J. Phys. A: Math. Gen. 26 (1993) 6893–6914. Printed in the UK

Statistical mechanics calculation of Vapnik–Chervonenkis bounds for perceptrons

A Engel and W Fink

Institut für Theoretische Physik, Georg-August-Universität, 37073 Göttingen, Bunsenstrasse 9, Federal Republic of Germany

Received 16 September 1993

Abstract. Using the replica technique we calculate the maximal possible difference between the learning and the generalization error of a perceptron learning a linearly separable Boolean classification from examples. We consider both spherical and Ising constraints on the couplings of the perceptron, investigate learnable as well as unlearnable problems and study the special situation where the class of perceptrons considered is restricted to the version space. The results are compared with the Vapnik–Chervonenkis bound and variants thereof. We find that these bounds are asymptotically tight within logarithmic corrections.

1. Introduction

Learning tasks of information processing such as recognition, classification and categorization from examples is a far reaching concept of general interest. In recent years statistical physics has contributed to the understanding of general features of this problem by analysing model situations which, on the one hand, are complex enough to show some of the generic aspects of the problem and, on the other hand, are simple enough to allow a mathematical treatment. A particularly suitable scenario in this respect is given by a 'student' perceptron trained to classify input patterns on the basis of examples provided by a 'teacher' perceptron [1, 2]. A perceptron is the simplest feed-forward network consisting of N input units S_j, $j = 1, \ldots, N$, and one output unit S_0 connected to the inputs by real-valued couplings J_i (see figure 1). We will only consider binary units $S_j = \pm 1$. For any input $\{S_j\}$, the perceptron determines the output according to

$$S_0 = \operatorname{sign}\left(\frac{1}{\sqrt{N}} \sum_j J_j S_j\right) \tag{1}$$

thereby providing a binary classification of all 2^N possible input patterns $\{S_j\}$. Let us assume that one particular classification is fixed by a teacher or target perceptron $\boldsymbol{T} = \{T_j\}$. The aim for the student perceptron $\boldsymbol{J} = \{J_j\}$ is either to reproduce or, if this is impossible due to different restrictions on the vectors $\boldsymbol{T}$ and $\boldsymbol{J}$, to approximate this target classification as faithfully as possible. The simplest procedure would of course be to set $J_j = T_j$ for all $j = 1, \ldots, N$, corresponding in a sense to explicit programming. More interesting, however, is the situation where the individual values T_j of the teacher couplings are not accessible to the student (figure 1). The classification ζ_T^μ of special patterns $\boldsymbol{\xi}^\mu = \{\xi_j^\mu\}$, $\mu = 1, \ldots, p$, by the teacher is known instead and the student is expected to infer the complete rule $\boldsymbol{T}$ from these examples. The patterns $\boldsymbol{\xi}^\mu$ together with their classification ζ_J^μ by the teacher

0305-4470/93/226893+22$07.50

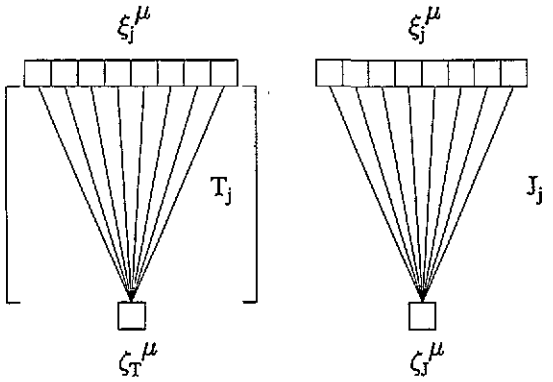

Figure 1. Setup of the two-perceptron scenario: A student perceptron J tries to learn the binary classification defined by the teacher perceptron T on the basis of input–output examples provided by the teacher.

form the so-called training set. The general strategy of the student will be to modify the coupling vector J in order to produce nearly the same classification of the patterns of the training set as the teacher. The hope is that for a sufficiently large training set this will result in an alignment of the vectors T and J which would imply that the student also classifies a pattern *not contained in the training set* similar to the teacher, i.e. that he will generalize from the examples to the rule. If the patterns of the training set are independently drawn at random from the set of all possible patterns according to some probability distribution $P(\xi^\mu)$ the quantities of central interest are the training error $\nu_J(p)$ and the generalization error e_J. The training error $\nu_J(p)$ denotes the fraction of patterns of the training set which the student classifies differently from the teacher. The generalization error is the probability that an arbitrary pattern drawn at random with *the same probability distribution* P (which therefore may or may not belong to the training set) is classified differently by teacher and student. The aim is to make this generalization error as small as possible.

Using methods of statistical mechanics of neural networks, the learning and generalization error have been determined for various combinations of learning rules, pattern statistics and network architectures in the limit $N \to \infty$, $p \to \infty, \alpha = p/N =$ constant [3–5]. In this limit the fluctuations in the performance due to the random nature of most of the learning rules used are suppressed and one ends up with *typical* results for the learning and generalization errors. This means that for one particular realization of both the patterns in the training set and the learning schedule one finds with probability 1 the values of the statistical mechanics analysis for $\nu_J(p)$ and e_J.

Besides these rather recent studies in statistical physics there is, however, a much longer line of investigations of the same problem in mathematical statistics which so far has remained fairly unrelated to the work of the statistical mechanics community. From the point of view of mathematical statistics, the generalization is just a variant of the general problem of convergence of frequencies to probabilities: the learning error is the frequency of mistakes of the student on a test set, whereas the generalization error denotes the corresponding probability. For finite N, one has clearly $\nu_J(p) \to e_J$ for $p \to \infty$. If one were able to characterize the fluctuations of $\nu_J(p)$ around e_J for finite p it would be possible to give estimates of e_J on the basis of $\nu_J(p)$. Since learning rules are designed to make $\nu_J(p)$ very small (ideally zero), the decrease of e_J with the size p of the training set could be quantified.

In fact, it is easy to describe the fluctuations of $\nu_J(p)$ around its limit e_J for a *fixed* student vector J. Since J and the patterns of the test set are uncorrelated, $\nu_J(p)$ obeys a Bernoulli distribution and we have by virtue of the Hoeffding inequality [6]

$$\text{Prob}\{|\nu_J(p) - e_J| > \epsilon\} \leqslant \delta(\epsilon, p) = 2\,\mathrm{e}^{-2\epsilon^2 p}. \tag{2}$$

Hence, the probability that $\nu_J(p)$ deviates from e_J by an amount ϵ is bounded by $\delta(\epsilon, p)$. Note that a constant bound corresponds to $\epsilon \sim 1/\sqrt{p}$, as is familiar from the central limit theorem. The occurence of two parameters, the tolerated error ϵ and the confidence δ, in equation (2) is the trademark of the concept of *probably approximately correct* (PAC) learning [7] frequently studied in computational science approaches.

The Hoeffding inequality, however, is not sufficient to describe the generalization problem. The reason is that the student vectors J of interest are chosen in order to make $\nu_J(p)$ small and in particular are modified if new patterns are added to the training set. This gives rise to strong correlation between the J's and the ξ^μ's; in fact, these correlations are the central aim of learning. In this way the *test* set really becomes a *training* set and it is impossible to apply the Hoeffding inequality.

A way to characterize the convergence of the training error to the generalization error is to find a bound for the difference between them, which is *uniform on the set of all possible students* (e.g. all perceptrons with real-valued coupling vector J). Such a bound is naturally given by a bound for the *maximal* possible difference

$$\max_J |\nu_J(p) - e_J| \tag{3}$$

between $\nu_J(p)$ and e_J. Estimating this maximum is often refered to as *worst-case analysis*. In a by now classical study, Vapnik and Chervonenkis have derived such a uniform bound of the form [8]

$$\text{Prob}\{\max_J |\nu_J(p) - e_J| > \epsilon\} \leqslant \delta^{\text{VC}}(\epsilon, p). \tag{4}$$

By definition, this bound applies to all possible student vectors J, in particular to those designed on the basis of the training set ξ^μ. Hence, knowing the behaviour of $\nu_J(p)$ for some particular learning rule, one can deduce bounds for the corresponding generalization error. If, for example, the learning rule achieves $\nu_J(p) = 0$ for all p, the probability that the generalization error will exceed some small parameter ϵ decreases as $\delta^{\text{VC}}(\epsilon, p)$.

Due to its generality, the Vapnik–Chervonenkis (VC) theorem is central in the mathematical theory of learning from examples and has been applied to several interesting situations [9–11]. There have also been several refinements since its original derivation [12]. A very nice introduction into the subject of uniform convergence bounds for physicists is provided by [13].

In the present paper we make contact between the investigations of the generalization problem in the mathematical statistics and statistical mechanics community, respectively, by testing the VC bound against the *actual performance of the worst student* under several circumstances. Note that this is a partial worst-case scenario only, whereas the VC theorem also holds for the worst possible choice of the teacher T and of the distribution $P(\xi^\mu)$ of the patterns forming the training set [14]. Our results will on the one hand highlight the tightness of the VC bounds in this case and on the other hand quantify the difference in performance between the worst and the typical student. This will also indicate, therefore, at least for the perceptron, whether a worst-case analysis is really too pessimistic, as is frequently claimed. We first discuss in section 2 the form of the VC theorem for perceptrons in the statistical mechanics limit $N \to \infty$, where we mainly follow [13]. Then we show in section 3 how one can calculate the performance of the *worst* student (instead of that

of the *typical* one) using methods of statistical mechanics. Sections 4 and 5 contain the corresponding replica calculation, where details are relegated to the appendix. In section 6 we discuss the special case that all student vectors belong to the version space, i.e. they perform on the training set exactly like the teacher. Section 7 is dévoted to student vectors restricted to the hypercube, $J_j = \pm 1$. Here we also discuss the situation of an unrealizable rule, in which no student J with $\nu_J(p) = 0$ for sufficiently large p exists. This can be viewed as a first step towards a worst-case analysis including the choice of the teacher. Finally, section 8 contains our conclusions. Part of the results of sections 4–6 have already been reported in a recent letter [15].

2. VC bound for large perceptrons

The VC theorem provides a uniform bound for the convergence of the learning error $\nu_J(p)$ to the generalization error e_J for a whole class $\mathcal{P}$ of classifiers J according to [8]:

$$\text{Prob}\left\{\max_{J\in\mathcal{P}} |\nu_J(p) - e_J| > \epsilon\right\} \leqslant c\Delta(2p)\mathrm{e}^{-p\epsilon^2}. \tag{5}$$

Here $\Delta(m)$ is the so-called growth function defined as the maximal possible number of different classifications of m patterns that can be induced by classifiers of class $\mathcal{P}$, and c is a constant slightly larger than 6. The bound is uniform on $\mathcal{P}$ since it refers to the worst possible choice of $J \in \mathcal{P}$. Hence, it applies in particular to the J's of interest, namely those that were designed on the basis of the training set to produce small values of the learning error $\nu_J(p)$. The VC bound (5) is of little use if $\Delta(m)$ always grow exponentially with m. On the other hand, considering simple examples, one easily realizes that for small m the behaviour $\Delta(m) = 2^m$ is rather typical. The strength of the VC theorem lies in the fact that for *any* class of classifiers $\mathcal{P}$ there exists an integer number d^{VC} (which may be infinite) called the VC dimension of this class such that [8, 16]

$$\Delta(m) \begin{cases} = 2^m & \text{if} \quad m \leqslant d^{\text{VC}} \\ \leqslant \sum_{i=0}^{d^{\text{VC}}} \binom{m}{i} & \text{if} \quad m \geqslant d^{\text{VC}} \end{cases}. \tag{6}$$

Hence, for $p \gg d^{\text{VC}}$, $\Delta(2p)$ grows only as a power of p and the bound (5) becomes effective. In the present paper we deal with perceptrons specified by a vector $J \in I\mathbb{R}^N$ in the thermodynamic limit $N \to \infty$, $p \to \infty$, $\alpha = p/N =$ constant. For perceptrons without threshold the VC dimension is known to be $d^{\text{VC}} = N$; moreover, for $p > N$ the *exact* result $\Delta(m) = 2\sum_{i=0}^{N-1}\binom{m-1}{i}$ holds [17]. Using the Stirling formula, replacing the sum by an integral and using peak integration, we then get for $N \to \infty$ and $\alpha \geqslant 1$

$$\Delta(2\alpha N) = 2\sum_{i=0}^{N-1}\binom{2\alpha N - 1}{i} \sim \exp\{N[2\alpha\log(2\alpha) - (2\alpha - 1)\log(2\alpha - 1)]\} \tag{7}$$

which, when combined with (5), yields

$$\text{Prob}\left\{\max_{J\in\mathcal{P}} |\nu_J(p) - e_J| > \epsilon\right\} \leqslant c\,\exp\left\{N[2\alpha\log(2\alpha) - (2\alpha - 1)\log(2\alpha - 1) - \alpha\epsilon^2]\right\}. \tag{8}$$

For $N \to \infty$ this means that $\max_{J\in\mathcal{P}} |\nu_J(p) - e_J|$ is bounded by a threshold

$$\epsilon^{\text{VC}}(\alpha) = \sqrt{2\log 2\alpha - (2 - 1/\alpha)\log(2\alpha - 1)} \tag{9}$$

with probability 1. Note that $\epsilon^{\text{VC}}(\alpha) \sim \sqrt{\log(\alpha)/\alpha}$ for large α.

It is this *bound* $\epsilon^{\text{VC}}(\alpha)$ for the maximal difference between learning and generalization error for the class of large perceptrons which we would like to test, in what follows, against the *actual performance* of the worst student. Let us finally note that for perceptrons one

has $\nu_{-J}(p) = 1 - \nu_J(p)$ and $e_{-J} = 1 - e_J$. Hence if the maximum of $(\nu_J(p) - e_J)$ is realized for a perceptron J^* then the minimum of $(\nu_J(p) - e_J)$ is realized for $-J^*$ and $\max(\nu_J(p) - e_J) = -\min(\nu_J(p) - e_J) = \max|\nu_J(p) - e_J|$. It is therefore possible to drop the absolute value. As one usually minimizes $\nu_J(p)$ on the training set in order to get low values for e_J as well, we will refer to the perceptron maximizing $(e_J - \nu_J(p))$ as the worst student.

3. Worst-case analysis using statistical mechanics

Our aim in this section is to calculate explicitly

$$\epsilon(\alpha) = \max_J (e_J - \nu_J(\alpha)) \tag{10}$$

as a function of the training set size α and to compare it with the threshold $\epsilon^{\text{VC}}(\alpha)$ resulting from the VC theorem. Note that the asymptotic form (8) of the VC bound suggests that $\epsilon(\alpha)$ is self-averaging for $N \to \infty$.

It is well known that in the situation of a student perceptron J generalizing a teacher perceptron T, where both teacher and student obey the spherical normalization

$$T^2 := \sum_j T_j^2 = J^2 = N \tag{11}$$

the generalization error e_J depends on J only through the variable

$$R = \frac{1}{N} T \cdot J := \frac{1}{N} \sum_j T_j J_j \tag{12}$$

according to [18, 19]

$$e_J = \frac{1}{\pi} \cos^{-1} R. \tag{13}$$

It is therefore convenient to split the calculation of $\max_J(e_J - \nu_J(\alpha))$ into two steps. First we determine the minimal possible training error $\nu_{\min}(\alpha, R)$ of a perceptron J with overlap R with the teacher T. Then the maximal possible value $\epsilon(\alpha, R)$ of $(e_J - \nu_J(\alpha))$ for these J's is given by

$$\epsilon(\alpha, R) = \frac{1}{\pi} \cos^{-1} R - \nu_{\min}(\alpha, R). \tag{14}$$

In the second step we maximize with respect to R to get

$$\epsilon(\alpha) = \max_R \epsilon(\alpha, R). \tag{15}$$

To determine $\nu_{\min}(\alpha, R)$, we introduce the (extensive) energy $E(J, \alpha)$ of a perceptron J according to

$$E(J, \alpha) = N\nu_J(\alpha) = \frac{1}{\alpha} \sum_\mu \theta\left(-\left(\frac{1}{\sqrt{N}} J \cdot \xi^\mu\right)\left(\frac{1}{\sqrt{N}} T \cdot \xi^\mu\right)\right) \tag{16}$$

and calculate the free energy density

$$f(\alpha, \beta, R) = -\lim_{N\to\infty} \frac{1}{\beta N} \left\langle\!\left\langle \log \int \mathrm{d}\mu_R(J)\, e^{-\beta E(J,\alpha)} \right\rangle\!\right\rangle \tag{17}$$

where the average is taken over the statistics of the patterns ξ^μ, $\mu = 1, \ldots, p$, forming the training set according to

$$P(\xi^\mu) = \prod_{i,\mu} \left[\tfrac{1}{2}\delta(\xi_i^\mu - 1) + \tfrac{1}{2}\delta(\xi_i^\mu + 1)\right] \tag{18}$$

and

$$\mathrm{d}\mu_R(\boldsymbol{J}) = \prod_i \mathrm{d}J_i\ \delta(N - \boldsymbol{J}^2)\ \delta(NR - \boldsymbol{J}\cdot\boldsymbol{T}) \tag{19}$$

incorporates the constraints (11) and (12). For large values of β the free energy is dominated by $\boldsymbol{J}$ vectors with small energies. Accordingly

$$v_{\min}(\alpha, R) = \lim_{\beta\to\infty} f(\alpha, \beta, R). \tag{20}$$

The calculation of $f(\alpha, \beta, R)$ can be done using the replica trick and standard techniques from the statistical mechanics of neural networks [2, 19, 20]. Some of the intermediate steps are sketched in the appendix. The result is

$$f(\alpha, \beta, R) = -\min_{q^{ab}} \lim_{n\to 0} \left[\frac{1}{2\beta}(1 + \log 2\pi) + \frac{1}{2\beta n}\mathrm{Tr}\log(\delta^{ab} + q^{ab} - R^2) + \frac{\alpha}{\beta n} G(R, q^{ab}) \right] \tag{21}$$

where the minimum is over the $n(n-1)/2$ order parameters $q^{ab} = q^{ba}$, $a \neq b$ and

$$G(R, q^{ab}) = \log 2 \int_0^\infty \mathrm{D}u \int \prod_{a=1}^n \frac{\mathrm{d}\lambda_a\, \mathrm{d}x_a}{2\pi} \exp\left\{ \mathrm{i}\sum_a x_a(\lambda_a - uR) - \frac{1}{2}\sum_a x_a^2 - \frac{1}{2}\sum_{(a,b)} x_a x_b q^{ab} + \frac{R^2}{2}\sum_{a,b} x_a x_b - \beta/\alpha \sum_a \theta(-\lambda_a) \right\} \tag{22}$$

with $\mathrm{D}u := (\mathrm{d}u/\sqrt{2\pi})\mathrm{e}^{-u^2/2}$. *The crucial step in the replica calculation is to find the* correct minimum with respect to the $n \times n$ order parameter matrix q^{ab} in equation (21) in the limit $n \to 0$. This can only be accomplished using special ansätze for the structure of the matrix q^{ab}. The simplest one is that of replica symmetry, to which we turn in the next section.

4. Replica symmetry

The replica-symmetric ansatz for the order parameter matrix q^{ab} is of the form $q^{ab} = q$, $a \neq b$. It is then possible to simplify considerably the expression for the free energy (21), (22) with the result (see appendix)

$$f(\alpha, \beta, R) = -\min_q \left[\frac{1}{2\beta}(1 + \log 2\pi) + \frac{1}{2\beta}\log(1-q) + \frac{q - R^2}{2\beta(1-q)} + \frac{2\alpha}{\beta}\int \mathrm{D}t\, H\left(-\frac{Rt}{\sqrt{1-R^2}}\right) \log \int \frac{\mathrm{d}\lambda}{\sqrt{2\pi(1-q)}} \mathrm{e}^{-\beta V(\lambda)} \right] \tag{23}$$

where

$$V(\lambda) = \frac{(\lambda + \sqrt{q}t)^2}{2x} + \frac{1}{\alpha}\theta(\lambda) \tag{24}$$

$H(x) := \int_x^\infty \mathrm{D}t$ and $x := \beta(1-q)$. In order to find $v_{\min}(\alpha, R)$ we have to take the limit $\beta \to \infty$. Then the λ-integral is dominated by the minimum $V(\lambda_0)$ of $V(\lambda)$ and we find

$$v_{\min}^{\mathrm{RS}}(\alpha, R) = -\min_x \left[\frac{1 - R^2}{2x} - 2\alpha \int \mathrm{D}t\, H\left(-\frac{Rt}{\sqrt{1-R^2}}\right) V(\lambda_0) \right]. \tag{25}$$

Using

$$\begin{array}{lll} \lambda_0(t) = -t & V(\lambda_0) = 0 & \text{if} \quad 0 < t \\ \lambda_0(t) = 0 & V(\lambda_0) = t^2/2x & \text{if} \quad -\sqrt{2x/\alpha} < t < 0 \\ \lambda_0(t) = -t & V(\lambda_0) = 1/\alpha & \text{if} \quad t < -\sqrt{2x/\alpha} \end{array} \tag{26}$$

we finally get

$$\nu_{\min}^{\mathrm{RS}}(\alpha, R) = -\min_x \left[\frac{1-R^2}{2x} - \frac{1}{\pi}\cos^{-1} R + \int_0^{\sqrt{\frac{2x}{\alpha}}} \mathrm{D}t \left(2 - \frac{\alpha t^2}{x}\right) H\left(\frac{Rt}{\sqrt{1-R^2}}\right) \right]. \tag{27}$$

The value of x minimizing the RHS of (27) is given by the solution of

$$0 = -\frac{1-R^2}{2x^2} + \frac{\alpha}{x^2} \int_0^{\sqrt{\frac{2x}{\alpha}}} \mathrm{D}t\, t^2 H\left(\frac{Rt}{\sqrt{1-R^2}}\right). \tag{28}$$

Obviously $x = \infty$ corresponding to $\nu_{\min}^{\mathrm{RS}}(\alpha, R) = 0$ is always a solution of (28). However for $R < R_{\min}(\alpha)$ there is another solution $x < \infty$ giving rise to a negative value of the square bracket in (27). The result for $\nu_{\min}(\alpha, R)$ is then

$$\nu_{\min}(\alpha, R) = \frac{1}{\pi}\cos^{-1} R - 2 \int_0^{\sqrt{\frac{2x}{\alpha}}} \mathrm{D}t\, H\left(\frac{Rt}{\sqrt{1-R^2}}\right) > 0. \tag{29}$$

$R_{\min}$ is given by the (smaller) solution of

$$1 - R^2 = \alpha \int_0^{\infty} \mathrm{D}t\, t^2 H\left(\frac{Rt}{\sqrt{1-R^2}}\right) = \frac{\alpha}{\pi}\left(\cos^{-1} R - R\sqrt{1-R^2}\right). \tag{30}$$

The interpretation of these results is as follows. Due to the spherical constraint (11), all possible student vectors $\boldsymbol{J}$ lie on the surface of an N-dimensional sphere with radius $\sqrt{N}$. Let us take $\boldsymbol{T}$ as the north pole of this sphere. For every value of α there is a subset of the sphere, called the version space $\mathcal{V}$, containing all student vectors $\boldsymbol{J}$ that classify all training examples exactly like the teacher. If part of the rim defined by $\boldsymbol{J} \cdot \boldsymbol{T} = NR$ belongs to this subset we obviously get $\nu_{\min}(\alpha, R) = 0$. Since by definition the teacher $\boldsymbol{T}$ belongs to the version space but the vector $\boldsymbol{T}$ does not, there must be a critical value $R_{\min}(\alpha)$ of R such that for $R < R_{\min}$ no student $\boldsymbol{J}$ out of the version space with $\boldsymbol{J} \cdot \boldsymbol{T} = NR$ exists and accordingly $\nu_{\min}(\alpha, R) > 0$. Hence, $R_{\min}$ as given by (30) marks the 'most southern tip' of the version space formed by the student with still zero training error but the worst generalization ability. $R_{\min}(\alpha)$ is plotted in figure 2. Note that $R_{\min}(\alpha = 0) = -1$ and $R_{\min}(\alpha \to \infty) \to 1$, in accordance with intuition. Moreover, $R_{\min}(\alpha = 2) = 0$. A student vector $\boldsymbol{J}$ confined to the plane perpendicular to the teacher cannot take advantage of the correlations between input and output induced by the teacher and hence has to learn random classifications. This becomes impossible beyond the Gardner threshold $\alpha_c = 2$ and hence $R_{\min} = 0$ for $\alpha = 2$.

In figure 3, $\nu_{\min}^{\mathrm{RS}}(\alpha, R)$ is plotted as a function of R for three different values of α. For $R > R_{\min}(\alpha)$ one has $\nu_{\min}^{\mathrm{RS}}(\alpha, R) = 0$. At $R_{\min}(\alpha)$, $\nu_{\min}$ becomes positive and tends to 1 for $R \to -1$ for all α.

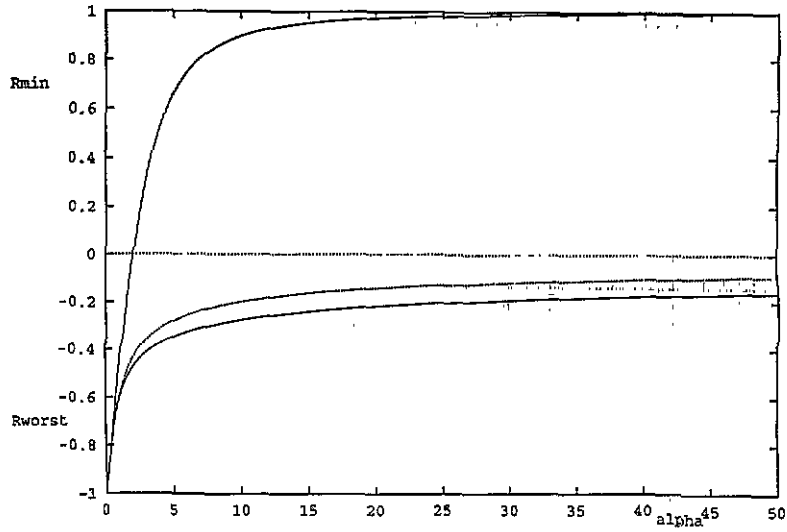

Figure 2. *Minimal possible overlap $R_{\min}$ of a student vector J still belonging to the version space (top) and overlap R_{worst} of the student vector that realizes the largest difference between* training and generalization error (bottom) as a function of α in replica symmetry (full lines) and one-step replica symmetry breaking (dotted lines).

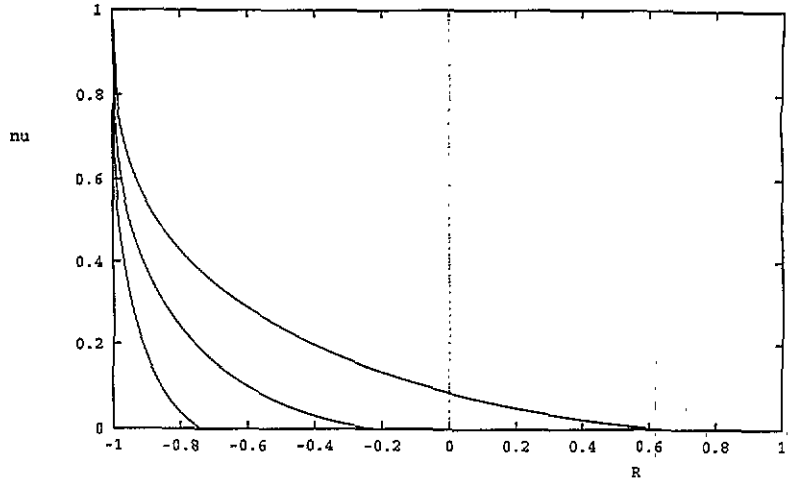

Figure 3. Minimal possible training error $\nu_{\min}(\alpha, R)$ as a function of the overlap R with the teacher for $\alpha = 5$, $\alpha = 1.5$ and $\alpha = 0.5$ (from top to bottom).

From (15), (14) and (29) we finally get the maximal possible difference $\epsilon(\alpha, R)$ between the training and generalization error for a given R in replica symmetry

$$\epsilon^{\text{RS}}(\alpha, R) = \begin{cases} 2\int_0^{\sqrt{\frac{2x}{\alpha}}} \mathrm{D}t\, H\left(\frac{Rt}{\sqrt{1-R^2}}\right) & \text{if} \quad -1 \leqslant R \leqslant R_{\min} \\ \frac{1}{\pi}\cos^{-1} R & \text{if} \quad R_{\min} \leqslant R \leqslant 1 \end{cases} \tag{31}$$

where x is given by the finite solution of equation (28). Figure 4 shows a plot of $\epsilon^{\text{RS}}(\alpha, R)$ as a function of R for the three values of α in figure 3. For every value of α there is a unique maximum $\epsilon^{\text{RS}}(\alpha)$ of $\epsilon^{\text{RS}}(\alpha, R)$ with respect to R. Moreover, since $\partial \nu_{\min}/\partial R = 0$

for $R = R_{\min}$, as follows from (29), this maximum lies always *outside the version space*, i.e. for a value R_{worst} of R with $\nu_{\min}^{\text{RS}}(\alpha, R) > 0$. We have included $R_{\text{worst}}(\alpha)$ in figure 2. Note that contrary to $R_{\min}(\alpha)$, $R_{\text{worst}}(\alpha)$ is always negative and tends to 0 for $\alpha \to \infty$. This can also be anticipated from figure 4.

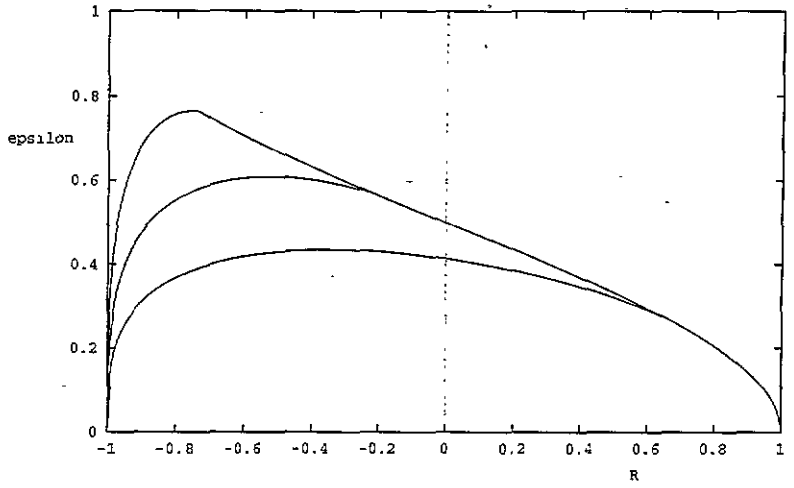

Figure 4. Maximal possible difference between the generalization and training error as a function of R for the same values of α as in figure 3 (now from bottom to top).

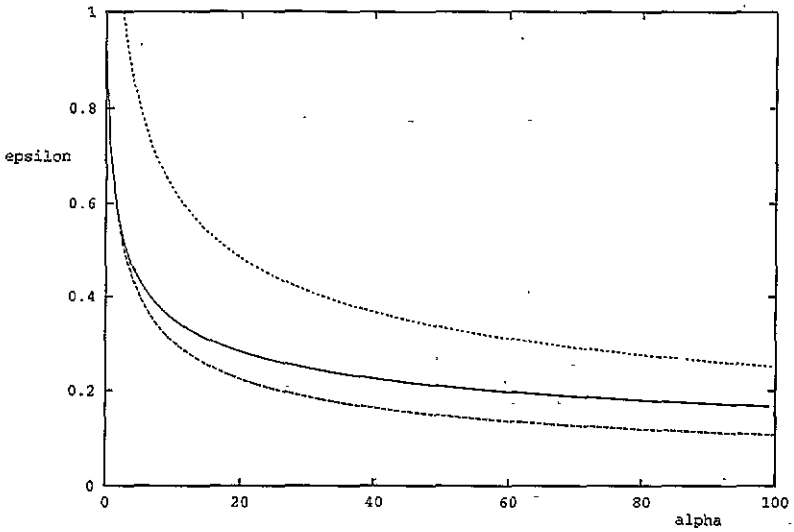

Figure 5. Maximal possible difference between generalization and training error as a function of the training set size α in replica symmetry (full line) and one-step replica symmetry breaking (dashed line). The dotted line is the rigourous upper bound provided by the VC theorem.

The resulting behaviour of $\epsilon^{\text{RS}}(\alpha)$ is shown in figure 5 together with the VC bound (9). As can be seen, the replica-symmetric result is well below the rigorous upper bound provided by the VC theorem for the values of α shown. On the other hand, solving the extremum equations for x and R asymptotically for $\alpha \to \infty$ one finds $x \sim \alpha^{1/3}$, $R \sim \alpha^{-1/3} \to 0$ and $\epsilon^{\text{RS}}(\alpha) \sim 0.78\alpha^{-1/3}$. This, however, violates the VC bound for large α (in fact, the respective

curves of figure 5 cross at $\alpha \cong 4500$) and, hence, the replica-symmetric calculation must be wrong, at least for large α. In fact, testing the local stability of the replica-symmetric saddle point [19, 21], we find local instability for $\alpha > \alpha_{AT}$, where α_{AT} is the solution of

$$\epsilon^{RS}(\alpha_{AT}) = \frac{1}{\alpha_{AT}} \tag{32}$$

yielding $\alpha_{AT} \approx 1.70$. This instability should not come as a surprise since the maximum of $\epsilon^{RS}(\alpha, R)$ with respect to R is always realized at a point with $\nu_{min}(\alpha, R) > 0$. However, as we know from the theory of storing random classifications, a replica symmetry (with the cost function used here) is broken whenever perfect storage becomes impossible [22, 23]. One may therefore suspect that the replica-symmetric solution is wrong even below α_{AT}. To clarify these points we investigate, in the next section, $\epsilon(\alpha)$ in one-step replica symmetry breaking.

5. Replica symmetry breaking

In order to obtain refined results for $\alpha > \alpha_{AT}$ and to test the *global* stability of replica symmetry, we calculate $\epsilon(\alpha)$ in this section using the ansatz of one-step replica symmetry breaking for the order parameter matrix q^{ab}. This ansatz is defined by the three parameters $q_0, q_1 > q_0$ and m according to $q^{aa} = 0$, $q^{ab} = q_1$ if $|a-b| < m$, otherwise $q^{ab} = q_0$. Plugging this into equation (21) we find after some algebra

$$\begin{aligned} f(\alpha, \beta, R) = -\min_{q_0, q_1, m} \Bigg\{ & \frac{1}{2\beta}(1 + \log 2\pi) + \frac{q_0 - R^2}{2\beta(1 - q_1 + m\Delta q)} + \frac{m-1}{2\beta m}\log(1-q_1) \\ & + \frac{1}{2\beta m}\log(1 - q_1 + m\Delta q) + \frac{2\alpha}{\beta m}\int \mathrm{D}t_0\, H\left(-\frac{Rt_0}{\sqrt{q_0 - R^2}}\right) \\ & \times \log \int \mathrm{D}t_1 \left[\int \frac{\mathrm{d}\lambda}{\sqrt{2\pi(1-q_1)}} \exp(-\beta V(\lambda))\right]^m \Bigg\} \end{aligned} \tag{33}$$

where now

$$V(\lambda) = \frac{(\lambda + t_0\sqrt{q_0} + t_1\sqrt{\Delta q})^2}{2x} + \frac{1}{\alpha}\theta(\lambda) \tag{34}$$

$x = \beta(1 - q_1)$ and $\Delta q = q_1 - q_0$. A non-trivial limit of $f(\alpha, \beta, R)$ for $\beta \to \infty$ results if $x = O(1)$ and $m \to 0$ with $\beta m := cx = O(1)$. We then get

$$\begin{aligned} \nu_{min}^{RSB}(\alpha, R) = -\min_{q_0, c, x} \Bigg[& \frac{q_0 - R^2}{2x(1 + c(1-q_0))} + \frac{1}{2cx}\log(1 + c(1-q_0)) \\ & + \frac{2\alpha}{cx}\int \mathrm{D}t_0\, H\left(-\frac{Rt_0}{\sqrt{q_0 - R^2}}\right)\log \int \mathrm{D}t_1 \exp(-cxV(\lambda_0))\Bigg] \end{aligned} \tag{35}$$

where $V(\lambda_0)$ again denotes the minimum of $V(\lambda)$. We have performed the three-dimensional minimization in equation (35) numerically and used the result of a numerical determination of $\max_R (\frac{1}{\pi}\cos^{-1} R - \nu_{min}^{RSB}(\alpha, R))$ to get the curve $\epsilon^{RSB}(\alpha)$ which is included in figure 5. As can be seen, $\epsilon^{RSB}(\alpha) < \epsilon^{RS}(\alpha)$ for all α, as should be expected. Moreover, the replica symmetry broken solution exists also for $\alpha < \alpha_{AT}$ so that the replica-symmetric solution, though locally stable, is incorrect. For $\alpha \leqslant 1$, q_0 is very near to unity and it is difficult to distinguish numerically between the replica-symmetric solution and the one using one-step replica symmetry breaking. Nevertheless we expect from our experience with the problem of storing random classifications [22, 23] that for all $\alpha > 0$ there is a one-step RSB solution giving smaller values for $\epsilon(\alpha)$ than the corresponding replica-symmetric result. Hence for

the problem at hand the RS solution must be rejected everywhere, although it gives a very good approximation for $\epsilon(\alpha)$ if $\alpha \leqslant 1$. As stated already in the last section, the reason for this is that the maximum of $\epsilon(\alpha, R)$ always lies outside the version space (though very near to it for small α, see figure 4). The presence of errors, however, makes the solution space disconnected and this implies the breaking of replica symmetry.

Finally it is important to see how the asymptotic behaviour of $\epsilon(\alpha)$ for $\alpha \to \infty$ is modified by replica symmetry breaking. The results of the numerical minimization suggest $cx/\alpha \to 0$ and $c(1-q) \to \infty$ for $\alpha \to \infty$. Using these ansätze we find, to leading order,

$$\epsilon^{\mathrm{RSB}}(\alpha) \cong \mathrm{extr}_{R,q_0,c,m}\left[\frac{q_0 - R^2}{2x(1+c(1-q_0))} + \frac{1}{2cx}\log(1+c(1-q_0)) + \frac{\sqrt{2}}{4\pi}\frac{cx}{\alpha}\sqrt{\frac{1-q_0}{q_0}} + \frac{2}{3}\sqrt{\frac{x}{\alpha\pi}} - \frac{2x}{\alpha\pi}\frac{R}{\sqrt{1-R^2}}\right]. \tag{36}$$

From this one gets the self-consistent saddle point $x \sim 9/4(\log\alpha)^{-3/2}$, $(1-q_0) \sim 3/2(\log\alpha)^{-1}$, $c \sim 4\sqrt{\pi}/9\sqrt{\alpha}(\log\alpha)^{9/4}$ and $R = -9/\sqrt{\pi}\alpha^{-1/2}(\log\alpha)^{-7/4}$, yielding

$$\epsilon^{\mathrm{RSB}}(\alpha) \sim 1.02\frac{(\log\alpha)^{1/4}}{\sqrt{\alpha}}. \tag{37}$$

It is reassuring that, contrary to the replica-symmetric result, $\epsilon^{\mathrm{RSB}}(\alpha)$ remains below the VC bound $\epsilon^{\mathrm{VC}}(\alpha) \sim \sqrt{\log\alpha/\alpha}$ for large α. On the other hand, it is difficult to estimate whether $\epsilon^{\mathrm{RSB}}(\alpha)$ indeed characterizes the performance of the worst student or if there are substantial corrections due to higher-order breakings of replica symmetry. In fact, the asymptotic result $q_0 \to 1$ for $\alpha \to \infty$ suggests that one-step replica symmetry breaking is not sufficient, since we expect the smallest overlap scale ($q(x=0)$ in the full replica symmetry breaking scheme of Parisi) to tend to zero for $\alpha \to \infty$ (see also section 7).

6. Restriction to the version space

For any value of α there is a non-empty set of perceptrons producing exactly the same output for all patterns of the training set as the teacher. A simple but efficient learning rule consists of choosing the student vector $\boldsymbol{J}$ at random from this set, the so-called version space $\mathcal{V}$. Both the VC theorem and the statistical mechanics analysis are special in this case. The convergence of frequencies to probabilities is much faster if the frequencies are close to zero. This allows us to improve the VC bound. By definition, one now has $\nu_J(\alpha) = 0$ and therefore the VC bound is a bound for the generalization error e_J itself. The result is [10, 11]

$$\mathrm{Prob}\left\{\max_{J\in\mathcal{V}} e_J > \epsilon\right\} \leqslant 4\Delta(2p)2^{-p\epsilon} \tag{38}$$

to be compared with (5). The convergence with p is now much faster, since we have $\epsilon \sim 1/p$ instead of $\epsilon \sim 1/\sqrt{p}$. For $p = \alpha N$ and $N \to \infty$ we find for $\alpha > 1$ by analogy to (8) the new threshold [13]

$$\epsilon_{\mathcal{V}}^{\mathrm{VC}}(\alpha) = [2\log 2\alpha - (2-1/\alpha)\log(2\alpha-1)](\log 2)^{-1} \tag{39}$$

which implies for large α

$$\epsilon_{\mathcal{V}}^{\mathrm{VC}}(\alpha) \sim \frac{\log\alpha}{\alpha\log 2}. \tag{40}$$

For the statistical mechanics analysis, we have to determine the largest generalization error of a vector J still belonging to the version space $\mathcal{V}$. In replica symmetry it is given by

$$\epsilon_{\mathcal{V}}^{\mathrm{RS}}(\alpha) = \max_{J \in \mathcal{V}} e_J = \frac{1}{\pi} \cos^{-1} R_{\min} \tag{41}$$

where $R_{\min}$ marks, as explained in section 4, the 'southernmost tip' of the version space and is given by equation (30). The resulting behaviour of $\epsilon_{\mathcal{V}}^{\mathrm{RS}}(\alpha)$ is shown in figure 6 together with the VC bound. Again, statistical mechanics gives a result that is well below the rigorous upper bound provided by the VC theorem for the α-values shown. Solving (30) for $\alpha \to \infty$ one finds $R_{\min} \sim 1 - 9\pi^2/8\alpha^2$, implying $\epsilon_{\mathcal{V}}^{\mathrm{RS}}(\alpha) \sim 3/(2\alpha)$, so that in the present case the RS result remains below the upper bound for *all* values of α.

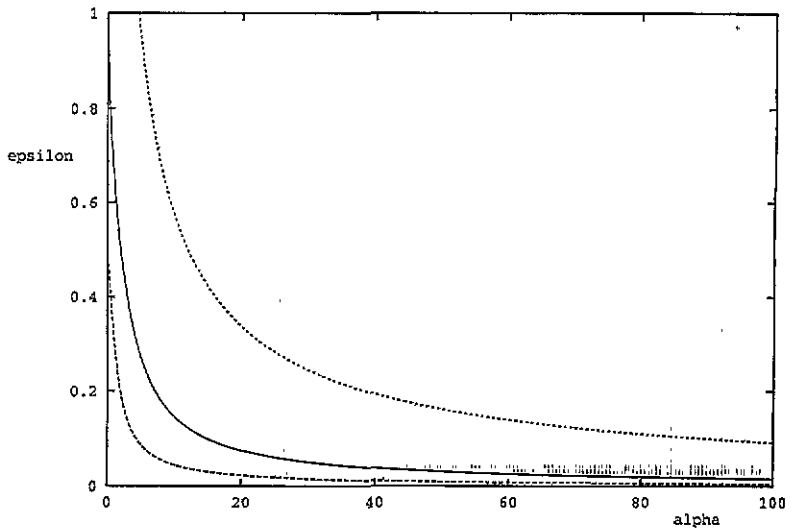

Figure 6. Generalization error as a function of α for the worst (full line) and typical (dashed line) student out of the version space. The dotted line is the rigorous upper bound provided by a variant of the VC theorem for this situation.

The version space is known to be convex and it is therefore tempting to assume that replica symmetry holds for all values of α. However, in order for RS to hold, it is not the version space itself which must be connected but the part of the rims given by $NR = \sum_j J_j T_j$ that belong to it. In figure 7 we have shown schematically for two values of R which part of the respective rims are cut if the first pattern is learnt. Furthermore, it is shown on the right-hand side that for $R < 0$ two patterns will always leave a connected part of the rim in the version space, whereas for $R > 0$ it is possible that two patterns cut the rim in disconnected pieces belonging to $\mathcal{V}$. Hence, for $R > 0$ we would expect that replica symmetry is broken.

We have tested the local stability of the RS solution with the result that it becomes locally unstable for $\alpha > \alpha_{\mathrm{AT}}$ with α_{AT} again given by (32). In the present case this gives $\alpha_{\mathrm{AT}} = 2$. Since $R_{\min}(\alpha = 2) = 0$, this is in perfect agreement with the geometric argument above. To our knowledge, this is the first case where one can verify by an independent geometric argument that replica symmetry breaks down when the solution space splits into disconnected pieces.

In order to quantify the implications of RSB for $\alpha > 2$ we have investigated the solution with one-step RSB. The determination of $R_{\min}(\alpha)$ in section 4 is not easily generalized to

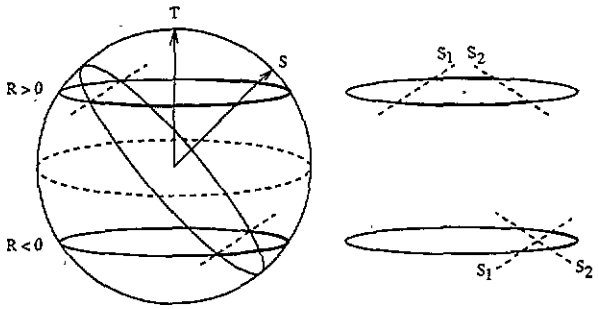

Figure 7. Three-dimensional sketch of the student space. Two rings corresponding to a fixed overlap R with the teacher T are shown. If a pattern S is added to the training set, the hyperplane perpendicular to it cuts that part of the rings which is not compatible with the teacher classification (left). Two patterns always leave a connected part of the ring if $R < 0$. For $R > 0$, however, it is possible that the part of the ring belonging to the version space is disconnected. This gives rise to replica symmetry breaking for $R > 0$ (right).

replica symmetry breaking. It is more convenient to use a different but equivalent approach.† To this end one calculates the typical value of the fractional volume $V(R,\alpha)$ of the part of a rim corresponding to R that belongs to the version space:

$$V(R,\alpha) = \frac{1}{V(R,0)} \int \mathrm{d}J\, \delta(N-(J)^2)\delta(NR - \boldsymbol{J}\cdot\boldsymbol{T}) \prod_{\mu=1}^{\alpha N} \theta\Big(\frac{1}{N}(\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu)\,(\boldsymbol{T}\cdot\boldsymbol{\xi}^\mu)\Big) \tag{42}$$

where the total volume of the rim for large N is given by

$$V(R,0) = \exp\left(\tfrac{1}{2}N[1+\log 2\pi + \log(1-R^2)]\right). \tag{43}$$

For given R, $V(R,\alpha)$ is a monotonously decreasing function of α and there is a threshold value $\alpha_{\max}(R)$ such that $V(R,\alpha_{\max}(R)) = 0$, much like in the original Gardner calculation [24]. $\alpha_{\max}(R)$ is the inverse function of $R_{\min}(\alpha)$. Calculating $V_{\mathrm{typ}}(R,\alpha) \cong \exp(\langle\langle \log V(R,\alpha)\rangle\rangle)$ using the replica trick and replica symmetry, it is easy to find back equation (30) for $R_{\min}(\alpha)$. Using the ansatz of one-step replica symmetry breaking as given in section 5 and introducing the same scaled parameters as introduced there, one gets after some algebra

$$\alpha_{\max}(R) = \min_{q,c}\left[-\{\frac{c(q-R^2)/[1+c(1-q)]\} + \log(1+c(1-q))}{4g(q,c,R)}\right] \tag{44}$$

where the function $g(q,c,R)$ is given by

$$g(q,c,R) = \int \mathrm{D}t\, H\Big(\frac{Rt}{\sqrt{q-R^2}}\Big)\, \log\Big[H\Big(\frac{\sqrt{q}\,t}{\sqrt{1-q}}\Big) + \frac{1}{\sqrt{1+c(1-q)}} \times \exp\Big(-\frac{cqt^2}{2(1+c(1-q))}\Big)\, H\Big(\frac{-\sqrt{q}\,t}{\sqrt{1-q}\sqrt{1+c(1-q)}}\Big)\Big]. \tag{45}$$

Performing the minimization in (44) numerically, one finds that the RSB solution branches off continuously from the RS solution at $\alpha = \alpha_{\mathrm{AT}} = 2$. Moreover, the difference between the two results is very small, at most 1% at $\alpha \cong 5.1$ (cf figure 2). Hence, RS gives a very accurate approximation for $R_{\min}(\alpha)$ and $\epsilon_V(\alpha)$ for all values of α. Most importantly, the asymptotic behaviour for large α is not modified by RSB. The results of the numerical minimization in equation (44) suggest $c(1-q) \to \infty$ and $c(1-q)^2 \to 0$ for $\alpha \to \infty$

† We thank Marc Bouten for pointing this out to us.

from which one can obtain the asymptotic behaviour of $g(q, c, R)$ and finally $\alpha_{\max}(R)$ in a self-consistent way, again with the result $\epsilon_V(\alpha) \sim 3/(2\alpha)$. Since usually the first step of replica symmetry breaking gives the largest correction of the RS result it seems very unlikely that higher-order breakings of replica symmetry will modify this asymptotic behaviour.

In figure 6 we have also included the generalization error $\epsilon_V^{\text{typ}}(\alpha)$ of the *typical* student out of the version space as determined in [2]. The corresponding asymptotic behaviour is $\epsilon_V^{\text{typ}}(\alpha) \sim 0.625/\alpha$, to be compared with $\epsilon_V(\alpha) \sim 3/2\alpha$ for the worst student. We thus find that there is only a factor of 2.4 between the performance of the typical and the worst student. The $\log\alpha$ in the VC bound $\epsilon_V^{\text{VC}} \sim \log\alpha/(\alpha \log 2)$ stems from replacing the maximum of e_J by the respective sum. This is a rather crude upper bound if all terms are of comparable magnitude as in the present case. In the situation where the students are restricted to the version space, the VC bound is therefore not tight but overestimates the generalization error of the worst student (asymptotically) by a factor of $2\log\alpha/(3\log 2)$.

It is possible, however, to improve the VC bound by using inequalities from information theory [14, 25] (see also [5]). One then finds that, irrespective of the distribution of patterns forming the training set, the generalization error of a student vector drawn at random from the version space is asymptotically bounded by $2/\alpha$. Our result, $\epsilon_V(\alpha) \sim 3/2\alpha$, shows that this is indeed an excellent bound and that the generalization performance cannot deteriorate significantly if one uses probability distributions other than (18) for the patterns ξ^μ.

7. Ising student

Some new aspects of the generalization problem can be studied within the simple scenario of two perceptrons if the coupling vectors J and T have to obey additional restrictions besides the spherical constraint (11). A simple example is the so-called Ising perceptron where the couplings are binary $J_j = \pm 1$. The generalization behaviour of the *typical* Ising student is known to be drastically different from the spherical case: at $\alpha_c = 1.245$ a discontinuous (first-order) transition occurs from rather poor ($e \sim 0.26$) to perfect ($e = 0$) generalization [26]. Moreover, considering an Ising student generalizing a spherical teacher, one can study a simple example of an unrealizable rule where for sufficiently large α the version space is empty and the asymptotic value of the generalization error is different from zero ($e(\alpha \to \infty) = 2.06$ in this case [3]). In the present section we will investigate the performance of the *worst* student in these situations.

Following the same line of reasoning as in section 3, we again have to determine the free energy (17) with the modification

$$\int \mathrm{d}\mu_R(J) \to \mathrm{Tr}_{\{J_j\}}\delta(NR - JT) \tag{46}$$

where the trace is over all the 2^N possible student coupling vectors J. Let us first consider the case of a realizable rule where the teacher is an Ising perceptron too. In replica symmetry we get, instead of (23) (see appendix),

$$f^{\text{RS}}(\alpha, \beta, R) = -\frac{1}{\beta}\mathrm{extr}_{q,F,G}\left[-RG - \frac{F}{2}(1-q) + \int \mathrm{D}z \log 2\cosh\left(\sqrt{F}z + G\right)\right.$$
$$\left. +2\alpha \int \mathrm{D}t\, H\left(\frac{Rt}{\sqrt{q-R^2}}\right)\log\left[\mathrm{e}^{-\beta/\alpha} + (1-\mathrm{e}^{-\beta/\alpha})H\left(\frac{\sqrt{q}t}{\sqrt{1-q}}\right)\right]\right]. \tag{47}$$

The minimal training error $\nu_{\min}(\alpha, R)$ is again given by $\lim_{\beta\to\infty} f(\alpha, \beta, R)$. Similarly to the spherical case, there are values of α and R for which $\nu_{\min}(\alpha, R) = 0$. In fact we can determine from (47) the entropy $s^{\text{RS}}(\alpha, \beta, R)$ and find that its zero-temperature value

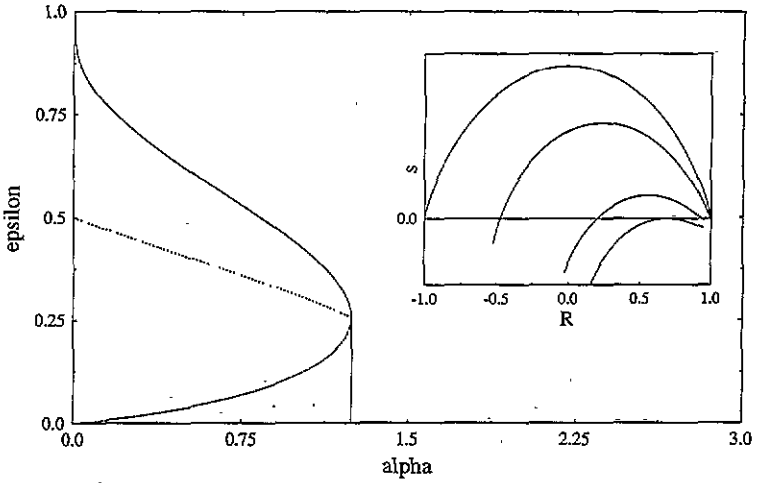

Figure 8. Generalization error of the worst (full line), typical (dotted line) and best (dashed line) Ising student from the version space. The inset shows the ground-state entropy as a function of the overlap R with the teacher for $\alpha = 0$, 0.4, 1.0, and 1.245 (from top to bottom).

$s_0(\alpha, R) = \lim_{\beta\to\infty} s^{RS}(\alpha, \beta, R)$ coincides with $-\lim_{\beta\to\infty} \beta f^{RS}(\alpha, \beta, R)$. This implies that the ground-state energy, i.e. $\nu_{min}(\alpha, R)$, is zero and $s_0(\alpha, R)$ is thus nothing but the logarithm of the number of student vectors J that realize zero training error for a given value of R. We have plotted $s_0(\alpha, R)$ as a function of R for different values of α in the inset of figure 8. Values of R with representatives out of the version space have $s_0(\alpha, R) > 0$. It is clear that the version space shrinks with increasing α and that for all $\alpha > 0$ there is a *gap* between the point $R = 1$ (the teacher) and the remaining R-values in the version space. Hence, in contrast to the spherical case there is now also the *best* student in addition to the worst one. The reason for the fact—surprising at first sight—that one can leave the version space by increasing the overlap with the teacher is as follows. Owing to the discrete nature of the coupling vectors, the number of available students grows exponentially with decreasing overlap. Hence there are few students with large values of R and, although they are classifying most patterns as the teacher, they are likely to be eliminated by a single mistake on the training set. On the other hand, smaller values of the overlap are represented by very many different coupling vectors and, even though these perform worse individually, some of them can survive. If the overlap becomes ultimately too small, the growing number of representations cannot compensate for the degrading individual performance and one is leaving the version space again.

The behaviour of the entropy $s_0(\alpha, R)$ explains the generalization error of the typical, worst and best student, as shown in figure 8 (note that the overlap of the typical students is given by the maximum of $s_0(\alpha, R)$). Starting from 0.5, 1 and 0, respectively, at $\alpha = 0$, the generalization errors converge to each other and meet at $\alpha \sim 1.245$, where the part of the version space different from the teacher disappears. Here the discontinuous transition to perfect generalization takes place and there is no difference between the best, worst and typical student any more since the version space has shrunk to a single point. A generalization error equal to zero trivially obeys any positive bound and hence no interesting comparison with the version space VC theorem, as in section 6, is possible here.

To analyse a more general case where students with non-zero training error are also included, one has to consider R-values for which the (replica-symmetric) ground-state

entropy $s_0(\alpha, R)$ is negative. As is well known [21, 27], a negative entropy in a random system with discrete configuration space signals the breakdown of replica symmetry. In the present case a one-step RSB solution with $q_1 = 1$ exists for $\beta \geqslant \beta_g$, where β_g is the inverse temperature at which the replica-symmetric entropy vanishes:

$$s^{\mathrm{RS}}(\alpha, \beta_g, R) = 0. \tag{48}$$

The low-temperature phase is completely frozen ($q_1 = 1$) and therefore one has [3, 26, 28]

$$\nu_{\min}(\alpha, R) = f^{\mathrm{RSB}}(\alpha, 0, R) = f^{\mathrm{RS}}(\alpha, \beta_g, R). \tag{49}$$

Determining numerically the value $\beta_g(\alpha, R)$ at which the entropy resulting from (47) becomes zero, we can now calculate $\nu_{\min}(\alpha, R)$ and then determine $R_{\mathrm{worst}}(\alpha)$ and $\epsilon(\alpha)$ from (15). Similarly to the spherical case, we find $R_{\mathrm{worst}}(\alpha) < 0$ and $R_{\mathrm{worst}}(\alpha) \to 0$ for $\alpha \to \infty$. The resulting behaviour of $\epsilon(\alpha)$ is shown in figure 9. A comparison with the VC theorem is somewhat ambiguous since, to our knowledge, the VC dimension of the Ising perceptron is not known. It must clearly be smaller than, or equal to, N; a recent numerical study using exact enumeration techniques strongly suggests $d^{\mathrm{VC}}_{\mathrm{Ising}} = N/2$ [29]. We have included in figure 9 both the bound (9) corresponding to $d^{\mathrm{VC}} = N$ and

$$\epsilon^{\mathrm{VC}}_{\mathrm{Ising}}(\alpha) = \sqrt{2\log 2\alpha + \tfrac{1}{2}\alpha \log 2 - (2 - \tfrac{1}{2}\alpha)\log(2\alpha - \tfrac{1}{2})} \tag{50}$$

resulting from $d^{\mathrm{VC}} = N/2$. Our result for $\epsilon(\alpha)$ is well below the values of both expressions and hence the bound is not tight enough to infer the actual value of $d^{\mathrm{VC}}_{\mathrm{Ising}}$ from it.

Let us finally turn to the case of an unrealizable rule by considering an Ising student learning examples provided by a spherical teacher. The replica-symmetric free energy is now given by (see appendix)

$$f(\alpha, \beta, R) = -\frac{1}{\beta}\mathrm{extr}_{q,F}\left[-\frac{R^2}{2(1-q)} - \frac{F}{2}(1-q) + \int \mathrm{D}z \log 2\cosh\left(\sqrt{F}z + G\right) + 2\alpha \int \mathrm{D}t\, H\left(\frac{Rt}{\sqrt{q - R^2}}\right) \log\left[\mathrm{e}^{-\beta/\alpha} + (1 - \mathrm{e}^{-\beta/\alpha}) H\left(\frac{\sqrt{q}t}{\sqrt{1-q}}\right)\right]\right] \tag{51}$$

i.e. it is only slightly different from (47). The differences stem from the fact that one has to explicitly average over the distribution of vectors T (cf appendix). Therefore, one can now study the dependence of the worst-case performance on this prior distribution of concepts to be learned.

Calculating again the corresponding entropy we determine β_g from (48) and use (49). The results for $R_{\mathrm{worst}}(\alpha)$ and $\epsilon(\alpha)$ are practically identical to those obtained for the realizable case discussed above. Slight differences (less than 1%) appear only for small values of α ($\alpha < 1$). Since the resulting curves are indistinguishable from those of figure 9, we did not plot them seperately. The worst-case performance of an Ising student is therefore very similar for a realizable rule and an unrealizable rule given by a spherical teacher. This is not really surprising if one takes into account that for $\alpha > 1$ the worst student has rather small overlap with the teacher. The precise values of the teacher couplings are decisive for student vectors strongly aligned with the teacher but they are not likely to be important for student vectors roughly perpendicular to it.

Interestingly, this argument holds also for the student and we will therefore speculate that the worst-case performance of an Ising student generalizing a spherical teacher should not be much different from that of a spherical student doing the same. In figure 10 we have plotted the values of $R_{\mathrm{worst}}(\alpha)$ and $\epsilon(\alpha)$ for both scenarios. The interesting point is that due to similarities with the random energy model [30, 31], there is some belief that for the

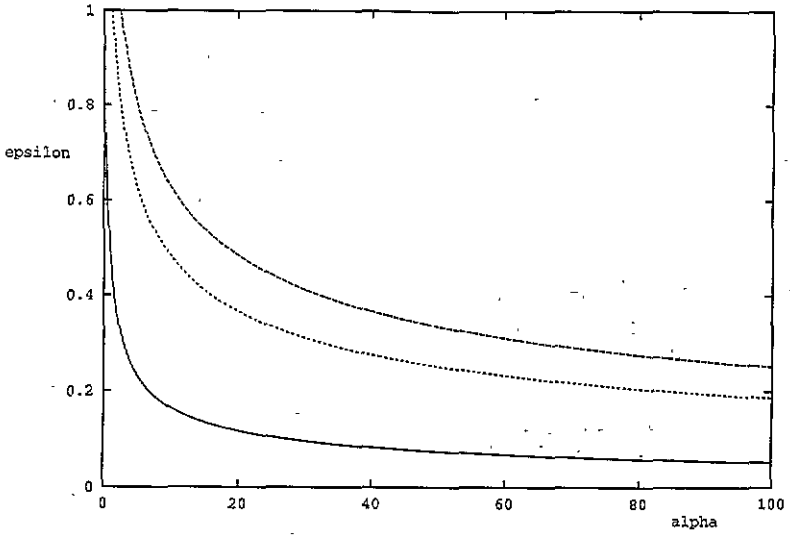

Figure 9. Maximal possible difference between generalization and training error as a function of the training set size α for an Ising student generalizing an Ising teacher. The dashed line is the VC bound corresponding to $d_{\text{Ising}}^{\text{VC}} = N$, the dotted line corresponds to $d_{\text{Ising}}^{\text{VC}} = N/2$.

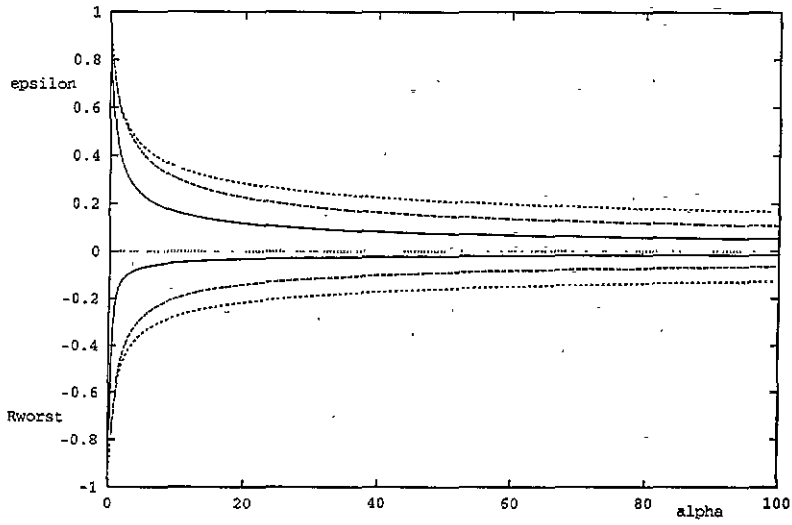

Figure 10. Maximal possible difference between generalization and training error (top), and overlap of the worst student with the teacher (bottom), as calculated from the statistical mechanics approach for a spherical student in replica symmetry (dotted lines), for one-step replica symmetry breaking (dashed lines) and for an Ising student in one-step replica symmetry breaking (full lines).

case of an Ising student the solution in one-step RSB is *exact* [28, 32]. The results of this section obtained on the basis of equation (47) are therefore probably the best estimates for the worst-case performance in the two-perceptron scenario that one can get using statistical mechanics. From figure 10 one infers that the expected modifications from higher-order RSB for the spherical case are quite substantial. In fact, for $\alpha \to \infty$ and $R \to 0$ we find from (51) a self-consistent asymptotic solution $q \to 0.5115$, $F \to 1.097$, $\beta_g \sim 3.21\sqrt{\alpha}$ implying $\epsilon(\alpha) \sim 0.5278/\sqrt{\alpha}$.

8. Summary

The generalization performance of neural network models learning from examples can be characterized in different ways. Whereas investigations using statistical mechanics have so far been concentrating on the typical behaviour studies in mathematical statistics, computer sciences have often focused on the worst-case situation in order to derive uniform convergence bounds for the quantities of interest. In the present paper we have shown how one can calculate the generalization performance of the *worst* student perceptron learning a linearly seperable Boolean function provided by a teacher perceptron with the help of standard techniques of statistical mechanics of neural networks. The worst student is here defined as the one with the largest difference between training and generalization error. Our results describe the *actual* performance of the worst student in the thermodynamic limit; hence, they yield a crucial test for *bounds* on the worst-case behaviour as, for example, provided by the Vapnik–Chervonenkis (VC) theorem.

We have found that for all values of the training set size α the overlap R between the worst student and the teacher is negative. For $\alpha \to \infty$, R tends to zero from below. Moreover, the worst student never belongs to the version space formed by all students with zero training error. This implies that replica symmetry is always broken. In fact, the replica-symmetric result for the difference between the training and generalization errors of the worst student violates the rigorous upper bound provided by the VC theorem for large values of α. The results in one-step replica symmetry breaking obey these bounds, while in the case of a student perceptron with couplings restricted to the values ± 1 (Ising student) they are probably exact. We have also studied the performance of the worst student out of the version space, i.e. the perceptron with the worst generalization ability among those that score perfectly on the training set. Again, replica symmetry is broken, however the quantitative implications are insignificant and the asymptotic behaviour of the generalization error for large values of α remains unchanged. In the case of an Ising student there is a discontinuous transition to perfect generalization at $\alpha = 1.245$ for both the worst and the typical student. Beyond this value the version space only consists of the teacher himself, so that the typical and worst performances become identical.

We have found that the maximal possible difference between the learning and generalization errors decays asmptotically as $0.5278/\sqrt{\alpha}$. If one restricts the class of perceptrons considered to the version space, the maximum possible generalization error decreases with α asymptotically as $3/(2\alpha)$. We thus found that the VC bounds giving $\sqrt{\log \alpha/\alpha}$ and $\log \alpha/\alpha$, respectively, overestimate the worst-case performance by logarithmic factors in the training set size α. This can be traced back to the replacement of the supremum over all perceptrons by the respective sum in the derivation of these bounds, which is a rather crude step in the case where the worst and the typical behaviour are not as dramatically different as seems to be the case in the two-perceptron scenario investigated here. In particular, the VC bound is not tight enough to infer from our calculations the so-far unknown value of the VC dimension of the Ising perceptron.

The investigation of the performance of the worst student is a partial worst-case analysis only, since the worst teacher and the worst possible distribution of training set patterns could be considered as well. As a first step to include the choice of the teacher, we have also studied the difference between learning and generalization errors for the worst student trying to learn an unrealizable rule as given by an Ising student generalizing a spherical teacher. The results are practically identical to the realizable case where an Ising student learns from an Ising teacher. This is probably due to the fact that the overlap R of the worst student is nearly zero, whereas the weight mismatch between teacher and student is known to be

crucial only if they are fairly aligned with each other. It would be interesting to extend these investigations to the case of a nonlinearly separable target rule.

Acknowledgments

We are grateful to Christian van den Broeck for sharing his insight into the subject of uniform convergence bounds with us and for many stimulating discussions. We have also benefitted from discussions with Reiner Kree, Manfred Opper and Wolfgang Kinzel.

Appendix

In this appendix we sketch the calculation of the free energy (17)

$$f(\alpha,\beta,R) = -\lim_{N\to\infty}\frac{1}{\beta N}\left\langle\!\!\left\langle \log\int\prod_i \mathrm{d}J_i\;\delta(N-\boldsymbol{J}^2)\;\delta(NR-\boldsymbol{J}\cdot\boldsymbol{T})\right.\right.$$
$$\left.\left.\times\exp\left[-\frac{\beta}{\alpha}\sum_\mu\theta\left(-\left(\frac{1}{\sqrt{N}}\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu\right)\left(\frac{1}{\sqrt{N}}\boldsymbol{T}\cdot\boldsymbol{\xi}^\mu\right)\right)\right]\right\rangle\!\!\right\rangle \tag{A1}$$

using standard techniques [2, 3, 24]. Employing the replica trick,

$$\langle\!\langle \log Z\rangle\!\rangle = \lim_{n\to 0}(\langle\!\langle Z^n\rangle\!\rangle - 1)/n$$

we have to determine

$$\langle\!\langle Z^n\rangle\!\rangle = \left\langle\!\!\left\langle\int\prod_{i,a}\mathrm{d}J_i^a\;\delta(N-(\boldsymbol{J}^a)^2)\;\delta(NR-\boldsymbol{J}^a\cdot\boldsymbol{T})\right.\right.$$
$$\left.\left.\times\exp\left[-\frac{\beta}{\alpha}\sum_{\mu,a}\theta\left(-\left(\frac{1}{\sqrt{N}}\boldsymbol{J}^a\cdot\boldsymbol{\xi}^\mu\right)\left(\frac{1}{\sqrt{N}}\boldsymbol{T}\cdot\boldsymbol{\xi}^\mu\right)\right)\right]\right\rangle\!\!\right\rangle. \tag{A2}$$

Introducing auxiliary variables $u_\mu = \frac{1}{\sqrt{N}}\boldsymbol{T}\boldsymbol{\xi}^\mu$ and using

$$\exp\left(-\frac{\beta}{\alpha}\sum_{\mu,a}\theta(-y_\mu^a)\right) = \prod_{\mu,a}\left[\theta(y_\mu^a) + \mathrm{e}^{-\beta/\alpha}\theta(-y_\mu^a)\right]$$
$$= \prod_{\mu,a}\left(\int_0^\infty\frac{\mathrm{d}\lambda_\mu^a}{2\pi}\int_{-\infty}^\infty\mathrm{d}x_\mu^a + \mathrm{e}^{-\beta/\alpha}\int_{-\infty}^0\frac{\mathrm{d}\lambda_\mu^a}{2\pi}\int_{-\infty}^\infty\mathrm{d}x_\mu^a\right)\exp(\mathrm{i}x_\mu^a\lambda_\mu^a - \mathrm{i}x_\mu^a y_\mu^a) \tag{A3}$$

one can perform the average over the patterns $\boldsymbol{\xi}^\mu$. The integrand in (A2) then becomes

$$\prod_a\delta(N-(\boldsymbol{J}^a)^2)\;\delta(NR-\boldsymbol{J}^a\cdot\boldsymbol{T})\exp\left(\mathrm{i}\sum_\mu s_\mu u_\mu + \mathrm{i}\sum_{\mu,a}x_\mu^a\lambda_\mu^a - \frac{1}{2}\sum_\mu s_\mu^2\right.$$
$$\left.-R\sum_\mu s_\mu u_\mu\sum_a x_\mu^a - \frac{1}{2}\sum_\mu u_\mu^2(x_\mu^a)^2 - \frac{1}{2}\sum_\mu u_\mu^2\sum_{a,b}x_\mu^a x_\mu^b\frac{1}{N}\sum_j J_j^a J_j^b\right). \tag{A4}$$

If we introduce the order parameter matrix

$$q^{ab} = \frac{1}{N}\sum_j J_j^a J_j^b \tag{A5}$$

the J-integrals decouple from the λ-, x-, u- and s-integrals. Using integral representations for the δ-functions in (A4) and for the new one enforcing (A5), the J integrals factorize in

j and the λ-, x-, u- and s-integrals in μ. The s-integral can now be calculated. In this way one finds

$$\langle\langle Z^n\rangle\rangle = \int \prod_a \frac{\mathrm{d}E^a}{4\pi}\frac{\mathrm{d}G^a}{2\pi}\prod_{a<b}\frac{\mathrm{d}q^{ab}\,\mathrm{d}F^{ab}}{2\pi}\exp\Big(N\Big[\frac{1}{2}\sum_a E^a + \sum_{a<b} F^{ab}q^{ab}R\sum_a G^a + \alpha G(R,q^{ab}) + \frac{1}{N}\sum_j G_j^{(2)}(E^a,F^{ab},G^a)\Big]\Big) \tag{A6}$$

with $G(R,q^{ab})$ given by (22)

$$G(R,q^{ab}) = \log 2\int_0^\infty \mathrm{D}u\int\prod_{a=1}^n\frac{\mathrm{d}\lambda_a\,\mathrm{d}x_a}{2\pi}\exp\Big(\mathrm{i}\sum_a x_a(\lambda_a - uR) - \frac{1}{2}\sum_a x_a^2 - \frac{1}{2}\sum_{(a,b)} x_a x_b q^{ab} + \frac{R^2}{2}\sum_{a,b} x_a x_b - \beta/\alpha\sum_a\theta(-\lambda_a)\Big) \tag{A7}$$

and

$$G_j^{(2)}(E^a,F^{ab},G^a) = \log\int\prod_a \mathrm{d}J^a\exp\Big(-\frac{1}{2}\sum_a E_a(J^a)^2 - \sum_{a<b}F^{ab}J^aJ^b - T_j\sum_a J^aG^a\Big). \tag{A8}$$

In the limit $N\to\infty$, the order parameter integrals in (A6) can be found by the saddle-point method. The saddle-point equations for E^a, F^{ab} and G^a are algebraic and these parameters can be expressed via R and q^{ab}. Performing this substitution, one also realizes that the result depends on the teacher-perceptron vector T only through $\sum_j T_j^2 = N$. Hence a separate average of $f(\alpha,\beta,R)$ over the randomly chosen teacher T is superfluous in this case. Having eliminated E^a, F^{ab} and G^a, the only remaining order parameters are the q^{ab}, so that one ends up with (21).

In the replica symmetry, $q^{ab} = q$, $a\neq b$, and expression (A7) can be simplified as follows:

$$\begin{aligned} G(q,R) &= \log 2\int_0^\infty \mathrm{D}u\int\prod_a\frac{\mathrm{d}\lambda_a\mathrm{d}x_a}{2\pi}\exp\Big(\mathrm{i}\sum_a x_a(\lambda_a-uR) - \frac{1-q}{2}\sum_a x_a^2 - \frac{q-R^2}{2}\Big(\sum_a x_a\Big)^2 - \frac{\beta}{\alpha}\sum_a\theta(-\lambda_a)\Big)\\ &= \log 2\int_0^\infty \mathrm{D}u\int\mathrm{D}t\Big[\int\frac{\mathrm{d}\lambda\,\mathrm{d}x}{2\pi}\exp(\mathrm{i}x(\lambda-uR) - \frac{1-q}{2}x^2 - \mathrm{i}tx\sqrt{q-R^2} - \frac{\beta}{\alpha}\theta(-\lambda))\Big]^n\\ &= 2n\int_0^\infty\mathrm{D}u\int\mathrm{D}t\log\int\frac{\mathrm{d}\lambda}{\sqrt{2\pi(1-q)}}\\ &\quad\times\exp\Big(-\frac{(\lambda-uR-t\sqrt{q-R^2})^2}{2(1-q)} - \frac{\beta}{\alpha}\theta(-\lambda)\Big) \end{aligned} \tag{A9}$$

where we have neglected higher-order terms in n. Substituting $\lambda\to-\lambda$ and using

$$\int_0^\infty\mathrm{D}u\int\mathrm{D}t\,g(uR+t\sqrt{q-R^2}) = \int\mathrm{D}t\,H\Big(-\frac{Rt}{\sqrt{q-R^2}}\Big)g(\sqrt{q}t) \tag{A10}$$

one arrives at (23).

In the case of the Ising student $J_j = \pm 1$, one has to perform the replacement

$$\int \prod_j \mathrm{d}J_j\, \delta(N - J^2) \longrightarrow \mathrm{Tr}_{\{J_j\}} \tag{A11}$$

in (A2) and only expression (A8) for $G_j^{(2)}$ is modified. It now becomes

$$G_j^{(2)}(F^{ab}, G^a) = \log \mathrm{Tr}_{\{J^a\}} \exp\Big(\sum_{a<b} F^{ab} J^a J^b + T_j \sum_a G^a J^a\Big). \tag{A12}$$

If also $T_j = \pm 1$ (realizable rule), one can use the 'gauge' transformation $J^a \longrightarrow T_j J^a$ to make $G_j^{(2)}$ independent of T_j, i.e. as in the spherical case, an explicit average over the randomly chosen teacher is superfluous. In replica symmetry one then gets

$$\begin{aligned} G^{(2)} &= \log \mathrm{Tr}_{\{J^a\}} \exp\left(\frac{F}{2}\Big(\sum_a J^a\Big)^2 - n\frac{F}{2} + G\sum_a J^a\right) \\ &= -n\frac{F}{2} + \log \int \mathrm{D}\, z \mathrm{Tr}_{\{J^a\}} \exp\Big(\Big(\sqrt{F}z + G\Big)\sum_a J^a\Big) \\ &= -n\frac{F}{2} + \log \int \mathrm{D}\, z [2\cosh(\sqrt{F}z + G)]^n \\ &= -n\frac{F}{2} + n\int \mathrm{D}\, z \log 2\cosh\Big(\sqrt{F}z + G\Big) + \mathrm{O}(n^2) \end{aligned} \tag{A13}$$

which together with (A9) gives (47).

In the unrealizable case one has to explicitly average over the teacher couplings using $P(\{T_j\}) \sim \delta(\sum_j T_j^2 - N)$. For $N \to \infty$ this is equivalent to $P(\{T_j\}) \sim \exp(-\sum_j T_j^2/2)$ and one finds in replica symmetry

$$G^{(2)}(F, G) = -n\frac{F}{2} + n\int \mathrm{D}\, z \log 2\cosh\Big(\sqrt{F + G^2}z\Big). \tag{A14}$$

Using $F + G^2 \longrightarrow F$, the saddle-point equation for G becomes algebraic, $G = R/(1-q)$, after eliminating G, one is left with (51).

References

[1] Gardner E and Derrida B 1989 *J. Phys. A: Math. Gen.* **22** 1983

[2] Györgyi G and Tishby N 1990 *Workshop on Neural Networks and Spin Glasses* ed by K Theumann and W K Koeberle (Singapore: World Scientific)

[3] Seung M S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056

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

[5] Opper M and Kinzel W 1993 Statistical Mechanics of Generalization *Preprint University of Würzburg*

[6] Shorak G R and Welton J A 1986 *Empirical processes with applications to statistics* (Cambridge: Cambridge University Press)

[7] Valiant L G 1984 *Commun. ACM* **27** 1134

[8] Vapnik V N and Chervonenkis A Y 1971 *Th. Prob. Appl.* **16** 264

[9] Vapnik V N 1982 *Estimation of dependences based on empirical data* (Berlin: Springer)

[10] Baum E and Haussler D 1989 *Neural Computation* **1** 151

[11] Blumer A, Ehrenfeucht A, Haussler D and Warmuth M K 1989 *J.A.C.M.* **36** 929

[12] Anthony M and Biggs N 1992 *Computational Learning Theory* (Cambridge: Cambridge University Press)

[13] Parrondo J M and Van den Broeck C 1993 *J. Phys. A: Math. Gen.* **26** 2211

[14] Haussler D, Kearns M and Schapire R 1991 Bounds on the sample complexity of Bayesian learning using information theory and the vc dimension *Proceedings COLT '91* (San Mateo: Morgan Kaufmann)

[15] Engel A and van den Broeck C 1993 *Phys. Rev. Lett.* **71** 1772

[16] Sauer N 1972 *J. Comb. Th.* A **13** 145
[17] Cover T M 1965 *IEEE Trans. El. Comp.* **14** 326
[18] Opper M, Kinzel W, Kleinz J and Nehl R 1990 *J. Phys. A: Math. Gen.* **23** L581
[19] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[20] Griniasty M and Gutfreund H 1991 *J. Phys. A: Math. Gen.* **24** 715
[21] de Almeida J R L and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 129
[22] Erichsen R Jr and Theumann W K 1993 *J. Phys. A: Math. Gen.* **26** L61
[23] Majer P, Engel A and Zippelius A 1993 Perceptrons above saturation *Preprint University of Göttingen* (submitted to *J. Phys. A: Math. Gen.*)
[24] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[25] Vapnik V N 1979 *Theorie der Zeichenerkennung* (Berlin: Akademie)
[26] Györgyi G 1990 *Phys. Rev. Lett.* **64** 2957
[27] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
[28] Krauth W and Mezard M 1989 *J. Phys. France* **50** 3057
[29] Stambke G 1992 *Diploma Thesis* University of Giessen
[30] Derrida B 1981 *Phys. Rev.* B **24** 2613
[31] Gross D J, Mezard M 1984 *Nucl. Phys.* B **240** 431
[32] Barkai E, Kanter I 1991 *Europhys. Lett.* **14** 107