2001 *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)
[5] Watkin T L H, Rau A, Bollé D and van Mourik J 1992 *J. Physique I* **2** 167
[6] Botelho E, de Almeida R M C and Iglesias J R 1995 *J. Phys. A: Math. Gen.* **28** 1879
[7] Yoon H and Oh J-H 1998 *J. Phys. A: Math. Gen.* **31** 7771
[8] Hansel D, Mato G and Meunier C 1992 *Europhys. Lett.* **20** 471
[9] Kabashima Y 1994 *J. Phys. A: Math. Gen.* **27** 1917
[10] Copelli M and Caticha N 1995 *J. Phys. A: Math. Gen.* **28** 1615
[11] Kozłowski P and Bollé D 2001 *Disordered and Complex Systems* ed P Sollich *et al* (New York: AIP) p 49
[12] Bollé D and Kozłowski P 2001 *Phys. Rev.* E **64** 011915
[13] Bollé D and Kozłowski P 1999 *J. Phys. A: Math. Gen.* **32** 8577
[14] Kinouchi O and Caticha N 1992 *J. Phys. A: Math. Gen.* **25** 6243
[15] Opper M and Hausler D 1991 *Phys. Rev. Lett.* **66** 2677
[16] Kinzel W, Metzler R and Kanter I 2000 *J. Phys. A: Math. Gen.* **33** L141
[17] Metzler R, Kinzel W and Kanter I 2000 *Phys. Rev.* E **62** 2555