Generalizing with perceptrons in the case of structured phase- and pattern-spaces

1998 J. Phys. A: Math. Gen. 31 2519

(http://iopscience.iop.org/0305-4470/31/11/005)

J. Phys. A: Math. Gen. **31** (1998) 2519–2540. Printed in the UK PII: S0305-4470(98)87417-9

Generalizing with perceptrons in the case of structured phase- and pattern-spaces*

G Dirscherl†, B Schottky‡ and U Krey†

† Institut für Physik II der Universität Regensburg, Universitätsstrasse 31, D-93040 Regensburg, Germany

‡ Department of Computer Science and Applied Mathematics, Aston University, Birmingham B4 7ET, UK

Received 9 September 1997, in final form 15 December 1997

Abstract. We investigate the influence of different kinds of structures on the learning behaviour of a perceptron performing a classification task defined by a teacher rule. The underlying pattern distribution is permitted to have spatial correlations. The prior distribution for the teacher coupling vectors itself is assumed to be nonuniform. Thus, classification tasks of quite different difficulty are included. As learning algorithms we discuss Hebbian, Gibbs, and Bayesian learning with different priors, using methods from statistics and the replica formalism. We find that the Hebb rule is quite sensitive to the structure of the actual learning problem, failing asymptotically in most cases. In contrast, the behaviour of the more sophisticated methods of Gibbs and Bayes learning is influenced by the spatial correlations only in an intermediate regime of α, where α specifies the size of the training set. In view of the Bayesian case, we show how enhanced prior knowledge improves the performance.

1. Introduction

In the statistical physics of neural networks one of the most important paradigms is the *learning of a rule from examples* [1, 2]. The simplest case is where: (i) the rule can be represented by a 'teacher perceptron', while (ii) at the same time the neural network, which tries to learn the rule, is also given by a perceptron, called the 'student'. However, although much is known on this generalization problem, at least for single-layer perceptrons, see e.g. [1, 2] and references therein, two simplifying assumptions are usually made, namely that (a) the 'rule' itself, and (b) the examples, are both completely random, i.e. (a) without correlations between the components B_i, $i = 1, \dots, N$, of the teacher perceptron's coupling vector $\boldsymbol{B}$ connecting the N input units i to the output unit, and (b) without correlations between the components ξ_i^μ with different i and/or μ, respectively, of the inputs $\boldsymbol{\xi}^\mu$.

In practical cases there exist of course such correlations, i.e. both *spatial* correlations (e.g. in the 'rule' $\boldsymbol{B}$, i.e. between different components B_i of the teacher perceptron, and/or in the components ξ_i of the vectors $\boldsymbol{\xi}$ representing the inputs to be classified by the system) and also *semantic* correlations (e.g. two different inputs $\boldsymbol{\xi}^\mu$ and $\boldsymbol{\xi}^\nu$ may represent different 'handwritings' of the same word). Here we only mention that storage problems with semantic correlations have been treated in [3, 4] and concentrate in the following on *spatial* correlations, by assuming that all patterns $\boldsymbol{\xi}^\mu$ are drawn independently from the same non-trivial probability distribution, see below. In context with the simpler 'storage

[*] Based on the Diploma thesis of G Dirscherl, Regensburg 1996.

capacity problem', spatial correlations have already been treated in [3–5], but the 'correlated *generalization* problem' itself, which is the focus of our paper, has not yet been studied, as far as the authors know, except in a paper of Tarkowski and Lewenstein [6], where only the special case of Gibbs learning with uncorrelated teacher couplings was discussed.

In all these papers on correlated patterns, [3–6], only *single-layer perceptrons* have been considered, whereas for *uncorrelated patterns* the generalization problem has also been extensively treated for *multilayer perceptrons*. Although many interesting results, which may also be of practical relevance, have been obtained for these more realistic multilayer networks, see e.g. [7, 8], this was for uncorrelated systems and uncorrelated tasks only. Moreover, it has turned out in these and similar studies that multilayer networks cannot be treated successfully without a proper understanding of the behaviour of the *single-layer sub-perceptrons*, which are the building blocks of the multilayer systems. Therefore we concentrate here on those 'prerequisite single-layer perceptrons', treating the influence of spatial correlations on the generalization ability of these simplest neural networks. As we will see, this influence can be useful or detrimental, depending on the task and on the system. If possible, we mention explicitly in the text, or at the end in the discussion, which of our results can be transferred to multilayer systems and can perhaps be used in some kind of 'strategy'. Nevertheless one should stress here that the single-layer perceptron itself has recently become a quite popular and successful classifier in so-called *support vector machines* [9] and is more than just a toy model—thus far the motivation of the following.

In this paper we consider exclusively the case of so-called *batch learning*, i.e. the 'student system' is always trained with all examples, which are kept in mind without any preference, and is forced to classify not only the last training example, but *all* members of the training set correctly, whereas with the so-called 'online learning' (see e.g. [10]) at every training step a *new* pattern is presented to the student and the student only uses this newly added example in the training. Extending our work to multilayer perceptrons for 'batch learning' would be in fact rather expansible whereas it is much easier for the case of online learning. These questions are under investigation.

In the following, by analytical methods we study therefore the generalization problem 'with spatial structure' as specified below; a 'student perceptron' is considered, trying to learn by batch-algorithms a rule given by a 'spatially structured teacher perceptron'. The set of training examples itself is also spatially structured, and we study, how the student takes over the spatial correlations inherent in the training examples and in the teacher perceptron, and how the generalization ability depends on these parameters as a function of the size α of the training set. The main problem is of course, how the spatial structure can be used most effectively, implicitly or explicitly, by the learning process considered. As learning algorithms we study Hebbian learning, Gibbs learning, and Bayesian learning, using statistical methods and the replica formalism. Although the spatial structure of the patterns and of the teacher machine does not matter asymptotically for $\alpha \to \infty$ in the two last-mentioned cases (see below), we find that the correlations, as well as enhanced prior information in the Bayesian case, can be quite useful at *intermediate* values of α.

In view of the spatial structure considered below, we concentrate on the basic case of *segmentation*—or more general *quasisegmentation*, see below—of the system into a finite, or infinite, number of segments, which have a finite mutual correlation between the activity of the neurons belonging to the same (resp. different) segments, and similarly partitioned correlations (but with different strengths) of the synaptic couplings joining these neurons. Real data have such correlations, and it is usually part of preprocessing the data to detect such global dependencies, for example by *principal-component analysis* (PCA) see e.g. [11, ch 8], or [12]. Although in the simplest case we consider spatial correlations corresponding to just two segments of equal size, is a restriction, the basic properties can actually be investigated quite clearly. On the other hand it is rather natural to assume similar correlations in the classifying 'teacher rule' as well as in the patterns; this reflects the fact that similarities in the properties of typical patterns correspond to a similar impact on the classification labels of the patterns. This is again a property encountered in practice. More details are given below.

2. Basic definitions

We consider as usual a system with binary input patterns $\boldsymbol{\xi}^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)$, where the ξ_i^μ are ± 1. These input patterns generate at the teacher and student perceptrons, respectively, the so-called post-synaptic fields

$$h_B := \frac{1}{\sqrt{N}} \sum_{i=1}^{N} B_i \xi_i = \frac{1}{\sqrt{N}} \boldsymbol{B} \cdot \boldsymbol{\xi} \tag{1}$$

and

$$h_J := \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i \xi_i = \frac{1}{\sqrt{N}} \boldsymbol{J} \cdot \boldsymbol{\xi}. \tag{2}$$

The corresponding outputs are $\sigma_B := \operatorname{sign} h_B$, which is the 'correct output', given by the teacher, and $\sigma_J := \operatorname{sign} h_J$. The stability of the student's output—if it is correct—is given by the positive quantity $\kappa := \sigma_B \boldsymbol{J} \cdot \boldsymbol{\xi} / (|\boldsymbol{J}| \sqrt{N})$.

As usual, the *generalization ability* $g(\alpha)$ is defined as the probability that the student, after training, produces the same output as the teacher on a newly added random input, which does not belong to the training set. Here the 'newly added random input' is specified as follows. It should be different from the training inputs, but drawn from the same probability distribution, i.e. with the same spatial correlations (see below). The corresponding *error probability* is $\epsilon := 1 - g(\alpha)$. If there are no correlations, ϵ is given as usual by the overlap $r := (\boldsymbol{J} \cdot \boldsymbol{B}) / (|\boldsymbol{J}| \cdot |\boldsymbol{B}|)$ of the coupling vectors of the two perceptrons, by $g(\alpha) = 1 - (1/\pi) \arccos(r)$, see e.g. [1, 2]. With correlations, however, the following non-trivial pattern- and (teacher-)phase-space correlation matrices come into play for $i, j = 1, \ldots, N$:

$$C_{ij}^P \equiv C_{ji}^P := \langle \xi_i \xi_j \rangle_{\boldsymbol{\xi}} \qquad \text{and} \qquad C_{ij}^T \equiv C_{ji}^T := \langle B_i B_j \rangle_{\boldsymbol{B}}. \tag{3}$$

(For $i = j$ these correlations are of course trivial, i.e. $C_{ii}^P = 1$, $C_{ii}^T = \boldsymbol{B}^2/N$ (also $= 1$ without restriction).) The brackets $\langle \ldots \rangle_{\boldsymbol{\xi}}$ resp. $\langle \ldots \rangle_{\boldsymbol{B}}$ imply ensemble averages with the corresponding binomial (resp. Gaussian) probability densities, e.g.

$$P(\boldsymbol{B}) = [(2\pi)^N \operatorname{Det} \mathbf{C}^T]^{-1/2} \exp\left[-\tfrac{1}{2} \sum_{i,j=1}^{N} B_i (C_{ij}^T)^{-1} B_j \right]. \tag{4}$$

In the following we skip the sub-indexes $\boldsymbol{\xi}$ and $\boldsymbol{B}$ for simplicity, since we additionally assume that the system is *self-averaging*; i.e. for almost all configurations of the patterns $\boldsymbol{\xi}$ and of the teacher perceptron $\boldsymbol{B}$ considered, the same correlation matrices $\mathbf{C}^P$ and $\mathbf{C}^T$, and also the expressions defined below, can not only be obtained by the ensemble averages $\langle \ldots \rangle_{\boldsymbol{\xi}}$ (resp. $\langle \ldots \rangle_{\boldsymbol{B}}$), but also for *fixed* realization by averaging over equivalent pairs of sites (i, j) in the limit of infinitely large systems, $N \to \infty$, see below. Moreover, as already mentioned, we exclude semantic correlations by requiring that for different patterns $\boldsymbol{\xi}^\mu$ and

$\boldsymbol{\xi}^\nu$ one always has $\langle \xi_i^\mu \xi_j^\nu \rangle = 0$ for $i, j = 1, \ldots, N$. With these definitions one obtains additionally the important parameters

$$T := \langle (h_B)^2 \rangle = N^{-1} \sum_{i,j=1}^{N} \langle B_i B_j \xi_i \xi_j \rangle = N^{-1} \sum_{i,j=1}^{N} C_{ij}^T C_{ij}^P \tag{5}$$

$$S := \langle (h_J)^2 \rangle = N^{-1} \sum_{i,j=1}^{N} \langle J_i J_j \xi_i \xi_j \rangle = N^{-1} \sum_{i,j=1}^{N} C_{ij}^P \langle J_i J_j \rangle \tag{6}$$

and

$$R := \langle h_J \cdot h_B \rangle = N^{-1} \sum_{i,j=1}^{N} C_{ij}^P \langle J_i B_j \rangle. \tag{7}$$

Here T is fixed by the 'teacher rule' and the spatial pattern correlations, while S and R change in course of the learning process.

As already mentioned, our paper is motivated by the natural assumption that the spatial pattern correlations, and the phase-space correlations as well, i.e. spatial correlations in the couplings, correspond structurally to a *segmented system* in a similar way as words are segmented into letters, but recognized as a whole, [13]. Such a segmentation arises implicitly or explicitly in many application tasks. It is also natural to assume that pattern- and phase-space correlations are *segmented in the same way*, which means that the correlation matrices have *the same eigenvectors* $\boldsymbol{\epsilon}^k = (\epsilon_1^k, \ldots, \epsilon_N^k)$, with $k = 1, \ldots, N$, although the corresponding eigenvalues C_k^P and C_k^T may be drastically different [14, 6]. In fact, only this agreement of the eigenvectors is what we postulate in the following, when talking of '*the general quasisegmented case*'. Moreover, we often specialize below to '*the simplest segmented case*' by making the natural assumption of only two segments of the same size:

$$\boldsymbol{\xi} := (\boldsymbol{\xi}^0, \boldsymbol{\xi}^1) = (\xi_1^0, \ldots, \xi_{N/2}^0, \xi_1^1, \ldots, \xi_{N/2}^1) \tag{8}$$

with

$$\langle \xi_i^0 \xi_j^1 \rangle = \delta_{i,j} c_p \qquad \langle \xi_i^0 \xi_j^0 \rangle = \langle \xi_i^1 \xi_j^1 \rangle = \delta_{i,j} \tag{9}$$

and analogously $\boldsymbol{B} := (\boldsymbol{B}^0, \boldsymbol{B}^1) = (B_1^0, \ldots, B_{N/2}^0, B_1^1, \ldots, B_{N/2}^1)$ with

$$\langle B_i^0 B_j^1 \rangle = \delta_{i,j} c_t \qquad \langle B_i^0 B_j^0 \rangle = \langle B_i^1 B_j^1 \rangle = \delta_{i,j} \tag{10}$$

for $i, j = 1, \ldots, N/2$. The correlation parameters c_p and c_t have to be smaller than 1 in magnitude, otherwise they can be arbitrary real numbers. During the training process, also the *student* perceptron develops a similar segmentation with a correlation parameter c_s.

In the 'general quasisegmented case', the generalization ability $g(\alpha)$ is obtained from the three parameters T, S and R defined in equations (5)–(7), by $g = 2\int_0^\infty \mathrm{d}h_J \int_0^\infty \mathrm{d}h_B\, P(h_J, h_B)$, with $P(h_J, h_B) = (2\pi\sqrt{ST - R^2})^{-1} \exp[-(Sh_B^2 + Th_J^2 - 2Rh_B h_J)/(2(ST - R^2))]$. The result is

$$g = 1 - \frac{1}{\pi} \arccos\left(\frac{R}{\sqrt{S \cdot T}}\right). \tag{11}$$

For the 'simplest segmented case' defined through equations (8)–(10), this general result is specialized, by evaluation of S, T and R, to

$$g = 1 - \frac{1}{\pi} \arccos\left(\frac{r + c_p c_d}{\sqrt{(1 + c_p c_s)(1 + c_p c_t)}}\right) \tag{12}$$

where

$$c_d = \frac{1}{2}\left\langle \frac{\boldsymbol{B}^0 \cdot \boldsymbol{J}^1}{|\boldsymbol{B}^0| \cdot |\boldsymbol{J}^1|} + \frac{\boldsymbol{B}^1 \cdot \boldsymbol{J}^0}{|\boldsymbol{B}^1| \cdot |\boldsymbol{J}^0|} \right\rangle \tag{13}$$

is the cross correlation between the *different* segments of the student's and teacher's coupling vectors.

3. Hebbian learning

First, we briefly consider Hebbian learning, although this learning prescription generally fails for $\alpha \to \infty$ in the presence of correlations, which is not astonishing (see e.g. [15]) and strongly contrasts to Gibbs and Bayes learning (see below). However, as we will see, even in the presence of correlations the results for Hebbian learning are interesting, if the number $p := \alpha N$ of training examples is small compared with N, i.e. for $\alpha \ll 1$.

Hebbian learning is defined by the *one-shot prescription*

$$J_i = N^{-1/2} \sum_{\mu=1}^{p} \operatorname{sign}\left(\frac{\boldsymbol{B} \cdot \boldsymbol{\xi}^\mu}{\sqrt{N}}\right) \xi_i^\mu \tag{14}$$

which leads for the 'general quasisegmented case' to

$$S = \frac{\alpha}{N} \sum_{k=1}^{N} \left[(C_k^P)^2 + \frac{2\alpha}{\pi T} C_k^T (C_k^P)^3 \right] \tag{15}$$

and

$$R = \frac{\alpha}{N} \left(\frac{2}{\pi T}\right)^{1/2} \sum_{k=1}^{N} C_k^T (C_k^P)^2 \tag{16}$$

whereas T is fixed. Here C_k^T and C_k^P are the eigenvalues of the correlation matrices of equation (3). From these general results one can evaluate the generalization ability simply via equation (11). For the 'simplest segmented case' defined by equations (8)–(10) one obtains $g(\alpha)$ from equation (12); the final result for the error-probability $\epsilon = 1 - g$ is then

$$\epsilon(\alpha) = \frac{1}{\pi} \arccos \left[\frac{\alpha(1 + 2c_p c_t + c_p^2)}{\sqrt{\alpha \frac{\pi}{2}(1 + c_p c_t)^2 (1 + c_p^2) + \alpha^2 (1 + c_p c_t)(1 + 3c_p c_t + 3c_p^2 + c_p^3 c_t)}} \right]. \tag{17}$$

From this result for the 'simplest segmented case' the following general conclusions can be drawn.

- For small α, Hebbian learning is quite effective. The generalization error $\epsilon(\alpha)$ decreases rapidly with increasing α as $\epsilon(\alpha) = \frac{1}{2} - \mathrm{O}(\sqrt{\alpha})$.
- Moreover, from figure 1 one can see that the decrease of the generalization error is faster, if the correlations are 'useful' (i.e. for $c_t c_p > 0$); whether this is the case or not, does of course not depend on the student, but only on the given training examples. That is, if the choice of the training examples is the teacher's task, he (or she) should try to give examples which are *in accordance* with the spatial correlations inherent in the 'rule', such that $c_p c_t > 0$. On the other hand, what the student could do is to *monitor* the spatial correlations in the examples to obtain an estimate of c_p already for rather small α. Then by comparison of the 'monitored' values of $\epsilon(\alpha)$ and c_p with equation (17), one can estimate c_t (i.e. an important part of the rule to be discovered, which may be useful afterwards for

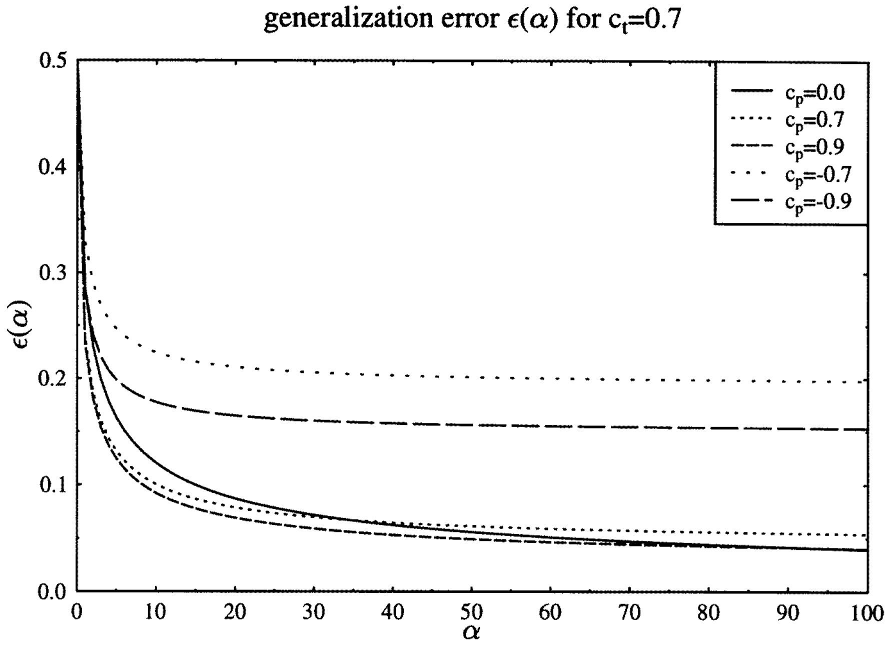

Figure 1. For Hebbian learning with a correlation parameter $c_t = 0.7$ of the two segments of the teacher perceptron, the generalization error $\epsilon(\alpha)$ is presented as a function of the reduced size $\alpha := p/N$ of the training set for different values of the pattern correlation parameter c_p.

Bayesian learning, see sections 5.2 and 5.3 below, where different priors are considered). Of course, for the 'general quasisegmented case' this may be illusionary.

- However, in the limit $\alpha \to \infty$, the error of the Hebbian learning prescription does not converge to zero, but to

$$\epsilon_\infty := \lim_{\alpha\to\infty} \epsilon(\alpha) = \frac{1}{\pi} \arccos \left[\frac{1 + 2c_p c_t + c_p^2}{\sqrt{(1 + c_p c_t)(1 + 3c_p c_t + 3c_p^2 + c_p^3 c_t)}} \right]. \tag{18}$$

This *residual generalization error* for Hebbian learning is due to the fact that the correct value for the student structure, $c_s = c_t$, is usually not achieved for $\alpha \to \infty$, although $\epsilon(\alpha)$, as obtained with the Hebb rule, decreases monotoneously with increasing α. Already at this point, we remark that, in contrast, for the Gibbs and Bayes algorithms $\epsilon(\alpha)$ always vanishes for $\alpha \to \infty$, and there the asymptotics of the limiting behaviour does not depend on the correlations at all (see below).

For the Hebbian case, the behaviour of ϵ_∞ as a function of c_p for different values of c_t is plotted in figure 2. Obviously, with Hebbian learning, correlations in the patterns usually lead to nonvanishing residual generalization error; moreover, as already mentioned, an *opposite sign* in the correlations of patterns and teacher vector, respectively, makes the learning task more difficult. (This observation will probably again transfer to more complicated networks.) Nevertheless, for fixed c_t, whatever the sign of $c_t c_p$ is, and although for sufficiently small values of $|c_p|$ the error increases $\propto |c_p|$ with increasing $|c_p|$, there is according to figure 2 finally a *decrease* down to 0 in the residual error, if $|c_p|$ increases beyond a certain value, which depends on c_t. This again is an important statement, which means that sufficiently strong spatial correlations in the patterns will almost always be

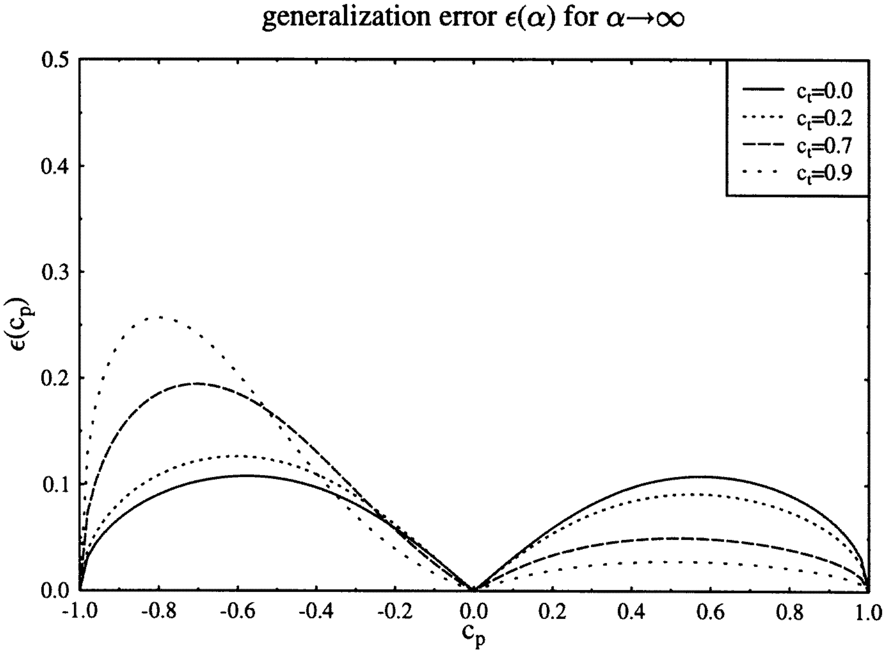

Figure 2. The limit of the generalization error for $\alpha \to \infty$ in the case of Hebbian learning is presented as a function of the pattern correlation parameter c_p for different values of the correlation c_t of the two segments of the teacher perceptron.

useful.

There are thus three limits where with Hebbian learning and fixed c_t a vanishing resisual generalization error is achieved for $\alpha \to \infty$, namely:

(i) for uncorrelated pattern spaces ($c_p = 0$); the value of c_t does not matter at all in this case, as can be seen already from equation (17), since then $\epsilon(\alpha) = \pi^{-1} \arccos[1 + (\pi/2\alpha)]^{-1/2}$, which vanishes for $\alpha \to \infty$ as $\epsilon = 1/\sqrt{2\pi\alpha}$;

(ii) for $c_p = \pm 1$, with $c_t \neq (-c_p)$; in this case the pattern segments are identical up to ± 1; this corresponds to an effective reduction of N to $N/2$, i.e. to a doubling of α, but otherwise the same result as for (i);

(iii) for $c_t = \pm 1$, with $c_p \neq (-c_t)$; in this case one has

$$\epsilon(\alpha) = \pi^{-1} \arccos \left\{ 1/\sqrt{1 + \pi(1 + c_p^2)/[2\alpha(1 \pm c_p)^2]} \right\}$$

which behaves for $\alpha \to \infty$ as $\sqrt{1 + c_p^2}/[\sqrt{2\pi\alpha}(1 \pm c_p)]$.

In contrast to (ii) and (iii), if c_t is not kept fixed, but if the point $(c_p, c_t) = (-1, 1)$ or $(1, -1)$ is approached with fixed slope $\partial c_t/\partial c_p = -x$, then, according to equation (18), the residual error ϵ_∞ is a decreasing function of x for $0 < x < \infty$, with $\epsilon_\infty = \frac{1}{2}$ (which corresponds to zero generalization ability) for $x = 0^+$, via $\epsilon_\infty = \frac{1}{4}$ for $x = 1$, to $\epsilon_\infty = 0$ for $x \to \infty$. At $x \equiv 0$, where ϵ_∞ vanishes, there is thus a discontinuity.

Except (i), these are just pretty artificial cases, so the Hebb rule fails, if correlated patterns are to be learned.

For $c_t = 0$, we have found that even the *modified* Hebb prescription of [14], which corresponds to the matrix transformation $\boldsymbol{J} \to \mathbf{K} \cdot \boldsymbol{J}$ with $\mathbf{K} = (\mathbf{C}^P + \nu \mathbf{I})^{-1}$, where the pattern correlation matrix $\mathbf{C}^P$ is given by equation (3), while $\mathbf{I}$ is the $N \times N$ unit matrix and

ν is an optimization parameter, would yield at most a $\sim 30\%$ reduction of the generalization error $\epsilon(\alpha)$, although for $\nu = \sqrt{1 - c_p^2}$ also c_s vanishes.

4. Gibbs learning

In the case of Gibbs learning, the student perceptron is drawn at random from the so-called *version space* $\mathcal{V}$, which consists exactly of all perceptrons which classify the training examples correctly. Tarkowski and Lewenstein [6], treated storage and generalization of spatially and semantically correlated patterns in perceptrons, but only for the special case of Gibbs learning with *uncorrelated* teacher couplings ($\mathbf{C}^T = \mathbf{I}$ in equation (3)). We extend their approach to $\mathbf{C}^T \neq \mathbf{I}$ and correct some of their results (see below), using Gardner's [16, 17] replica method. With the teacher field $u_t := N^{-1/2}\boldsymbol{B} \cdot \boldsymbol{\xi}$ ($= h_B$ in equation (1)) and the different student fields $u_a := N^{-1/2}\boldsymbol{J}^a \cdot \boldsymbol{\xi}$, where $a = 1, 2, \ldots, n$ enumerates the replicas, one obtains for general quasisegmentation with equations (6) and (7) the following order parameters:

$$T := \langle u_t^2 \rangle = N^{-1} \sum_{k=1}^{N} C_k^P \tilde{B}_k^2 \tag{19}$$

$$R_a := \langle u_t u_a \rangle = N^{-1} \sum_{k=1}^{N} C_k^P \tilde{B}_k \tilde{J}_k^a \tag{20}$$

$$S_a := \langle u_a^2 \rangle = N^{-1} \sum_{k=1}^{N} C_k^P (\tilde{J}_k^a)^2 \tag{21}$$

$$Q_{ab} := \langle u_a u_b \rangle = N^{-1} \sum_{k=1}^{N} C_k^P \tilde{J}_k^a \tilde{J}_k^b. \tag{22}$$

Here the C_k^P are again the eigenvalues of the pattern correlation matrix $\mathbf{C}^P$, while $\tilde{B}_k$ and $\tilde{J}_k^a$ are the components of $\boldsymbol{B}$ (resp. $\boldsymbol{J}^a$) in the corresponding basis; the fields u_t and u_a can be generated from normally distributed, independent variables w, v_t and v_a by

$$u_t = \sqrt{T - \frac{R^2}{Q}}\, v_t - \frac{R}{\sqrt{Q}} w \tag{23}$$

$$u_a = \sqrt{S - Q}\, v_a - \sqrt{Q} w. \tag{24}$$

The general result for the free energy, evaluated with the replica trick assuming replica symmetry, which is exact in this case, is $F = \mathrm{Extr}\,[F_1 + \alpha F_2]$, where the *energy term* F_2 is

$$F_2 = 2 \int \mathrm{D}w\, H(x_1) \ln H(x_2) \tag{25}$$

with $x_1 := Rw(TQ - R^2)^{-1/2}$ and $x_2 = w(Q/(S - Q))^{1/2}$, where $\mathrm{D}w := (2\pi)^{-1/2}\,\mathrm{d}w\, \exp(-w^2/2)$ and $H(x) := \int_x^\infty \mathrm{D}w$. The *entropy term* F_1 is given by

$$F_1 = \ln(2\pi) - N^{-1} \sum_{k=1}^{N} \left\{ \ln[E + (F + H)C_k^P] + \frac{FC_k^P + G^2 (C_k^P)^2 C_k^T}{E + (F + H)C_k^P} \right\} + \frac{E}{2} + GR + \frac{HS + FQ}{2}. \tag{26}$$

Here E, F, G and H are additional order parameters conjugate to $|\boldsymbol{J}|$, Q, R and S, so that in all (since $|\boldsymbol{J}|$ is fixed) F has to be optimized for seven order parameters.

For our ‘simplest segmented systems’, see equations (8)–(10), the general results from equations (20)–(22), see also (11)–(13), specialize to

$$R_a = r^a + c_p c_d^a \qquad S_a = 1 + c_p c_s^a \qquad Q_{ab} = q^{ab} + c_p q_d^{ab} \tag{27}$$

with

$$\begin{aligned} &r^a = N^{-1}\boldsymbol{B}\cdot\boldsymbol{J}^a \qquad q^{ab} = N^{-1}\boldsymbol{J}^a\cdot\boldsymbol{J}^b \qquad c_s^a = 2N^{-1}\boldsymbol{J}^{0a}\cdot\boldsymbol{J}^{1a} \\ &c_d^a = N^{-1}(\boldsymbol{B}^0\cdot\boldsymbol{J}^{1a} + \boldsymbol{B}^1\cdot\boldsymbol{J}^{0a}) \\ &q_d^{ab} = N^{-1}(\boldsymbol{J}^{0a}\cdot\boldsymbol{J}^{1b} + \boldsymbol{J}^{1a}\cdot\boldsymbol{J}^{0b}). \end{aligned} \tag{28}$$

In view of the free energy, with the saddle-point approach and again with the replica symmetry assumption, the *entropy term* specializes to

$$\begin{aligned} F_1 = \frac{1}{2}\{&\ln(2\pi) + \ln[(q-1-q_d+c_s)(q-1+q_d-c_s)]\} \\ &- \frac{1-q+q_d c_s - c_s^2}{(q-1-q_d+c_s)(q-1+c_d+c_s)} \\ &+ \frac{(r^2+c_d^2)(q-1+c_t q_d - c_t c_s) + 2r\,c_d(c_t - q c_t + c_s - q_d)}{(q-1-q_d+c_s)(q-1+q_d+c_s)(1-c_t^2)} \end{aligned} \tag{29}$$

which depends only on the five parameters r, c_s, c_d, q, and q_d, but not on c_p, whereas the *energy* contribution specializes to

$$F_2 = 4\int \mathrm{D}w\, H(x_1 w)\ln H(x_2 w) \tag{30}$$

with

$$x_1 = \frac{r + c_p c_d}{\sqrt{(1+c_p c_t)(q+c_p q_d) - (r+c_p c_d)^2}} \tag{31}$$

$$x_2 = \sqrt{\frac{q + c_p c_d}{1 + c_s c_p - (q + c_p q_d)}}. \tag{32}$$

Using the conditions $\partial F/\partial r = \partial F/\partial c_s = \partial F/\partial c_d = \partial F/\partial q = \partial F/\partial q_d = 0$ one obtains the evolution of all interesting quantities.

A major difference to the Hebb case can be seen from the asymptotic behaviour for $\alpha \to \infty$. For unstructured teacher perceptron ($c_t = 0$), the entropy term can be simplified, since then $q = r$, $q_d = c_d$ and $c_s = 0$. So one obtains asymptotically $c_d \to c_p \cdot (1 - r)$ and

$$r \to 1 - \frac{1}{\alpha^2 C^2 (1 - c_p^2)} \tag{33}$$

with $C = (2\pi)^{-1/2}\int \mathrm{d}x\, H(x)\ln H(x) \approx -0.360\,324$.

Thus, with Gibbs learning in the case $c_t = 0$ a perfect overlap, and thus perfect generalization, is reached for all values of c_p, in contrast to Hebbian learning, where this was only the case when $c_p = 0$ (except some limiting cases, see above). However, the prefactor of the $1/\alpha^2$-behaviour of equation (33) is proportional to $(1 - c_p^2)$, which means that asymptotically for the overlap r, but not for the generalization ability itself (see below), spatial pattern correlations are still slightly detrimental for the Gibbs case with $c_t = 0$, but only for the just-mentioned prefactor, whereas the ‘residual error’ itself now vanishes, in contrast to the Hebb case.

Let us now concentrate on the generalization error. In figure 3 this quantity is plotted for several values of $|c_p|$ ($c_t = 0$ fixed), showing that the error becomes smaller with

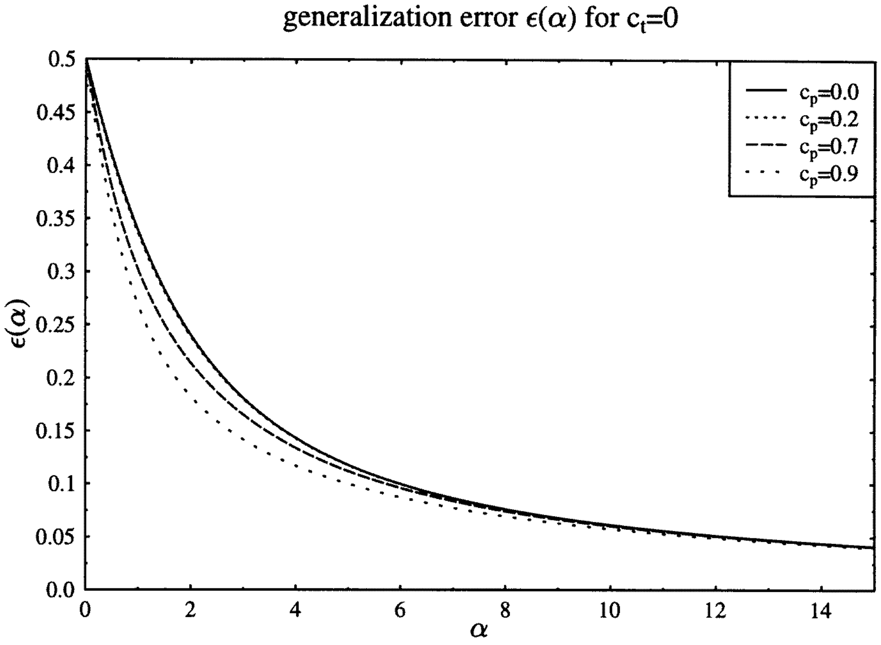

Figure 3. For the case of Gibbs learning, the generalization error $\epsilon(\alpha)$ is presented as a function of the reduced size $\alpha := p/N$ of the training set, for $c_t = 0$ and different values of $|c_p|$.

increasing $|c_p|$ for all α. In other words, the more structured the pattern space, the easier it is to actually learn the classification task given by the teacher rule. This is in contrast to the behaviour of r (see above) but intuitively reasonable, and can be understood slightly more thoroughly by the following consideration.

If we perform a coordinate transformation in the pattern and phase space to diagonalize the correlation matrix (of the patterns) we have two eigenvalues $1 \pm c_p$ determining the variance of the corresponding sites. This means that the sites with $1-|c_p|$ are less significant than those with $1+|c_p|$. Thus, the student can concentrate on the $N/2$ latter ones to learn the task. Since these are only half as many as the whole set, learning can be performed faster. In the extreme case of $|c_p| = 1$ the dimension of the system is effectively reduced to $N/2$, leading to a rescaling of α with the factor of 2. It is clear that this reasoning can be transferred to more general segmentations and more complex architectures.

The above considerations provide an alternative view on the learning problem investigated here as well, i.e. pattern sets which can be decomposed into components of different magnitude. Data preprocessing using principal component analysis techniques uses such structures in practical applications [11, 12]. Thus, correlations should be helpful in general.

Nevertheless, looking at the asymptotic behaviour for $\alpha \to \infty$ of the generalization error, which for Gibbs learning with $c_t = 0$ is

$$\lim_{\alpha\to\infty} \varepsilon(\alpha) = \frac{1}{\pi}\arccos(r + c_p c_d) = \frac{1}{\pi}\arccos\left(1 - \frac{1}{\alpha^2 C^2}\right) \approx \frac{0.625}{\alpha} \tag{34}$$

we have a result which is independent of the pattern correlations at all. So, for large α, structure in the patterns has no advantage in terms of the generalization error. Actually, the fact that the improved generalization ability due to structure in the pattern space is confined

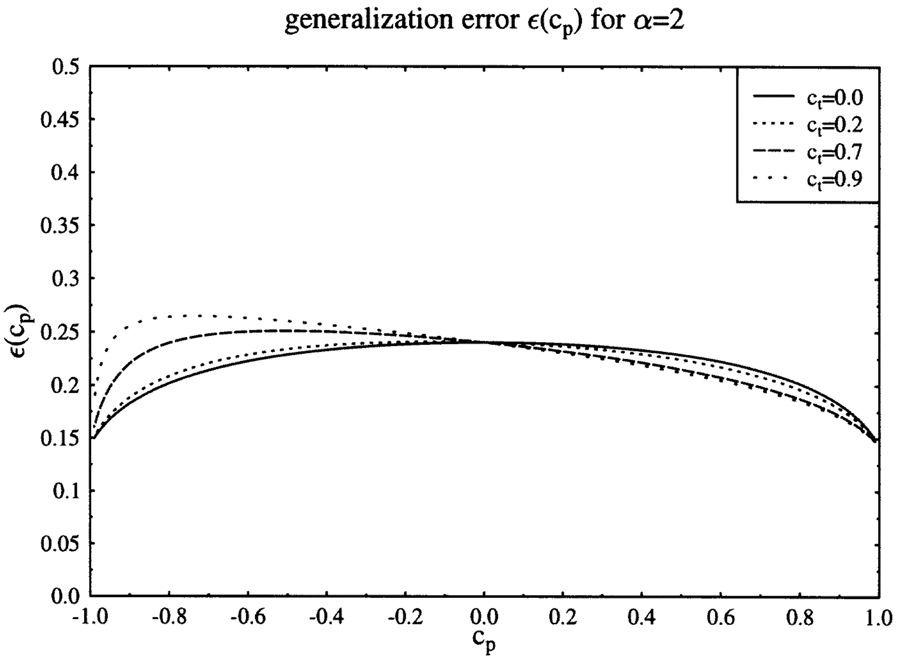

Figure 4. For the case of Gibbs learning and $\alpha = 2$, the generalization error $\epsilon(c_p)$ is presented as a function of the pattern correlation parameter c_p for different values of c_t.

to an intermediate α-regime can easily be understood. To reach perfect generalization, the sites with eigenvalue $1 - |c_p|$, which are less significant at first, become important for $\alpha \to \infty$ to achieve the ultimate 'fine adjustment'.

In the case of a correlated teacher vector ($c_t \neq 0$) things change slightly. Figure 4 shows the dependence of the generalization error on c_p for several values of c_t and fixed $\alpha = 2$ (which is something like an intermediate value). We see that structure in the patterns can actually worsen the generalization ability, if the structure is in the opposite direction to the teacher correlation, i.e. for $c_p c_t < 0$. This resembles the behaviour of the Hebb rule, where such a type of learning problems are also difficult, and again the result can probably be transferred to more general situations.

Looking at the simultaneously diagonalized correlation matrices the reason for this becomes clear. Sites with the smaller variance $1 - |c_p|$, concerning the patterns, are related to teacher sites with the larger eigenvalue $1 + |c_t|$, and therefore their loss in significance (due to a small value $1 - |c_p|$) is somehow compensated by the larger weights of the teacher vector.

Although not analytically shown, we expect from numerical evidence perfect generalization in the limit $\alpha \to \infty$ to be achieved for $c_t \neq 0$ as well, again with the law given in (34). This means that correlations in the system *asymptotically* neither improve nor worsen the generalization behaviour if one uses good enough learning rules.

Let us now break down the behaviour into the contributions from the several order parameters. Figures 5(a) and (b) show the evolution of $r(\alpha)$ for different values of c_p with $c_t = 0$ and $c_t = 0.9$, respectively. For $c_t = 0$ a higher correlation $|c_p|$ leads to a smaller overlap. For $c_t = 0.9$ the behaviour depends on the sign of c_p as well. For small α the overlap $r(\alpha)$ is larger for $c_p c_t > 0$ than for $c_p c_t < 0$; but for larger values of α the relation

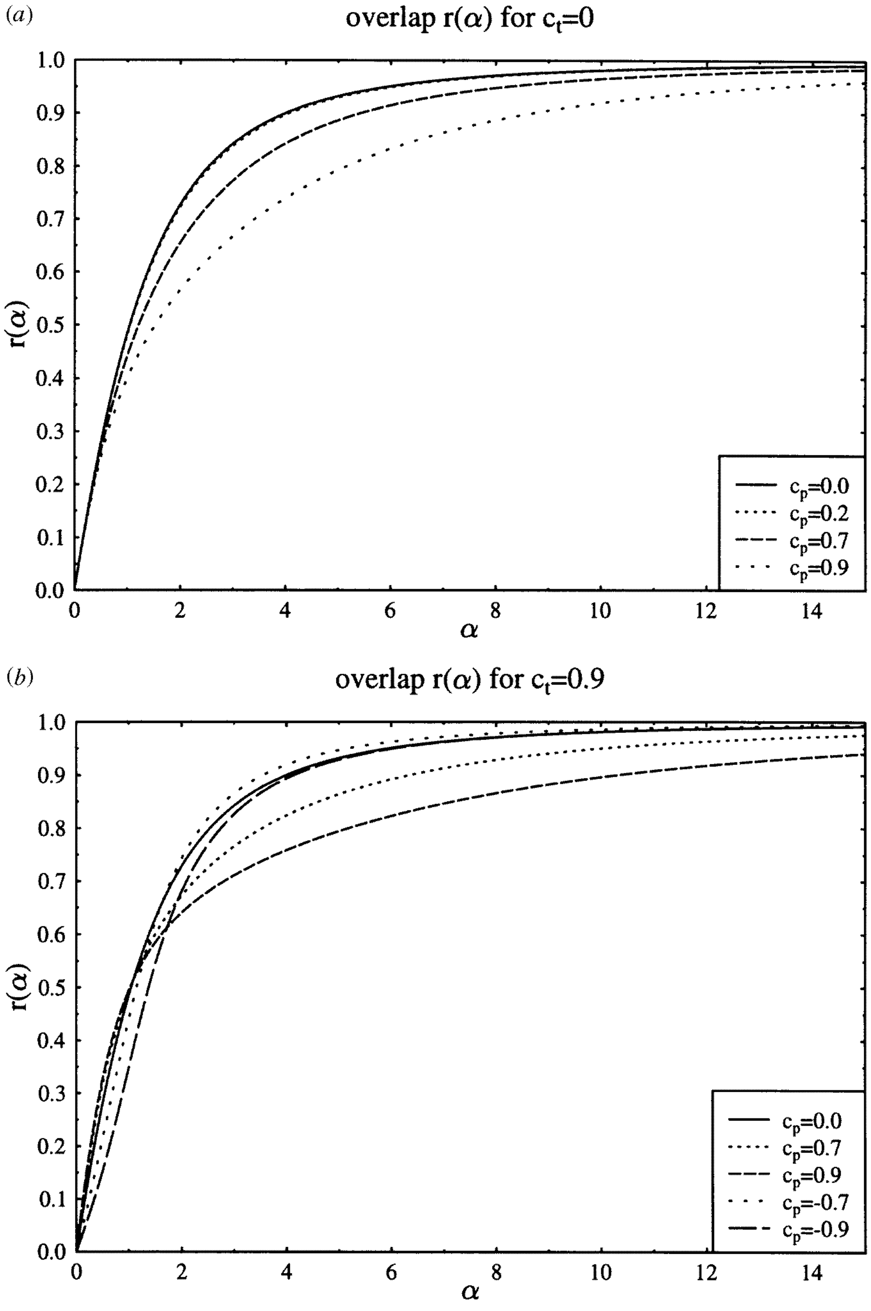

Figure 5. (*a*), (*b*) For Gibbs learning with $c_t = 0$ and $c_t = 0.9$, respectively, the normalized overlap $r(\alpha)$ of the coupling vectors of the teacher's and student's perceptron is presented as a function of the reduced size $\alpha := p/N$ of the training set.

is *opposite*. To understand this 'crossing behaviour' we have to notice that the magnitude of the local fields, and so of the stability of the patterns, is enhanced (reduced) for $c_pc_t > 0$ ($c_pc_t < 0$).

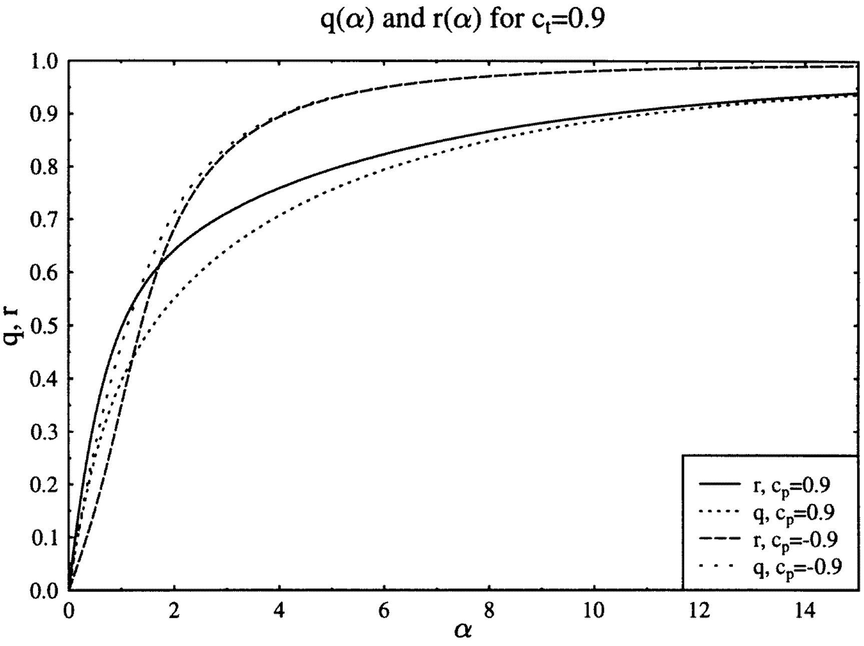

Figure 6. For Gibbs learning with $c_t = 0.9$, the order parameters $q(\alpha)$, which is the typical overlap between the coupling vectors of two different student perceptrons, and $r(\alpha)$, which is the overlap between the coupling vectors of a typical student and the teacher, are presented as a function of the reduced size $\alpha := p/N$ of the training set for the two cases of $c_p = \pm 0.9$.

• For $\alpha \ll 1$, a small stability (small on average) merely leads to a small bias of the version space away from the true teacher vector (since the training patterns lie near the classification boundary). The direction of this small bias is naturally such that the $c_p c_t > 0$ case yields higher overlap.

• For $\alpha \gg 1$ the biasing effect of the small stability disappears, since the patterns cover the space somehow dense. On the other hand, for $c_p c_t > 0$ the phase space of the solutions is now more confined (q is smaller) because of the constraint of a higher stability (see the evolution of $q(\alpha)$ in figure 6) of the possible solutions. This leads to a smaller overlap r for $c_p c_t < 0$ in case of $\alpha \gg 1$.

Figure 7 shows the evolution of the student structure $c_s(\alpha)$ for a teacher correlation $c_t = 0.9$. It is interesting to see that opposite correlations in the patterns (compared with the teacher) forces the student to adopt the teacher structure rather rapidly with a similar explanation as given above for the evolution of $r(\alpha)$.

The evolution of $c_d(\alpha)$ with c_p (figure 8 for the case $c_t = 0$) is nonmonotonic, which generally occurs if $|c_p| > |c_t|$. Asymptotically of course, the value $c_d = c_t$ is approached. So a high correlation in the patterns (e.g. $c_p \gtrsim 0.7$) induce strong correlations of $c_d(\alpha)$ in an intermediate region around $\alpha \sim 1$, which improve (worsen) the generalization ability in this regime for $c_p c_t > 0$ ($c_p c_t < 0$).

Finally we should mention that the *independence* of $\epsilon(\alpha \to \infty)$ on the pattern correlations c_p, which we have shown analytically in equation (34) for $c_t = 0$, corrects a different result of Tarkowski and Lewenstein [6]. For $c_t \neq 0$ and $c_p \neq 0$, because of the large number of order parameters, we did not yet succeed in calculating the limiting behaviour analytically, although it is probably unchanged. Again, in view of the results of

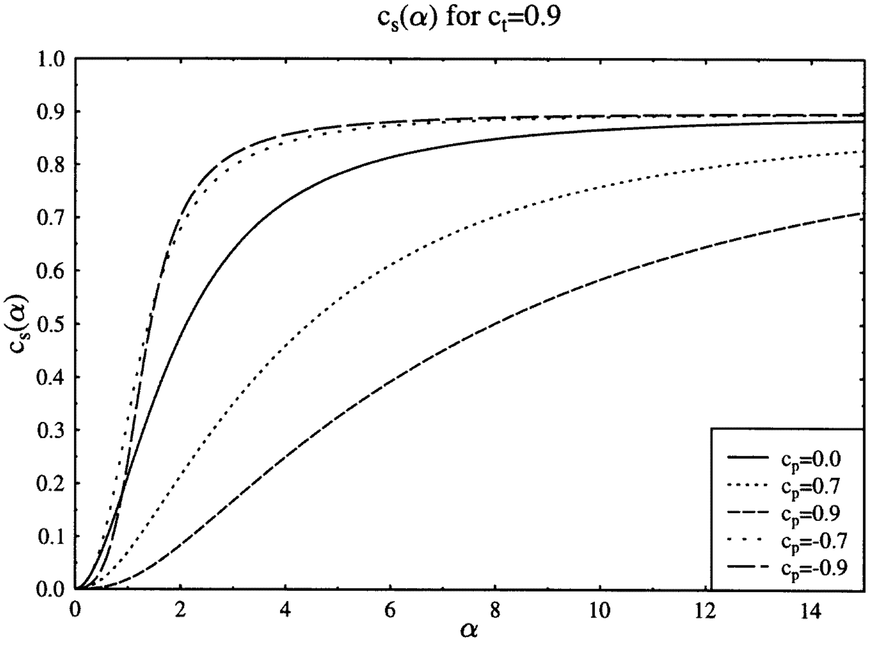

Figure 7. For Gibbs learning, the evolution of the correlation parameter $c_s(\alpha)$ between the two segments of the student perceptron, as it develops as a function of the reduced size $\alpha := p/N$ of the training set, is presented over α for $c_t = 0.9$ and $c_p = 0$, ±0.7 and ±0.9.

[7, 8], the result should also apply to the more complicated multilayer architectures treated in these papers, and should also be valid in the presence of certain classes of noise.

In the following section we treat Bayesian learning with different priors, while the results for AdaTron learning, which leads to maximal stability but not to optimal generalization, will be discussed in a separate paper.

5. Bayesian learning

Bayesian methods are succesfully used for learning in neural networks, see [18, 19, 12]. In this approach a pattern is classified with the purpose to minimize the probability of a 'wrong answer'. The framework requires the specification of a prior belief about the possible networks and a noise model defining their answer behaviour.

More precisely, the *noise model* $p(s|\boldsymbol{J}, \boldsymbol{\xi})$ defines the conditional probability of getting the answer s (correct or not) on a given pattern $\boldsymbol{\xi}$ for a general classifying automaton $\boldsymbol{J}$ ranging over some sample space. The probability $p(D|\boldsymbol{J})$ of the data D comprising the whole training set is typically given by simply *multiplying* all probabilities for the single members of the training set, i.e. pairs of training-questions with 'correct answers', thus assuming that these pairs are given independently of each other, i.e. without semantical correlations, whereas *spatial* correlations may be included.

The so-called *prior* $p(\boldsymbol{J})$ defines the probability that the vector $\boldsymbol{J}$ describes the automaton, before the evidence of any data is taken into account, i.e. on the basis of

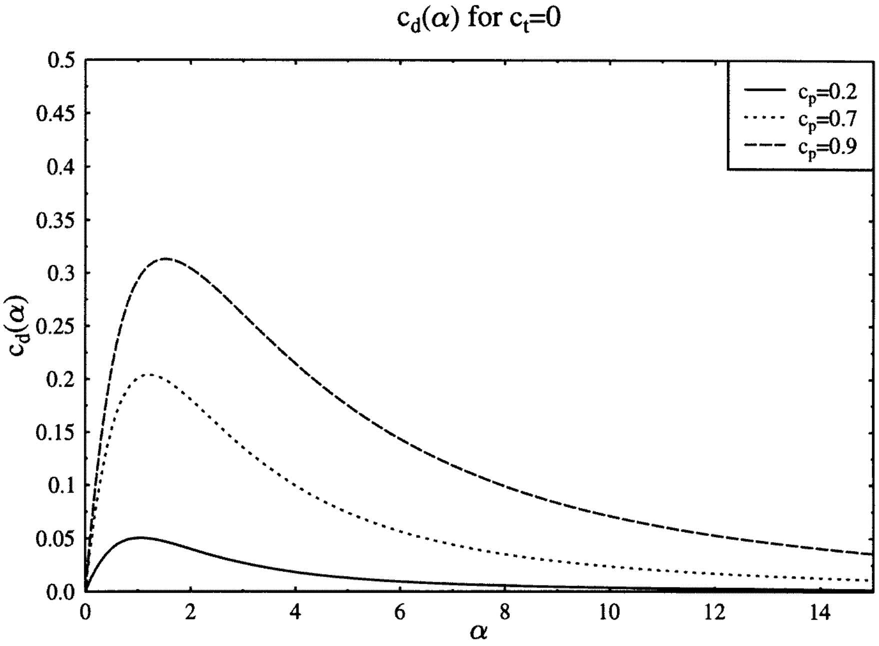

Figure 8. For Gibbs learning, the evolution of the cross-correlation parameter c_d of two different segments of the teacher's and student's perceptron, see equation (13), as it develops as a function of the reduced size $\alpha := p/N$ of the training set, is presented over α for $c_t = 0.9$ and $c_p = 0.2$, 0.7 and 0.9.

some prior knowledge. Using the Bayes theorem we obtain

$$\mathcal{P}(\boldsymbol{J}|D) = \frac{p(\boldsymbol{J})p(D|\boldsymbol{J})}{\mathcal{P}(D)} \tag{35}$$

as the *a postiori probability* of $\boldsymbol{J}$ after absorbing the evidence of the training data. Here the so-called *evidence of the model* $\mathcal{P}(D) := \sum_{\boldsymbol{J}} p(\boldsymbol{J})\, p(D|\boldsymbol{J})$ serves for normalization. The 'most probable correct answer' s' on a *test-question* $\boldsymbol{\xi}'$ is then given by the weighted majority vote due to $\mathcal{P}(\boldsymbol{J}|D)$ from (35). Here again we assume that the same spatial correlations C_{ij}^P, see equation (3), apply to both the training-questions and to the test-questions, while in both cases the 'correct answers' are given by the same 'teacher automaton' $\boldsymbol{B}$, which is not specified explicitly in equation (35) and principally can have an architecture different from that of the 'student automaton' $\boldsymbol{J}$ (although in our case we assume the same architecture). Of course, we also assume that the 'student' uses the same 'noise model' for both training and afterwards.

In practice, a good choice of the noise model and the prior (which include the choice of the architecture used) is a crucial point for getting good generalization behaviour. One possibility for proper model selection is to calculate the 'evidence' of several possible models [18, 19].

Methods from statistical mechanics can be used to investigate systems in the thermodynamic limit, see [20], and concerning model selection [21]. The purpose of this section is to compare the behaviour of Bayesian learning to Gibbs learning in the case of structured spaces on the one hand, and to investigate the influence of different priors on the other. As priors we use:

(1) a *uniform prior* over all normalized student coupling vectors;

(2) a *restricted prior* permitting only those student weight vectors, which have the correct (and in this case assumed as known) correlation $c_s = c_t$. (If a sufficient number of training examples is given, the 'student' can get knowledge of c_t by monitoring the spatial statistics c_p of the questions posed by the teacher and applying Hebbian learning for some time, i.e. for finite α, see above.)

Since we are considering here a deterministic classification, the appropriate 'noise model' gives probability 1 for the correct answer (due to the coupling vector $\boldsymbol{J}$ and the perceptron mapping rule) and 0 otherwise.

One should stress that these choices contain a rather large amount of prior knowledge about the possible teacher rules which is not in the same way available in practical problems.

5.1. Relation to the Gibbs case

In this case it is rather easy to derive the Bayes properties from the already calculated quantities for the Gibbs case. This is possible since one can construct a *perceptron* from the Gibbsian version space $\mathcal{V}$ which performs like the Bayesian classification, namely the *central-point (CP) perceptron*. If the M members $\boldsymbol{J}_l$ of the version space carry identical *a priori* probabilities, the CP-perceptron is simply

$$\boldsymbol{J}^{\mathrm{CP}} = \lim_{M\to\infty} \frac{1}{K} \sum_{l=1}^{M} \boldsymbol{J}_l. \tag{36}$$

Here K is chosen, such that $|\boldsymbol{J}^{\mathrm{CP}}| = N^{1/2}$. Therefore

$$K^2 = \lim_{M\to\infty} \frac{1}{N} \sum_{l,m=1}^{M} \boldsymbol{J}_l \cdot \boldsymbol{J}_m = \lim_{M\to\infty} [M + M(M-1)\cdot q]. \tag{37}$$

So one gets for the overlap

$$r^{\mathrm{CP}} = \frac{\boldsymbol{J}^{\mathrm{CP}} \cdot \boldsymbol{B}}{N} = \lim_{M\to\infty} \frac{1}{NM\sqrt{q}} \sum_{l=1}^{M} \boldsymbol{J}_l \cdot \boldsymbol{B}. \tag{38}$$

Since $r := \lim_{M\to\infty}(NM)^{-1} \sum \boldsymbol{J}_l \cdot \boldsymbol{B}$ is the overlap for the case of Gibbs learning, we have in this way the simple relations

$$r^{\mathrm{CP}} = \frac{r}{\sqrt{q}} \qquad c_d^{\mathrm{CP}} = \frac{c_d}{\sqrt{q}}. \tag{39}$$

Additionally, one needs the correlation between the two different segments of the CP student perceptrons:

$$\begin{aligned} c_s^{\mathrm{CP}} &= \frac{2}{N} (\boldsymbol{J}^{\mathrm{CP}})^0 \cdot (\boldsymbol{J}^{\mathrm{CP}})^1 = \lim_{M\to\infty} \frac{2}{NM^2 q} \sum_{l,m=1}^{M} \boldsymbol{J}_l^0 \cdot \boldsymbol{J}_m^1 \\ &= \frac{Mc_s + M(M-1)q_d}{M^2 q} \to \frac{q_d}{q}. \end{aligned} \tag{40}$$

The fact that the CP perceptron reaches the same generalization ability as the Bayes classification follows from

$$\sigma^{\mathrm{CP}} = \mathrm{sign}\left(\frac{1}{M\sqrt{qN}} \sum_{l=1}^{M} \boldsymbol{J}_l \cdot \boldsymbol{\xi} \right) = \mathrm{sign}(\langle h_J \rangle) \tag{41}$$

and

$$\sigma^{\text{Bayes}} = \text{sign}\left[\sum_{l=1}^{M} \text{sign}\left(\frac{\boldsymbol{J}_l \cdot \boldsymbol{\xi}}{\sqrt{N}}\right)\right] = \text{sign}(\langle \text{sign}(h_J) \rangle). \tag{42}$$

In [1, 22, 20] it was proved for the case of vanishing pattern—and teacher—correlations ($c_p = c_t = 0$) that the generalization abilities obtained with the CP perceptron, equation (41), and the corresponding Bayes algorithm, equation (42), respectively, agree for almost all $\boldsymbol{\xi}$ in the limit $M \to \infty$, where additionally $M \ll N$ is assumed. Probably the agreement of the generalization abilities is also true, if pattern- and teacher-correlations are included.

We mention hear that for *two-layer perceptrons*, in contrast to this case, the CP automaton does *not* reach the generalization ability of the Bayes process, except for the parity machine. The reason for this exception is due to the 'chequered' structure of the mapping in the second layer of the parity machine (each flip of the output of only one hidden node changes the final classification from (+1) to (−1) and *vice versa*). This leads to the fact that for the parity machine exploring the phase-space *around* the CP solution by the Bayesian method gives just the same result as the CP solution itself. The interested reader will find more details in [7].

In the following we call the CP solution 'CP_1-perceptron' if the uniform prior (1) is used, 'CP_2-perceptron' if only students with structure $c_s = c_t$ are permitted, prior (2).

5.2. *Uniform prior*

The learning curves $\epsilon(\alpha)$ for this prior are shown in figure 9 for several c_p and $c_t = 0$. For comparison the performance of the Gibbs algorithm is shown as well ($c_p = 0$, Gibbs).

The improvement compared with Gibbs learning is significant and remains asymptotically, i.e. one obtains for $c_t = 0$ (and probably also for $c_t \neq 0$) a behaviour again independent from c_p, namely [20]:

$$\lim_{\alpha\to\infty} \epsilon^{\text{Bayes}}(\alpha) \approx \frac{0.44}{\alpha}. \tag{43}$$

The influence of pattern correlations is similar to the Gibbs case.

Figures 10(*a*) and (*b*) present results for the overlap $r(\alpha)$ between teacher- and CP_1-perceptron for $c_t = 0$ and $c_t = 0.9$ as a function of c_p. Here, one finds a similar behaviour to the preceding section, but now somewhat more pronounced, namely: (i) for $c_t = 0$ the overlap decreases with increasing c_p; (ii) for $c_t \neq 0$ there is a *crossing* of the results near $\alpha \sim 2$, and (iii) different signs of c_p and c_t lead to higher values of r for large α; probably this behaviour generalizes again to multilayer networks, see [7, 8]. Figure 11 deals with the internal structure of the CP_1-perceptron, i.e. the internal overlap $c_s(\alpha)$ of its two segments is presented, again for $c_t = 0.9$, for various values of c_p. For $\alpha \to \infty$, $c_s(\alpha)$ converges to the internal structure of the teacher perceptron, i.e. $c_s(\alpha) \to c_t$. The most prominent difference to the case of Gibbs learning is that here in the opposite limit $\alpha \to 0$ the CP_1-perceptron takes the value of the spatial correlation of the *patterns*, i.e. $c_s(\alpha \to 0) \to c_p$. This has already been observed with the Hebb rule, see above, and also with maximal-stability learning [14], in connection with the simpler *storage problem*.

5.3. *Restricted prior*

Now let us look at the result if an enhanced prior knowledge is given, i.e. the internal structure c_t of the teacher. The Bayesian inference based on this prior has the best possible

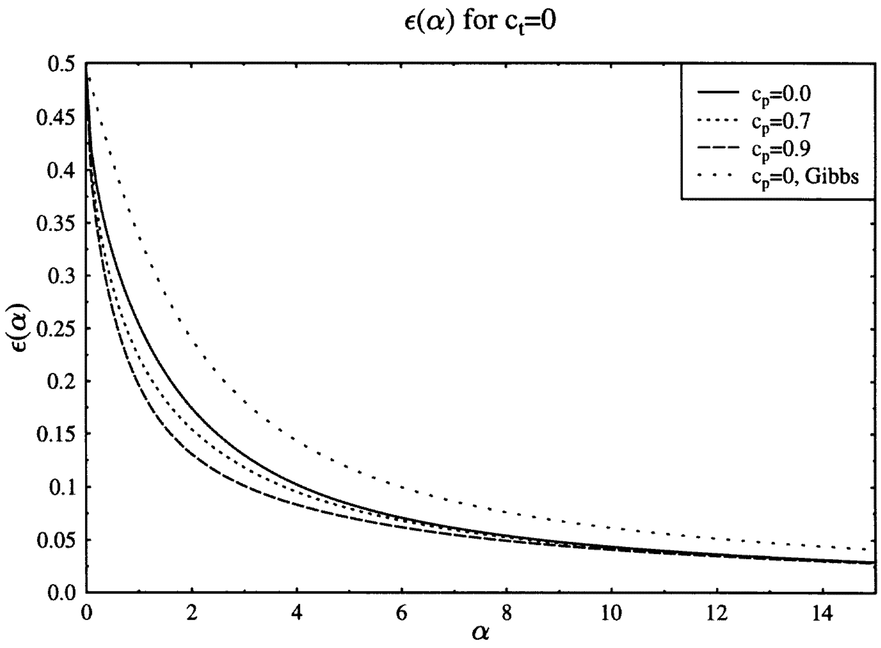

Figure 9. For Bayesian learning with *uniform prior*, i.e. the CP_1 perceptron, and $c_t = 0$, the generalization error $\epsilon(\alpha)$ is presented over the reduced size $\alpha := p/N$ of the training set, for $c_p = 0$, 0.7 and 0.9, and for comparison also for Gibbs learning with $c_p = 0$.

generalization performance since all available prior knowledge is used to minimize the error probability.

In the averaging process defined by equation (36) only those members $\boldsymbol{J}_l$ of the version space are now taken into account, which fulfil the constraint $c_s = c_t$, i.e. which have the same correlation between the segments as the teacher. In this case the teacher is a typical member of the restricted version space, so we have $q = r$ and $q_d = c_d$. Thus, the expression for the free energy simplifies for the CP_2-perceptron with equations (29) and (30) to

$$F = \mathrm{Extr}_{r,c_d}\left\{\frac{1}{2}[\ln(2\pi) + \ln((r-1+c_t-c_d)(r-1-c_t+c_d))] \right.$$
$$\left. +1 + \frac{c_t c_d - r}{c_t^2 - 1} + 4\alpha\int \mathrm{D}w\, H(x)\ln H(x)\right\} \tag{44}$$

with $x = (r + c_p c_d)^{1/2}(1 + c_p c_t - r - c_p c_d)^{-1/2}$. Extremizing with respect to r and c_d, one obtains the quantities describing $\mathcal{V}$ in this case, and from them the behaviour of the CP_2-perceptron.

To check the performance we choose a high teacher correlation, $c_t = 0.9$. (Clearly, for smaller c_t the expected advantage should decrease, since the actual restriction of the prior by imposing $c_s = c_t$ is reduced.) The results in figure 12 show the performance of the CP_1 and CP_2 perceptron as a function of α for $c_t = 0.9$. For intermediate values of α, we observe in fact a quite big improvement of the CP_2 results with respect to the CP_1 case.

However, it can be shown [23] that again asymptotically for $\alpha \to \infty$ the results are the same as for the uniform prior (1). This is a well known effect in Bayesian learning: for large sizes of the training set the evidence of the examples dominates the influence of the prior, which becomes increasingly irrelevant (as long as—in our case—the correct teacher

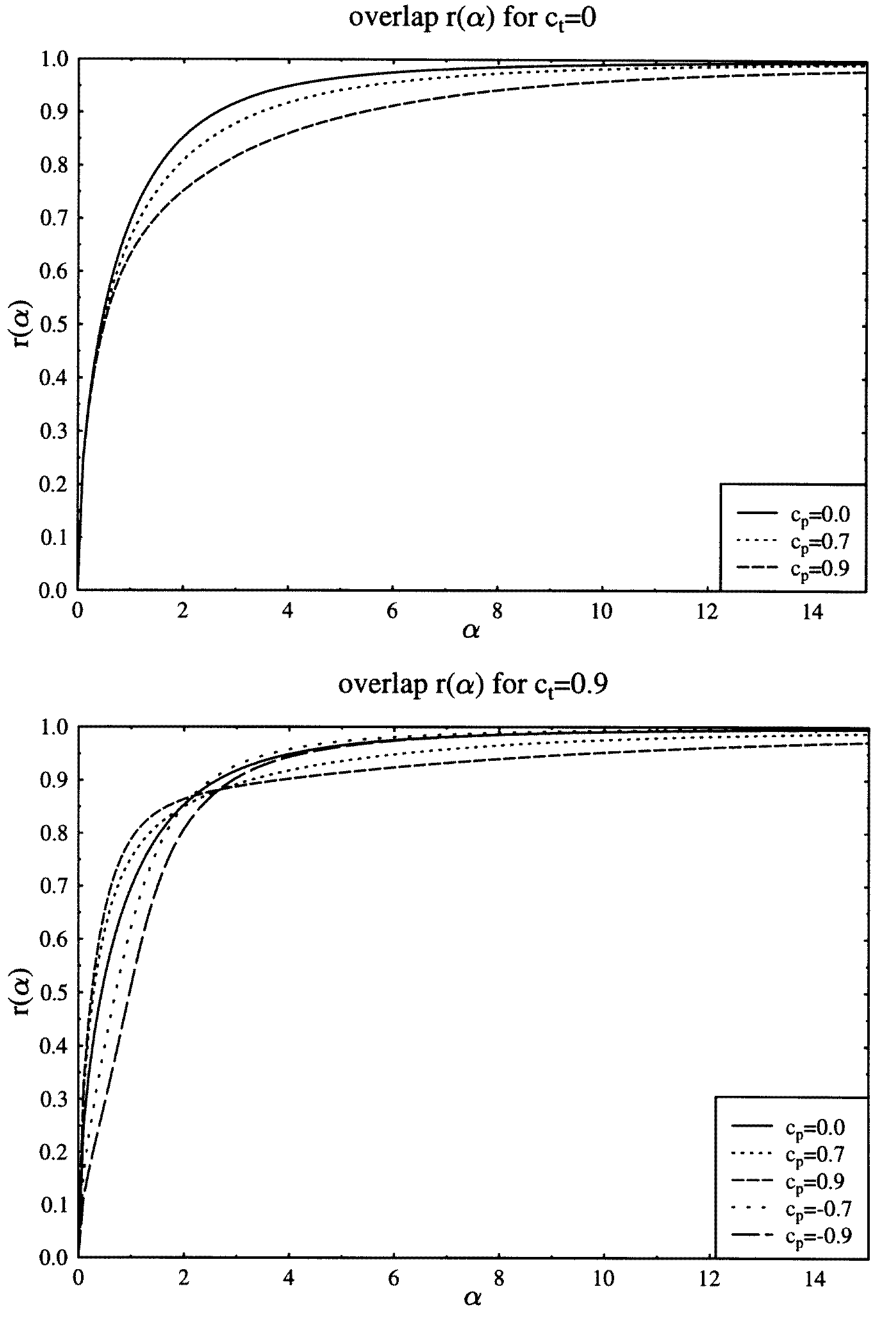

Figure 10. (*a*), (*b*) For Bayesian learning with *uniform prior*, i.e. the CP$_1$ perceptron, for the two cases $c_t = 0$ and $c_t = 0.9$, the overlap $r(\alpha)$ between the coupling vector of the teacher and the CP$_1$ student perceptron is presented as a function of the reduced size $\alpha := p/N$ of the training set, for pattern correlations $c_p = 0$, ± 0.7 and ± 0.9.

rule is included with finite probability).

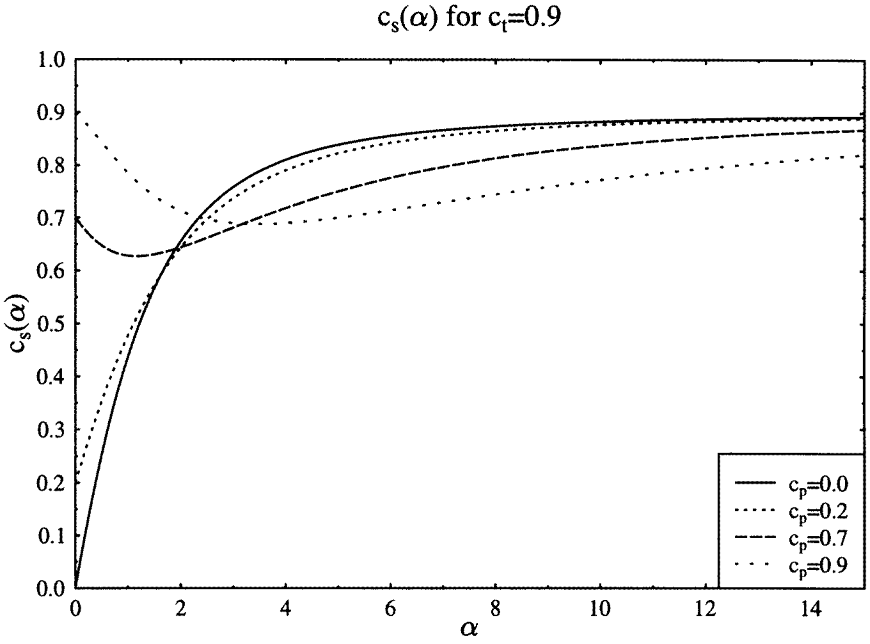

Figure 11. For Bayesian learning with *uniform prior*, i.e. the CP$_1$ perceptron, for $c_t = 0.9$, the evolution of the correlation $c_s(\alpha)$ between the two different segments of the CP$_1$ student perceptron is presented, as it evolves as a function of the reduced size $\alpha := p/N$ of the training set, for pattern correlations $c_p = 0$, 0.2, 0.7 and 0.9.

6. Conclusions

We have studied the generalization properties of student perceptrons, which try to learn a 'classification rule with spatial correlations', implemented by a teacher perceptron with built-in spatial correlations between the components of the coupling vector. 'Batch learning' is used, and the patterns are drawn from a spatially nonuniform distribution as well, allowing correlations between different sites, which can be different, however, from the above-mentioned spatial correlations of the teacher. We concentrated on the natural case of 'segmented perceptrons' and 'segmented patterns', where the correlations were those of corresponding sites in different segments, and where the different correlation matrices involved in our formalism had at least the same eigenvectors ('quasisegmented systems').

Using the replica method [16, 2] with a replica-symmetric ansatz, which is exact in this case, we obtained the behaviour of Gibbs and Bayesian learning in the thermodynamic limit. As a third learning algorithm we investigated the Hebb rule, and found that in the presence of correlations it is useful only for low loading and exceptional limiting cases of vanishing or extreme correlation. Otherwise there remains a residual error for $\alpha \to \infty$. However, due to its simplicity, the Hebb rule allows the easiest determination of the site-correlation measure c_t of the 'teacher rule' by monitoring the pattern correlation c_p and the generalization error for finite α and comparing with equation (17).

In contrast, for the Gibbs and Bayes cases we find that the structure of the patterns and of the teacher machines does not matter asymptotically for $\alpha \to \infty$, and perfect generalization is achieved. Nevertheless, in an intermediate α-regime the performance is quite sensitive to correlations which can improve or worsen the generalization ability.

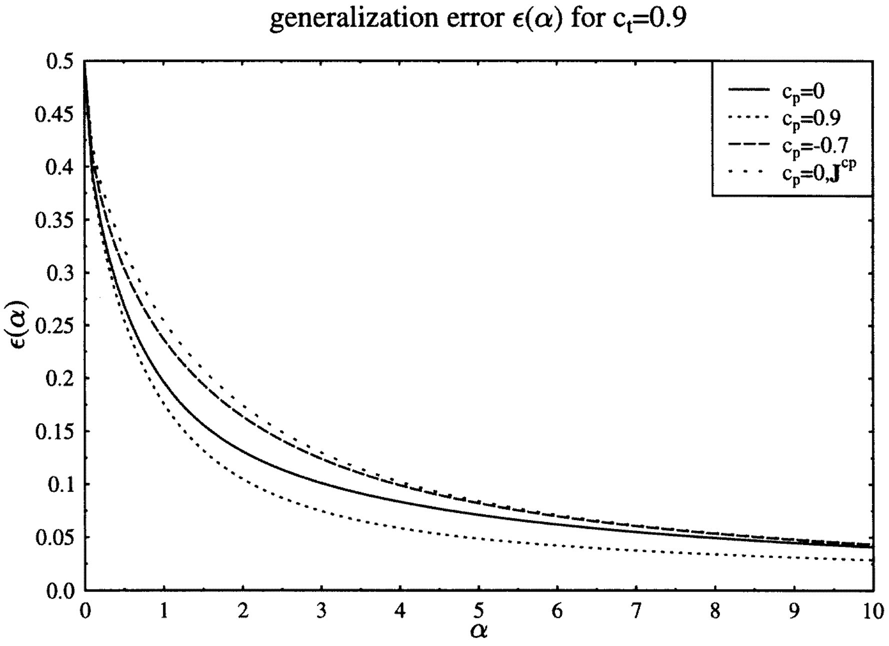

Figure 12. For Bayesian learning with *restricted prior*, i.e. the CP$_2$ perceptron, for $c_t = 0.9$, the generalization error $\epsilon(\alpha)$ is presented as a function of the reduced size $\alpha := p/N$ of the training set, for pattern correlations $c_p = 0$, 0.7, 0.9, and also, for comparison, with unrestricted prior (i.e. the CP$_1$ perceptron) and $c_p = 0$.

(We only mention at this place that we have verified some results by numerical implementation of the learning algorithms, which is difficult for Gibbs and Bayes processes. We simply used small systems, where the phase space was sampled by Monte Carlo methods; a more effective way allowing for larger systems is suggested in a recent preprint of Berg and Engel [24].)

Difficult learning cases are those with *opposite correlations* in the patterns and the teacher vector, respectively. For the Hebb rule the residual error is high, for the other learning rules the generalization error is high for intermediate α.

These effects can be understood better by viewing the scenario as a learning problem with different magnitudes for different components of patterns and teacher vectors. This consideration relates things to methods like principal component analysis. Here an interesting and practical extension would be to investigate the influence of noise, whose disturbing influence should depend on the relation between its size and the corresponding magnitudes of pattern and teacher-vector sites (see [8] for multilayer networks with noise, but still for uncorrelated patterns).

For the Bayesian case we investigated the influence of different priors, showing that improved prior knowledge (e.g. based on a knowledge of the just mentioned quantity c_t) enhances the performance, but again only for an intermediate regime of α. This corresponds to the well known fact that prior information loses significance for large training sets.

The case of maximum-stability learning, where the AdaTron algorithm of Anlauf and Biehl provides a fast and effective learning algorithm [25] and a related cavity method, will be the themes of a following paper.

Acknowledgments

The authors would like to thank F Gerl, M Probst, J Winkel, H Kirschner and B Vogl for useful discussions.

References

[1] Watkin T, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499

[2] Opper M and Kinzel W 1996 Statistical mechanics of generalization *Model of Neural Networks* ed L van Hemmen, E Domany and K Schulten (Berlin: Springer)

[3] Winkel J Ó, Gerl F and Krey U 1996 *Z. Phys.* B **100** 149

[4] Winkel J Ó, Schottky B, Gerl F and Krey U 1996 *Z. Phys.* B **101** 305

[5] Engel A 1990 *J. Phys. A: Math. Gen.* **23** 2587

[6] Tarkowski W and Lewenstein M 1993 *J. Phys. A: Math. Gen.* **26** 2453

[7] Schottky B 1995 *J. Phys. A: Math. Gen.* **28** 4515

[8] Schottky B and Krey U 1997 *J. Phys. A: Math. Gen.* **30** 8541

[9] Cortes C and Vapnik V 1995 *Mach. Learn.* **20** 273

[10] Biehl M 1994 *Europhys. Lett.* **28** 525
Biehl M 1990 *J. Phys. A: Math. Gen.* **23** 258

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

[12] Bishop C M 1995 *Neural Networks for Pattern Recognition* (Oxford: Clarendon)

[13] Pöppel G and Krey U 1989 *Z. Phys.* B **76** 589

[14] Monasson R 1992 *J. Phys. A: Math. Gen.* **25** 3701

[15] Buhmann J and Schulten K 1987 *Europhys. Lett.* **4** 1205

[16] Gardner E 1987 *Europhys. Lett.* **4** 481

[17] Opper M, Kinzel W, Klein J and Nehl R 1990 *J. Phys. A: Math. Gen.* **23** L581

[18] MacKay D J C 1991 *Neural Comput.* **4** 415

[19] MacKay D J C 1991 *Neural Comput.* **4** 698

[20] Opper M and Haussler D 1991 *Phys. Rev. Lett.* **66** 2677

[21] Bruce A D and Saad D 1994 *J. Phys. A: Math. Gen.* **27** 3355

[22] Watkin T L H 1993 *Europhys. Lett.* **21** 871

[23] Dirscherl G 1996 *Diploma Thesis* University of Regensbur, unpublished

[24] Berg J and Engel A 1997 *LANL-Preprint* cond-mat/9710154

[25] Anlauf J K and Biehl M 1989 *Europhys. Lett.* **10** 687