Gradient descent learning in perceptrons: A review of its possibilities

M. Bouten, J. Schietse, and C. Van den Broeck
Limburgs Universitair Centrum, B-3590 Diepenbeek, Belgium
(Received 28 November 1994)

We present a streamlined formalism which reduces the calculation of the generalization error for a perceptron, trained on random examples generated by a teacher perceptron, to a matter of simple algebra. The method is valid whenever the student perceptron can be identified as the unique minimum of a specific cost function. The asymptotic generalization error is calculated explicitly for a broad class of cost functions, and a specific cost function is singled out that leads to a generalization error extremely close to the one of the Bayes classifier.

PACS number(s): 87.10.+e, 02.50.−r, 75.10.Nr

I. INTRODUCTION

Many of the tasks that we routinely perform during our daily activity have been learned through experience. These tasks may appear to be simple, yet it is often very difficult to find algorithms on the basis of which they can be carried out reliably. This observation has prompted the search for systems that can learn from examples. Prominent among these are the so-called neural networks. To gain more insight into the mechanism of learning, a very simple scenario, namely, that of a student perceptron learning from examples generated by a teacher perceptron, has been investigated in great detail using the powerful techniques of statistical mechanics, see, e.g., [1–7]. This scenario can also be used as a test ground for new ideas. Furthermore, these theoretical insights can often be applied to multilayer architectures, especially so when the training algorithm is dealing with the separate perceptrons that constitute the network, see e.g. [8,9].

In this paper we will review the performance of gradient descent algorithms for the perceptron on the basis of a unified and streamlined presentation, which complements the one given recently for the capacity problem [10,11]. We will focus on the case of a cost function with a unique nondegenerate minimum. When this condition is met, the calculation of the generalization error and other quantities of interest such as the cost or the overlap distribution is a matter of simple algebra. The local stability of the replica symmetric solution, can be verified by the evaluation of a simple integral. The same results can be obtained using a cavity approach with the additional advantage that the parameters appearing in the replica calculations acquire a simple physical interpretation. Apart from illustrating the calculations for cost functions that have been discussed previously in the literature, we consider a general class for which we calculate the generalization error in detail. The aim is to identify a cost function for which a simple gradient descent algorithm can be applied, but which gives a lower generalization error than the currently used perceptron with optimal stability (which can be found through the adatron algorithm [12]). This purpose is realized beyond expectation since we find a cost function with a nondegenerate minimum that has a generalization error lying within 0.5% of the lowest possible result as given by the Bayes rule [13]. We also consider a class of cost functions for which the generalization error is expected to be large even though the training examples are correctly reproduced. This case is of interest in the more theoretical context of the so-called worst case scenario, which has been considered in detail in the computational science literature see, e.g., [14–16].

II. REPLICA CALCULATION AND CAVITY INTERPRETATION

We first briefly review the teacher-student perceptron scenario. A teacher perceptron, characterized by an N-dimensional weight vector $\mathbf{T}$, returns the classification

$$\xi_0^\mu = \mathrm{sgn}\left[\frac{\mathbf{T}\cdot\boldsymbol{\xi}^\mu}{\sqrt{N}}\right]$$

on a set of randomly selected training patterns $\boldsymbol{\xi}^\mu, \mu = 1, \ldots, p$. On the basis of this information one would like to select a student perceptron, with weight vector $\mathbf{J}$, such that it reproduces as closely as possible the classification of the teacher. One of the most common and practical procedures to select a student vector is to require that it gives the minimum value of an appropriately chosen cost function $E(\mathbf{J})$. If the minimum is unique, this $\mathbf{J}$ vector can be obtained by applying a gradient descent algorithm. We will restrict ourselves in this paper to a cost function in which the information about the different training patterns enters in an additive way:

$$E(\mathbf{J}) = \sum_{\mu=1}^{p} V(\lambda^\mu) \tag{1}$$

with

$$\lambda^\mu = \frac{\mathbf{J}\cdot\boldsymbol{\xi}^\mu \xi_0^\mu}{\sqrt{N}} . \tag{2}$$

This choice leads to a cost function which is extensive in the number p of patterns, which itself is chosen propor-

tional to the dimensionality N of the input space. The physics of the problem will involve the competition of this cost with an entropic term, namely, the number of $\mathbf{J}$ vectors that correspond to a given value of the cost. With the normalization condition $\mathbf{J}^2=N$, this number is exponentially large in N so that the corresponding entropy (being the log of this number) is also extensive. For simplicity of the calculations we also follow the convention that the other N-dimensional vectors such as $\mathbf{T}$ and ξ^μ also have a length equal to $\sqrt{N}$.

The basic question at hand is how the generalization performance depends on the choice of the "potential" function V. This performance is usually quantified by introducing the so-called generalization error $\varepsilon(\mathbf{J})$ defined as the probability that student and teacher disagree on a randomly chosen question $\mathbf{S}$. As is well known, the generalization error is essentially determined by the angle between student and teacher [3,4]:

$$\varepsilon(\mathbf{J})=\frac{1}{\pi}\arccos R \tag{3}$$

with

$$R=\frac{\mathbf{J}\cdot\mathbf{T}}{N} . \tag{4}$$

The following asymptotic forms are useful:

$$\varepsilon \underset{R\to 0}{\sim} \frac{1}{2}-\frac{R}{\pi}, \quad \varepsilon \underset{R\to 1}{\sim} \frac{\sqrt{1-R^2}}{\pi} .$$

In order to calculate ε, the formalism of statistical mechanics turns out to be a very elegant and powerful tool. One can associate the following partition function to the energy function E:

$$Z=\int d\mathbf{J}\, e^{-\beta E(\mathbf{J})}\delta(\mathbf{J}^2-N) . \tag{5}$$

This partition function is a random variable through its dependence on the randomly chosen training patterns. However, the corresponding free energy $F=Nf$ is extensive and expected to be self-averaging in the limit $N\to\infty$, $p\to\infty$ with $\alpha=p/N$ fixed:

$$-\beta N f=\ln Z=\langle \ln Z\rangle . \tag{6}$$

In this limit, the average over the patterns can be performed through the replica technique and one finds under the assumption of replica symmetry that

$$f=-\mathrm{Extr}_{q,R}\left\{\frac{q-R^2}{2\beta(1-q)}+\frac{\ln(1-q)}{2\beta}+\frac{\alpha}{\beta}\int_{-\infty}^{+\infty}Dt_1\int_{-\infty}^{+\infty}Dt_2\ln\int_{-\infty}^{+\infty}\frac{d\lambda}{\sqrt{2\pi(1-q)}}e^{-\beta V[\lambda\,\mathrm{sgn}(t_2)]-(\lambda-Rt_2-\sqrt{q-R^2}t_1)^2/2(1-q)}\right\} . \tag{7}$$

The meaning of the order parameters q and R is as usual: q is the overlap between two typical $\mathbf{J}$ vectors and R is the overlap between a typical $\mathbf{J}$ vector and the teacher vector $\mathbf{T}$. The word typical refers to the $\mathbf{J}$ vectors that give the exponentially dominant contribution to the free energy. The generalization error is obtained by inserting the value of R into Eq. (3).

In order to find the ground state we take the limit $\beta\to\infty$. As mentioned in the Introduction we will concentrate on the situation in which this ground state is nondegenerate. In this case, the calculations are extremely simplified. Indeed, if there is a unique minimum, the overlap q between the typical $\mathbf{J}$ vectors has to converge to 1, hence $q\to 1$, and the free energy thus reduces to the following simple expression:

$$f^{T=0}=e=-\mathrm{Extr}_{x,R}\left\{\frac{1-R^2}{2x}-2\alpha\int_{-\infty}^{+\infty}Dt_1\int_0^{\infty}Dt_2\min_\lambda\left[V(\lambda)+\frac{(\lambda-t)^2}{2x}\right]\right\} \tag{8}$$

with

$$x=\lim_{\beta\to\infty}\beta(1-q) , \tag{9}$$

$$t=Rt_2+\sqrt{1-R^2}t_1 , \tag{10}$$

and $E=Ne$ is the ground state cost value.

In order to make the connection with the cavity approach [17], which will allow us to give physical meaning to the parameters λ, t, and x, we review the steps that lead to the solution contained in Eq. (8) in more detail as follows.

(1) Find the function $\lambda_0(t,x)$ which minimizes

$$V(\lambda)+\frac{(\lambda-t)^2}{2x} . \tag{11}$$

(2) The values of R and x in function of α are obtained from the extremum conditions (saddle point equations):

$$2\int_{-\infty}^{\infty}Dt_1\int_0^{\infty}Dt_2(\lambda_0-t)\frac{\partial t}{\partial R}=\frac{R}{\alpha} , \tag{12}$$

$$2\int_{-\infty}^{+\infty}Dt_1\int_0^{\infty}Dt_2(\lambda_0-t)^2=\frac{1-R^2}{\alpha} . \tag{13}$$

A more useful form of these equations is obtained by an orthogonal transformation on the integration variables:

$$\sqrt{2/\pi}\int_{-\infty}^{\infty}Dt\,\lambda_0(\sqrt{1-R^2}\,t,x)=\frac{R}{\alpha} , \tag{14}$$

$$2\int_{-\infty}^{+\infty}Dt\,H\left[-\frac{Rt}{\sqrt{1-R^2}}\right][\lambda_0(t,x)-t]^2=\frac{1-R^2}{\alpha} . \tag{15}$$

As usual, Dt stands for the Gaussian measure $Dt=(dt/\sqrt{2\pi})e^{-t^2/2}$ and $H(u)=\int_u^\infty Dt$.

(3) The ground state cost value is given by

$$e=2\alpha\int_{-\infty}^{+\infty}Dt_1\int_0^\infty Dt_2 V[\lambda_0(t,x)]\ . \tag{16}$$

The function $\lambda_0(t,x)$, which minimizes (11), is identical to the one obtained in the analogous treatment of the capacity problem [10,11,18]. Its interpretation can be clarified as follows by applying the cavity method proposed by [17], see also [20,21]. Let $\mathbf{J}^*$ be the vector that minimizes the cost function E (or its intensive counterpart e) for a given set of p patterns $\xi^\mu,\mu=1,\ldots,p$. Consider a new randomly chosen pattern ξ, with classification $\xi_0=\mathrm{sgn}(\mathbf{T}\cdot\xi)$ and define the vector $\mathbf{u}=\xi\xi_0$. In order to find the new vector $\mathbf{J}=\mathbf{J}^*+\delta\mathbf{J}$ that minimizes the cost function including the new pattern we can proceed in two steps. First, we restrict ourselves to the search of an optimal $\mathbf{J}$ vector that has a given overlap $\lambda=(\mathbf{u}\cdot\mathbf{J})/\sqrt{N}$ with $\mathbf{u}$. The contribution to the cost of the new pattern is then clearly equal to $V(\lambda)$, cf. Eq. (1). However, since we expect that $\mathbf{J}$ will slightly deviate from $\mathbf{J}^*$, there is also an increased contribution to the cost arising from the original patterns, which can be evaluated by Taylor expansion of $E(\mathbf{J})$ around $\mathbf{J}^*$. Solving this simple variational problem (subject to the spherical constraint $\mathbf{J}^2=N$), one finds that the minimal value of this contribution is equal to $(\lambda-t)^2/2x$, with $t=(\mathbf{u}\cdot\mathbf{J}^*)/\sqrt{N}$ the overlap with the vector $\mathbf{J}^*$ prior to training of the new pattern, while x is a measure of the steepness of the cost function in the vicinity of $\mathbf{J}^*$. In the second step we find the optimal value of $\lambda=\lambda_0$ by minimizing $V(\lambda)+(\lambda-t)^2/2x$. We thereby recover the result (11), but with the additional insight that λ_0 is the overlap of the vector $\mathbf{u}$ with the $\mathbf{J}$ vector after training, given as a function of the overlap t prior to training. Note that all the patterns play a completely symmetric role so that this interpretation applies to any one of them. In the capacity problem [17], where the factor ξ_0 is randomly equal to $+1$ or -1, t is the overlap between $\mathbf{J}^*$ and a random vector $\mathbf{u}=\xi\xi_0$, hence it is a Gaussian random variable. In the case considered here, the vectors $\mathbf{u}$ and $\mathbf{J}^*$ are correlated through their dependence on the $\mathbf{T}$ vector. This correlation can be taken into account by decomposing ξ into its components parallel and orthogonal to $\mathbf{T}$:

$$\xi=\underbrace{\frac{\xi\cdot\mathbf{T}}{N}\mathbf{T}}_{\xi_\parallel}+\underbrace{\left[\xi-\frac{\xi\cdot\mathbf{T}}{N}\mathbf{T}\right]}_{\xi_\perp}\ . \tag{17}$$

One thus finds that

$$t=\frac{\mathrm{sgn}(\xi\cdot\mathbf{T})\xi\cdot\mathbf{J}^*}{\sqrt{N}}=Rt_2+\sqrt{1-R^2}t_1\ . \tag{18}$$

Here $R=(\mathbf{J}^*\cdot\mathbf{T})/N$ is the overlap between $\mathbf{J}^*$ and $\mathbf{T}$, $t_2=[\xi\cdot\mathbf{T}\,\mathrm{sgn}(\xi\cdot\mathbf{T})]/\sqrt{N}$ is the absolute value of a normal random variable, while $\sqrt{1-R^2}t_1=[\mathrm{sgn}(\xi\cdot\mathbf{T})\xi_\perp\cdot\mathbf{J}^*]/\sqrt{N}$ is an independent Gaussian random variable (since the orientation of $\xi_\perp$ is uncorrelated from that of $\mathbf{J}^*$ and $\xi_\parallel$) with second moment $1-R^2$. These results again agree with the ones obtained from the replica calculation, cf. the integration over the normal random variable t_1 from $-\infty$ to ∞ and the integration over the normal random variable t_2 from 0 to ∞ with an extra factor of 2 in Eq. (8).

In view of the physical meaning of the overlap λ_0, and of the permutation symmetry of all the training patterns, one can infer the following additional properties.

(4) Since λ_0 is a known function of the normal random variables t_1 and t_2 (and a known function of α through its dependence on x, but we do not write this dependence explicitly in the following for simplicity of notation), one obtains the following result for the probability density of this overlap (also called the aligned field [11,19]):

$$P(\lambda)=2\int_{-\infty}^{+\infty}Dt_1\int_0^\infty Dt_2\delta[\lambda-\lambda_0(t)]\ . \tag{19}$$

(3)′ The ground state cost can be rewritten under the following physically appealing form:

$$e=\alpha\int V(\lambda)P(\lambda)d\lambda=\lim_{N\to\infty}\frac{1}{N}\sum_{\mu=1}^{p}V(\lambda^\mu) \tag{20}$$

which is merely expressing the fact that this quantity is self-averaging.

(5) The training error ν is defined as the fraction of misclassified training patterns. Such a pattern clearly corresponds to an overlap $\lambda_0<0$. Therefore

$$\nu=\int_{-\infty}^{0}P(\lambda)d\lambda\ . \tag{21}$$

(6) It is also possible to perform a stability analysis of the replica symmetric solution. The resulting Almeida Thouless condition [22] for local stability of the replica symmetric solution takes on the following simple form in terms of the overlap λ_0:

$$2\alpha\int_{-\infty}^{+\infty}Dt\,H\left[-\frac{Rt}{\sqrt{1-R^2}}\right][\lambda_0'(t)-1]^2<1\ . \tag{22}$$

Note that a sufficient condition for symmetry breaking [23] is the presence of a discontinuity of the alignment $\lambda_0(t)$ as a function of t.

III. REVIEW OF PREVIOUS RESULTS

The results from the previous section can be applied to situations that have been discussed previously in the literature, see, e.g., [24], and that we now pass in revue. Although they do not fit in the scheme of an energy function with a nondegenerate minimum, we have included for later comparison the results of the Gibbs, Bayes, and worst case rule.

A. Hebb rule

For the specific choice

$$V(\lambda)=-\lambda \tag{23}$$

the minimum of the cost function $E(\mathbf{J})$ is known explicitly, namely,

$$\mathbf{J} \sim \sum_{\mu=1}^{p} \xi^{\mu} \xi_0^{\mu} \tag{24}$$

with the proper normalization constant ($\mathbf{J}^2 = N$). This choice corresponds to the familiar Hebb rule. In this case we find

$$\lambda_0(t,x) = t + x \tag{25}$$

and

$$R = \left[\frac{2\alpha}{2\alpha + \pi} \right]^{1/2} , \tag{26}$$

$$x = \left[\frac{\pi}{2} \right]^{1/2} \frac{R}{\alpha} , \tag{27}$$

which agrees with the results of [25]. We also mention the small and large α behaviors of the generalization error:

$$\begin{aligned} \varepsilon_H(\alpha) &= \frac{1}{2} - \frac{\sqrt{2\alpha}}{\pi^{3/2}} + O(\alpha) , \\ \varepsilon_H(\alpha) &\underset{\alpha \to \infty}{\sim} \frac{1}{\sqrt{2\pi\alpha}} \sim \frac{0.40}{\sqrt{\alpha}} . \end{aligned} \tag{28}$$

B. Adaline rule

This rule corresponds to gradient descent on the following potential function:

$$V(\lambda) = \tfrac{1}{2}(\lambda - K)^2 \quad (K > 0) . \tag{29}$$

One finds

$$\lambda_0(t,x) = \frac{t + Kx}{1 + x} \tag{30}$$

while the equations for x and R read

$$\begin{aligned} 1 - R^2 &= \alpha \left[\frac{x}{1+x} \right]^2 \left[K^2 + 1 - 2 \left[\frac{2}{\pi} \right]^{1/2} KR \right] , \\ R &= \left[\frac{2}{\pi} \right]^{1/2} \alpha \frac{x}{1+x} K . \end{aligned} \tag{31}$$

For a fixed value of K, these equations have no acceptable solution (i.e., $x \geq 0$ and $0 \leq R \leq 1$) for $\alpha < \alpha_c$ given by

$$\alpha_c = \frac{\pi}{4} \left[\frac{1 + K^2}{K^2} - \frac{\sqrt{(K^2+1)^2 - 8K^2/\pi}}{K^2} \right] . \tag{32}$$

In particular, one has that $\alpha_c = 1$ for $K = 0$. The breakdown of our formalism for $\alpha < \alpha_c$ is due to the fact that the minimum of $E(\mathbf{J})$ is degenerate in this case. For $\alpha \geq \alpha_c$, Eqs. (31) admit a unique solution that can be obtained analytically by identifying the physical acceptable solution of the third order equation for R. We only mention here the resulting asymptotic behavior

$$\varepsilon_A(\alpha) \underset{\alpha \to \infty}{\sim} \frac{1}{\sqrt{2\pi\alpha}} \frac{\sqrt{K^2 + 1 - 2\sqrt{2/\pi} K}}{K} . \tag{33}$$

This result differs from the Hebb rule by a factor with minimum value $\sqrt{1 - 2/\pi} \sim 0.6$, which is obtained for $K = \sqrt{\pi/2}$.

C. Pseudo-inverse rule

The $\mathbf{J}$ vector is obtained by minimizing a cost function similar to the one for the adaline, namely, [26]

$$E_{\mathrm{PI}}(\mathbf{J}) = \sum_{\mu=1}^{p} (\mathbf{J} \cdot \xi^{\mu} - \xi_0^{\mu})^2 \tag{34}$$

however without the normalization constraint $\mathbf{J}^2 = N$. The relation with the Adaline rule is clarified by looking for the minimum of this cost function in two steps, one in which the normalization of $\mathbf{J}^2$ is kept fixed to the value $1/K^2$ and the second one in which one minimizes over the choice of K. By a further change of variable $\mathbf{J} \to \mathbf{J}/(K\sqrt{N})$, one recovers the usual normalization condition $\mathbf{J}^2 = N$:

$$\mathrm{Min}_{\mathbf{J}} \{ E_{\mathrm{PI}}(\mathbf{J}) \} = \mathrm{Min}_K \left\{ \frac{1}{K^2} \mathrm{Min}_{\substack{\mathbf{J} \\ (\mathbf{J}^2 = N)}} \sum_{\mu} \left[\frac{\mathbf{J} \cdot \xi^{\mu} \xi_0^{\mu}}{\sqrt{N}} - K \right]^2 \right\} . \tag{35}$$

The partial minimization of the cost with K held constant is realized by the Adaline rule. The minimization with respect to K leads to the following optimal value of this parameter:

$$K = \left[\frac{\alpha - 1}{1 + \frac{2}{\pi}(\alpha - 2)} \right]^{1/2} \quad (\alpha > 1) . \tag{36}$$

Again the restriction to values $\alpha > 1$ follows from the fact that otherwise the ground state, corresponding to the solutions of $E_{\mathrm{PI}}(\mathbf{J}) = 0$, is degenerate. Note that due to the extra factor in $1/K^2$, the pseudo-inverse rule is not identical to the Adaline rule for an optimal choice of the parameter K. For $\alpha \to \infty$ however, we obtain that $K = \sqrt{\pi/2}$, and the pseudo-inverse coincides asymptotically with the best value of the adaline rule.

D. Maximum stability

The quantity λ^{μ} quantifies "conviction" with which the classification of the student perceptron $\mathbf{J}$ agrees with that of the teacher on example ξ^{μ}. Following the criterion of "maximum stability" (this name was coined in the context of the capacity problem, where it is also called the optimal perceptron) one looks for the $\mathbf{J}$ vector such that $\lambda^{\mu} \geq \kappa, \forall \mu$ for the largest possible value of κ. This can be realized by considering the following potential:

$$V(\lambda) = \begin{cases} \infty , & \lambda < \kappa , \\ 0 , & \lambda \geq \kappa , \end{cases} \tag{37}$$

and find the largest possible value of κ for which there exists a solution $\mathbf{J}$ with cost equal to zero. There are several ways to obtain the corresponding generalization error from our formalism. The simplest, although not the most transparent one, is to work directly with the above poten-

tial. One finds

$$\lambda_0(x,t)\begin{cases}=\kappa, & t\le\kappa\ ,\\ =t, & t\ge\kappa\ .\end{cases} \tag{38}$$

By inserting this result into the saddle point equations, Eqs. (12) and (13), one finds that the variable x disappears altogether and one obtains two equations determining the value of R and κ:

$$2\int Dt_1\int_{\kappa\le t}Dt_2(\kappa-t)\frac{\partial t}{\partial R}=\frac{R}{\alpha}\ , \tag{39}$$

$$2\int Dt_1\int_{\kappa\le t}Dt_2(\kappa-t)^2=\frac{1-R^2}{\alpha}\ . \tag{40}$$

We recall that t is defined in Eq. (10). These equations are identical to those given in [26]. The equation for κ arises from the fact that there is a unique value of this parameter, namely, precisely the one corresponding to "optimal stability," for which our formalism applies (i.e., there is a nondegenerate ground state with zero value of the cost). In order to obtain the generalization error, one has to numerically search for the solution of Eqs. (39) and (40). The asymptotic behavior can be found analytically:

$$\varepsilon_{\rm MS}(\alpha)\underset{\alpha\to\infty}{\sim}\frac{c}{\alpha\int_{-\infty}^{1}du\,(1-u)e^{-u^2/2c}}\sim\frac{0.5005}{\alpha} \tag{41}$$

with the constant c determined by the following transcendental equation:

$$\left[\frac{c}{2\pi}\right]^{1/2}=\frac{\int_{-\infty}^{1}du\,(1-u)^2H\left[-\frac{u}{\sqrt{c}}\right]}{\int_{-\infty}^{1}du\,(1-u)e^{-u^2/2c}}\ . \tag{42}$$

E. Gibbs rule

The version space is defined as the set of all **J** vectors that classify correctly the examples from the training set. The Gibbs rule corresponds to choosing at random a **J** vector from this space. The corresponding potential forbids the existence of errors, i.e., $\lambda^\mu<0$ is not allowed:

$$V(\lambda)=\begin{cases}\infty, & \lambda<0\ ,\\ 0, & \lambda\ge0\ .\end{cases} \tag{43}$$

Note that the ground state is degenerate so that one has to go back to the original formulas, cf. Eq. (7), to find the typical overlap $R_G(\alpha)$. The result is given by the solution of the following transcendental equation [1,3]:

$$\frac{R}{\sqrt{1-R}}=\frac{\alpha}{\pi}\int_{-\infty}^{\infty}Dt\frac{e^{-Rt^2/2}}{H(\sqrt{R}\,t)}\ . \tag{44}$$

The corresponding small and large α behaviors for the generalization error are, respectively,

$$\varepsilon_G(\alpha)=\frac{1}{2}-\frac{2\alpha}{\pi^2}+O(\alpha^2)\ ,$$
$$\varepsilon_G(\alpha)\underset{\alpha\to\infty}{\sim}\frac{1}{\alpha}\frac{2}{\int_{-\infty}^{\infty}Dt\frac{1}{H\left[\frac{t}{\sqrt{2}}\right]}}\sim\frac{0.625}{\alpha}\ . \tag{45}$$

F. Bayes rule

In the Bayes rule one classifies a new pattern following the majority vote of all the **J** vectors in the version space. It seems *a priori* unlikely that this rule can be represented by a specific member of the version space, which is the same independent of which question is being asked. Surprisingly, it turns out [27] that there is indeed such a member, namely, the perceptron with as **J** vector the center of mass of all the vectors from the version space. Because of convexity of this space, this perceptron is also a member of the version space. Note that this vector can in principle be identified on the basis of the information provided by the training examples, since the latter determine the version space unequivocally, but we are not aware of any fast converging algorithm that will do so. The generalization error for the Bayes rule was first derived in [13] and is given by

$$\varepsilon_B(\alpha)=\frac{1}{\pi}\arccos\sqrt{R_G(\alpha)} \tag{46}$$

with the following asymptotic results:

$$\varepsilon_B(\alpha)=\frac{1}{2}-\frac{\sqrt{2\alpha}}{\pi^{3/2}}+O(\alpha) \tag{47}$$

which is identical to the result obtained for the Hebb rule, cf. Eq. (28) and

$$\varepsilon_B(\alpha)\underset{\alpha\to\infty}{\sim}\frac{\varepsilon_G(\alpha)}{\sqrt{2}}\sim\frac{0.442}{\alpha}\ . \tag{48}$$

G. Worst case rule

The worst student from the version space is the one that has the smallest overlap with the teacher. This student can be identified if one knows both the version space and the teacher. One of the questions that we will address below is how to find a worst student exclusively based on the knowledge of the training set. The generalization error of the worst student was calculated in [28]. For large α values, a one-step replica symmetry breaking calculation predicts

$$\varepsilon_W(\alpha)\underset{\alpha\to\infty}{\sim}\frac{3}{2\alpha}\ . \tag{49}$$

IV. BOUNDS FOR THE GENERALIZATION ERROR ASSOCIATED TO A MONOTONIC POTENTIAL

In the next sections we will consider two classes of cost functions based on potentials, defined within the version space, that are monotonic increasing or decreasing functions of λ. For such potentials, it is possible to derive a bound for the corresponding generalization error. Consider first the case of a monotonic decreasing potential:

$$V(\lambda)=\infty,\quad \lambda<0\ , \tag{50}$$

$$V'(\lambda)<0,\quad \lambda>0\ . \tag{51}$$

The function $\lambda_0(t,x)$ that minimizes (11) is a solution of the following equation:

$$\lambda - t = -xV'(\lambda) \tag{52}$$

provided $\lambda \geq 0$ and it is zero otherwise. Consequently one finds that

$$\lambda_0(t,x) \geq 0, \quad t \leq 0 , \tag{53}$$

$$\lambda_0(t,x) \geq t, \quad t \geq 0 . \tag{54}$$

Using this in the saddle point equation Eq. (14) yields

$$\left[\frac{\pi}{2}\right]^{1/2} \frac{R}{\alpha} = \int_{-\infty}^{\infty} Dt\, \lambda_0(t\sqrt{1-R^2},x) \tag{55}$$

$$\geq \int_0^{\infty} Dt\, t\sqrt{1-R^2} = \frac{\sqrt{1-R^2}}{\sqrt{2\pi}} . \tag{56}$$

Hence

$$R^2 \geq \frac{\alpha^2}{\alpha^2+\pi^2} \tag{57}$$

and thus

$$\varepsilon(\alpha) \leq \varepsilon^*(\alpha), \quad \forall \alpha > 0 , \tag{58}$$

for every monotonic decreasing potential in the version space with

$$\varepsilon^*(\alpha) = \frac{1}{\pi} \arccos\left[\frac{\alpha^2}{\pi^2+\alpha^2}\right]^{1/2} . \tag{59}$$

In a similar way one finds that for a monotonic increasing potential of the type

$$V(\lambda) = \infty, \quad \lambda < 0 , \tag{60}$$

$$V'(\lambda) > 0, \quad \lambda > 0 , \tag{61}$$

that

$$\lambda_0(t,x) = 0, \quad t \leq 0 , \tag{62}$$

$$\lambda_0(t,x) \leq t, \quad t \geq 0 , \tag{63}$$

and it easily follows that the generalization error $\varepsilon(\alpha)$ is bounded from below by the same $\varepsilon^*(\alpha)$:

$$\varepsilon(\alpha) \geq \varepsilon^*(\alpha), \quad \forall \alpha > 1 , \tag{64}$$

for every monotonic increasing potential in the version space. Note the intriguing result that the asymptotic behavior

$$\varepsilon^*(\alpha) \approx 1/\alpha, \quad \alpha \to \infty , \tag{65}$$

is identical to that obtained by using the annealed approximation for the Gibbs rule see, e.g., [5].

V. A CLASS OF COST FUNCTIONS WHICH FAVOR LARGE OVERLAPS WITHIN THE VERSION SPACE

We consider the following general class of cost functions, cf. Fig. 1:

$$V_s^+(\lambda) = \begin{cases} +\infty, & \lambda < 0 , \\ -\dfrac{\lambda^s}{s}, & \lambda > 0 , \end{cases} \tag{66}$$

with s real and $\neq 0$. For $s = 0$ we define $V_s^+(\lambda)$ as

$$V_{s=0}^+(\lambda) = \begin{cases} +\infty, & \lambda < 0 , \\ -\ln\lambda, & \lambda > 0 . \end{cases} \tag{67}$$

We start by pointing out that the parameter s may not be larger than 2 in order to avoid the divergence of the integral determining the value of the free energy, cf. Eq. (6). Furthermore, we show in Appendix A that the cost function associated with the above potential is convex $\forall s \leq 1$, hence the minimum is unique and can be found by gradient descent. At $s = 1$, the curvature of the potential switches sign with the result that, for the values $1 < s \leq 2$, one of the conditions in our proof is not met. In fact we will find in this case that the local stability of the replica symmetric solution is violated.

We now set out to find the minimum of the cost function for different values of the parameter s. The infinite value of the cost for $\lambda < 0$ implies that we are looking for a **J** vector inside the version space. The function $V_s^+(\lambda)$ is, for all values of s, a monotonic decreasing function of λ on the positive axis and thus belongs to the class of potentials for which the upper bound $\varepsilon^*(\alpha)$ given in Eq. (59) applies. The specific form of the potential however allows to extract detailed information about the generalization error. The derivative $dV_s^+/d\lambda$ is, $\forall s < 2$, given by

$$\frac{dV_s^+}{d\lambda} = -\lambda^{s-1} \quad (\lambda > 0) . \tag{68}$$

This leads to the simple-looking equation determining the function $\lambda_0(t,x)$ that minimizes (11):

$$\lambda - t = x\lambda^{s-1} \quad (\lambda \geq 0) . \tag{69}$$

It is not possible to solve this equation for $\lambda_0(t,x)$ for general s, but it is trivial to solve it for the inverse function:

$$t(\lambda_0,x) = \lambda_0 - x\lambda_0^{s-1}, \quad \lambda_0 \geq 0 . \tag{70}$$

To obtain $\lambda_0(t,x)$ from it, we must consider two separate cases. For $s < 1$, the function $t(\lambda_0,x)$ is a monotonously increasing function on the interval $\lambda_0 > 0$ ranging from $-\infty$ at $\lambda_0 = 0$ to $+\infty$ as $\lambda_0 \to +\infty$. In this case, it is convenient to use λ_0 as new integration variable in the saddle point equations Eqs. (14) and (15) and to calculate the integrals numerically. For $1 < s < 2$, however, $t(\lambda_0)$ decreases from the value 0 at $\lambda_0 = 0$ to a minimum $t_m < 0$ at a certain value λ_c and then increases steadily for larger values of λ. In this case, the function $\lambda_0(t,x) \equiv 0$ for $t < t_m$. At $t = t_m$, it makes a finite jump to the value λ_c and follows further the increasing branch of the inverse function $t(\lambda_0,x)$. The integrals in Eqs. (14) and (15) must now be split in a part $-\infty < t < t_m$ where $\lambda_0(t,x) \equiv 0$ and in a part $t_m < t < +\infty$ where λ_0 can again be used as new integration variable. In this way we never need the explicit solution of Eq. (69) for $\lambda_0(t,x)$. With regard to the stability of the replica symmetric solution we find that

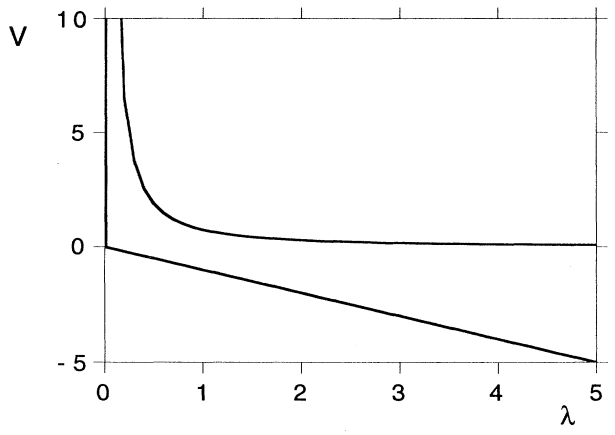

FIG. 1. Two representative examples of potentials $V_s^+(\lambda)$ (cf. Eqs. (66) and (67), which favor a large overlap with the training patterns, for $s=-1.35$ (upper curve) and $s=1$ (lower curve).

the stability condition (22) is satisfied for $s<1$. This was to be expected since we know that the minimum is unique and nondegenerate. The stability criterion is violated for $s>1$, as is immediately clear from the existence of a discontinuity of the function λ_0 at $t=t_m$.

Using the above procedure one can proceed to a numerical solution of Eqs. (14) and (15), and calculate the resulting generalization error for any value of s. As an example we plotted in Fig. 2 the resulting generalization error for $s=-1.35$, together with the results obtained by simulation using gradient descent for a system of size $N=50$. The agreement is quite satisfactory.

To get a more precise idea of how the generalization error depends on the parameter s we derive the exact asymptotic results for small α and large α values. This derivation is greatly simplified by the observation that the solution $\lambda_0(t,x)$ of Eq. (69) exhibits the following scaling behavior:

$$\lambda_0(t,x)=x^{1/(2-s)}\lambda_0\left[\frac{t}{x^{1/(2-s)}},1\right]. \tag{71}$$

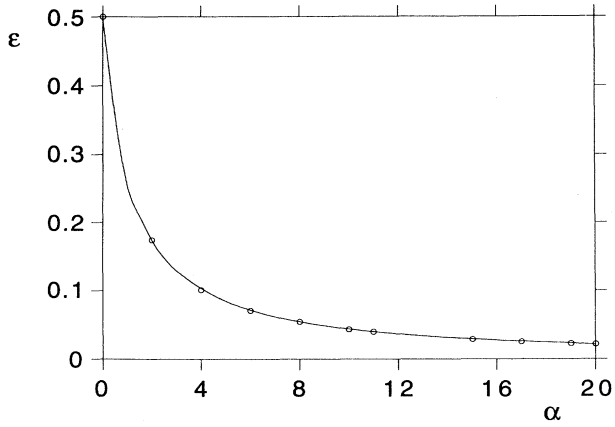

FIG. 2. The generalization error $\epsilon(\alpha)$ corresponding to a potential $V_s^+(\lambda)$ with $s=-1.35$ and results of a numerical simulation for $N=50$.

For α small, $R\to 0$ and $x\to\infty$ so that Eqs. (14) and (15) can be solved rather easily:

$$R=\left[\frac{2\alpha}{\pi}\right]^{1/2} \tag{72}$$

hence

$$\varepsilon(\alpha)\underset{\alpha\to 0}{\longrightarrow}\frac{1}{2}-\frac{\sqrt{2\alpha}}{\pi^{3/2}}. \tag{73}$$

These results are identical to the one obtained for the Hebb rule cf. Eq. (28). The calculation of the asymptotic behavior for $\alpha\to\infty$ is more involved. We briefly discuss some points of the calculation and refer to Appendix B for more details. In the limit $\alpha\to\infty$ we have that $R\to 1$ and $x\to 0$ so that Eqs. (14) and (15) reduce to two equations for the combination $A=\alpha\sqrt{1-R^2}/\pi$ and $B=x^{1/(2-s)}/\sqrt{1-R^2}$. The coefficient A is the interesting one because it is directly related to the asymptotic behavior of $\varepsilon(\alpha)$ by

$$\varepsilon(\alpha)\underset{\alpha\to\infty}{\longrightarrow}\frac{A}{\alpha}. \tag{74}$$

In solving the equations for A and B one must distinguish two cases. When $\frac{1}{2}\le s<2$ the equations can be solved analytically yielding $A=1$ and $B=0$. The asymptotic behavior thus saturates the upper bound, cf. Eq. (65). For $s<\frac{1}{2}$ on the other hand, it is found that $B\neq 0$ and the two equations can only be solved numerically. The value of A is represented in Fig. 3 as a function of s. As one moves from large to small s values, one observes that A first takes on a constant plateau value equal to 1 for $\frac{1}{2}\le s\le 2$, then decreases from the value 1 for $s=\frac{1}{2}$ to a minimum value of $A\approx 0.443$ for $s\approx -1.35$ after which it is again slightly increases and asymptotically approaches to the value $A\approx 0.50$ which is presumably identical to the asymptotic value for the perceptron with maximal stability cf. Eq. (41). The minimal value of 0.443 is surprisingly close to the result of the Bayes rule, cf. Eq. (48). In fact, the whole generalization curve for

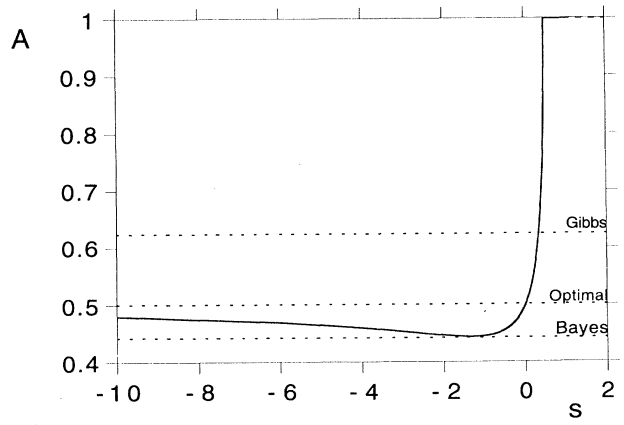

FIG. 3. The proportionality constant A, describing the asymptotic decay of the generalization error $\epsilon\sim A/\alpha$ [Eq. (74)], in function of the value of s for the case of a repulsive potential $V_s^+(\lambda)$. A broken line indicates that the replica symmetric solution is unstable.

$s \approx -1.35$, which is shown in Fig. 2, lies within 0.5% of the generalization error of the Bayes rule for all values of α. At this point it is pertinent to recall that the corresponding **J** vector is the unique nondegenerate minimum of the cost function and can thus be found straightforwardly by gradient descent techniques. For practical applications one may consider using an inverse power law potential $s = -1$, which is numerically less time consuming, and which still gives results within 1% of Bayes.

VI. A CLASS OF POTENTIALS WHICH FAVOR SMALL OVERLAPS WITHIN THE VERSION SPACE

Since we know that the version space is convex and that the best possible student that can be constructed on the basis of the training set is the center of mass, we infer that the students with larger generalization error can typically be found close to the boundaries of the version space. To verify this intuition we consider the following class of functions that favor small values of λ:

$$V_s^-(\lambda) = \begin{cases} +\infty, & \lambda < 0 , \\ +\dfrac{\lambda^s}{s}, & \lambda > 0 . \end{cases} \tag{75}$$

The parameter s now may take on positive values only. Negative values must be excluded as they would destroy the convergence of the integral over λ in (6). It is immediately clear that the lowest energy $E(\mathbf{J})$ will be 0 for all values of $\alpha < 1$. Indeed, for $\alpha < 1$, many different students will satisfy the conditions $\mathbf{J}\cdot\boldsymbol{\xi}^\mu = 0$ $(\mu = 1, \ldots, p)$ so that Eqs. (14) and (15), which are based on the assumption of a unique nondegenerate minimum, do not describe this case. We therefore limit ourselves here to $\alpha > 1$.

The function $V_s^-(\lambda)$ satisfies the conditions specified in Eq. (61), so that $\varepsilon^*(\alpha)$ gives a lower bound for the generalization error. More detailed information can be obtained as follows. The equation for $\lambda_0(t,x)$ now reads

$$\lambda - t = -x\lambda^{s-1} \quad (\lambda \geq 0) . \tag{76}$$

From its definition, $x \geq 0$. Since λ must be non-negative, it follows immediately that $\lambda_0(t,x) = 0$ for $t \leq 0$. Again it is easy to solve Eq. (76) for the inverse function

$$t(\lambda_0, x) = \lambda_0 + x\lambda_0^{s-1} \quad (\lambda_0 \geq 0) . \tag{77}$$

For $s > 1$, $t(\lambda_0, x)$ is zero at $\lambda_0 = 0$ and increases steadily with increasing λ_0. This defines the inverse function uniquely. For $0 < s < 1$ on the other hand, $t(\lambda_0, x)$ decreases from $+\infty$ at $\lambda_0 = 0$ to a minimum $t_m > 0$ at a certain value λ_c and then increases steadily for larger values of λ. In this case $\lambda_0(t,x) \equiv 0$ for $t < t_m$. At $t = t_m$ it makes a finite jump to the inverse function $t(\lambda_0, \alpha)$. Using these observations, the integrals in Eqs. (14) and (15) can be calculated by splitting the integration interval in a part $-\infty < t < t_m$ where $\lambda_0(t,x) = 0$ and a part $t_m < t < \infty$ where λ_0 can be used as an integration variable. The equations are then easily solved numerically for any value of α and s.

The most interesting point is the behavior of $\varepsilon(\alpha)$ for $\alpha \to \infty$. Here we proceed as in Appendix B by using the same scaling relation (71) as before. Again, for $s \geq \frac{1}{2}$, we obtain

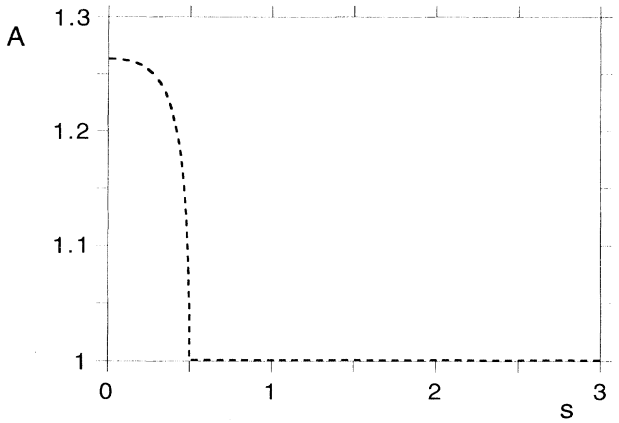

FIG. 4. Same result as in Fig. 3, but for the case of an attractive potential $V_s^-(\lambda)$.

$$\varepsilon(\alpha) \underset{\alpha\to\infty}{\sim} \frac{1}{\alpha} \tag{78}$$

so that we saturate the lower bound ε^*. For smaller values of s however, the asymptotic behavior is still proportional to $1/\alpha$, but the proportionality factor A is larger than 1 and reaches a maximum of approximately 1.28 in the limit $s \to 0$, cf. Fig. 4. This result should be compared with the one from the worst student of the version space, cf. Eq. (49).

We stress that the results of this section have been obtained assuming replica symmetry. It is however clear that replica symmetry must be broken for $0 < s < 1$, where the function $\lambda_0(t)$ has a discontinuity in function of t. Furthermore, a numerical evaluation of the integral appearing in (22) leads to the conclusion that replica symmetry is broken for all values of $s > 0$, except for the limiting values $\alpha \to 1$ and $\alpha \to \infty$, where the replica symmetric solution is marginally stable. A similar behavior of replica symmetry breaking between two limiting values of α was also observed for the worst case scenario [28]. In this case it was proven that replica symmetry and one-step replica symmetry breaking lead to an identical asymptotic behavior of the generalization error for $\alpha \to \infty$, which is therefore believed to be exact. For the same reason, we expect that the asymptotic results derived above are also exact.

VII. DISCUSSION

We would like to close with some thought-provoking comments about the form of the version space. As mentioned before, the Bayes result is reproduced by the **J** vector located at the center of mass of the version space. This center can be found very accurately by a cost function which penalizes heavily **J** vectors close to the boundaries of the version space. In this way we have discovered a simple gradient descent algorithm generating a unique minimum with generalization error almost identical to the Bayes one. Penalizing too much however leads to the location of the perceptron with optimal stability. This suggests that the form of the version space

deviates from one with inversion symmetry. Nevertheless, the fact that the center of mass of the version space yields the result of the Bayes rule implies that this space is cut in two pieces of exactly equal size by any large circle that passes through its center of mass. This is a property that we would normally associate to an object with inversion symmetry.

Both the teacher and the worst student lie at the boundary of the version space. It is therefore of some interest to study the effect of a cost function that favors **J** vectors close to this boundary. We found that for such cost functions the generalization error is larger with an asymptotic decay of at least $1/\alpha$. The fact that this result is worse than that of the typical student (Gibbs rule) agrees with the intuitive geometrical picture that the region of the boundary of the version space with high entropy (i.e., large "number" of students between R and $R+dR$ close to this boundary) will lie further away from the teacher than the region with high "bulk" entropy ("number" of students in the version space between R and $R+dR$). By maximally penalizing **J** vectors that are not on the boundary of the version space we obtain a generalization error that decreases asymptotically as $1.28/\alpha$, to be compared with the $1.5/\alpha$ behavior of the worst student.

As far as practical applications of the above results are concerned, one may wonder whether the difference between the optimal stability perceptron and the perceptron that one obtains as the minimum of a λ^s potential (with $s \simeq -1.35$) is significant. For random independent patterns, the difference in generalization error between both is small indeed. However, in real-life problems, where the patterns are not random, the version space will deviate significantly from an object with inversion symmetry, while the Bayes result is still reproduced by the center of mass [29]. We expect that in this case the optimal perceptron will perform much worse than the inverse-power law cost function. This question is currently under investigation.

ACKNOWLEDGMENTS

We thank the Program on Inter-University Attraction Poles, Prime Minister's Office, Belgian Government for financial support. One of the authors (C. V.d.B.) also acknowledges support from the NFWO Belgium.

APPENDIX A: THE COST FUNCTION ASSOCIATED WITH THE REPULSIVE POTENTIALS IS CONVEX ON THE SPHERE $J^2=N$ FOR $s \leq 1$

For $s \leq 1$ and $\lambda \geq 0$, the potential $V_s(\lambda) = -\lambda^s/s$ obeys the following two inequalities:

$$V[a\lambda_1+(1-a)\lambda_2] \leq aV(\lambda_1)+(1-a)V(\lambda_2) , \quad \forall a, \ 0 \leq a \leq 1 , \tag{A1}$$

$$V(\rho\lambda) \leq V(\lambda), \quad \forall \rho \geq 1 . \tag{A2}$$

Consider now any two vector $\mathbf{J}_1$ and $\mathbf{J}_2$ with λ_1 and $\lambda_2 \geq 0$ and $\mathbf{J}_1^2=\mathbf{J}_2^2=N$ and a vector $\mathbf{J}'=a\mathbf{J}_1+(1-a)\mathbf{J}_2, 0 \leq a \leq 1$, lying on the line that connects them. By applying inequality (A1) to every term of the sum defining $E(\mathbf{J}')$, one finds that

$$E(\mathbf{J}') \leq aE(\mathbf{J}_1)+(1-a)E(\mathbf{J}_2) . \tag{A3}$$

This proves the convexity of $E(\mathbf{J})$ within the sphere $\mathbf{J}^2=N$.

We now consider the vector **J** that is parallel and in the same direction of **J'**, but with the "proper" normalization $\mathbf{J}^2=N$. Clearly $\mathbf{J}=\rho\mathbf{J}'$ with $\rho \geq 1$, since **J'** lies inside the hypersphere ($\mathbf{J}' \leq N$) and Eq. (A2) immediately yields

$$E(\mathbf{J}) \leq E(\mathbf{J}') . \tag{A5}$$

Combining (A3) and (A4) then yields

$$E(\mathbf{J}) \leq aE(\mathbf{J}_1)+(1-a)E(\mathbf{J}_2) \tag{A5}$$

which proves the convexity of $E(\mathbf{J})$ on the surface of the sphere $J^2=N$.

APPENDIX B: ASYMPTOTIC FORM WHEN $\alpha \to \infty$ FOR $V_s^+(\lambda)$

Using the scaling relation (71) and introducing a new integration variable $u=t/\sqrt{1-R^2}$, Eq. (15) can be rewritten as

$$\frac{1}{2\alpha\sqrt{1-R^2}} = \left[\frac{x^{1/(2-s)}}{\sqrt{1-R^2}}\right]^2 \int_{-\infty}^{\infty} \frac{du}{\sqrt{2\pi}} e^{-1/2(1-R^2)u^2} H(-Ru) \left[\lambda_0\left[u\frac{\sqrt{1-R^2}}{x^{1/(2-s)}}, 1\right] - u\frac{\sqrt{1-R^2}}{x^{1/(2-s)}}\right]^2 . \tag{B1}$$

When $\alpha \to \infty$, $R \to 1$ and the Gaussian factor in the integrand tends to 1. Without the Gaussian, the convergence of the integral at the upper limit is no longer guaranteed in all cases. Indeed, from (69) one easily derives

$$\lim_{t\to\infty} \frac{\lambda_0(t,1)-t}{t^{s-1}} = \lim_{t\to\infty} \frac{\lambda_0^{s-1}}{t^{s-1}} = 1 . \tag{B2}$$

This means, when $R \to 1$, that the integral will diverge when $2(s-1) \geq -1$ or $s \geq \frac{1}{2}$ and converge when $s < \frac{1}{2}$.

(i) $\frac{1}{2} \leq s < 2$

Since the integral in (B1) diverges when $R \to 1$, it is necessary that $x^{1/(2-s)}/\sqrt{1-R^2}$ tends to zero in order to obtain a finite value for $\alpha\sqrt{1-R^2}$. We now look at the second equation Eq. (14) which we rewrite using (71) as

$$\left[\frac{\pi}{2}\right]^{1/2}\frac{R}{\alpha\sqrt{1-R^2}} = \frac{x^{1/(2-s)}}{\sqrt{1-R^2}}\int_{-\infty}^{\infty} Dt\,\lambda_0\left[t\frac{\sqrt{1-R^2}}{x^{1/(2-s)}},1\right]. \quad \text{(B3)}$$

When $R\to 1$ and $x^{1/(2-s)}/\sqrt{1-R^2}\to 0$ the RHS is undefined. We can, however, use (71) again and rewrite (B3) as

$$\left[\frac{\pi}{2}\right]^{1/2}\frac{R}{\alpha\sqrt{1-R^2}} = \int_{-\infty}^{\infty} Dt\,\lambda_0\left[t,\left[\frac{x^{1/(2-s)}}{\sqrt{1-R^2}}\right]^{2-s}\right].$$

The limit $R\to 1$ now yields immediately

$$\left[\frac{\pi}{2}\right]^{1/2}\frac{1}{\alpha\sqrt{1-R^2}} = \int_{-\infty}^{\infty} Dt\,\lambda_0(t,0) = \int_0^{\infty} Dt\,t = \frac{1}{\sqrt{2\pi}}$$

where we have used (69) for $x=0$ and $\lambda_0(t,x)\geq 0$ for all t.

(ii) $s<\frac{1}{2}$

Calling

$$A=\lim_{\alpha\to\infty}\frac{\alpha\sqrt{1-R^2}}{\pi},\quad B=\lim_{\alpha\to\infty}\frac{x^{1/(2-s)}}{\sqrt{1-R^2}},$$

we now immediately obtain two equations for A and B by taking the limit of (B1) and (B3):

$$\frac{1}{A}=\sqrt{2\pi}B^2\int_{-\infty}^{\infty} du\,H(-u)\left[\lambda_0\left[\frac{u}{B},1\right]-\frac{u}{B}\right]^2,$$

$$\frac{1}{A}=\sqrt{2\pi}B\int_{-\infty}^{\infty} Dt\,\lambda_0\left[\frac{t}{B},1\right],$$

which can be solved numerically for A and B.

[1] G. Gyorgyi and N. Tishby, in *Neural Networks and Spin Glasses,* edited by W. K. Theumann and R. Koberle (World Scientific, Singapore, 1990), pp. 3–36.

[2] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computing* (Addision-Wesley, Reading, Massachusetts, 1991).

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

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

[5] M. Opper and W. Kinzel, in *Physics of Neural Networks III,* edited by E. Domany, J. L. Van Hemmen, and K. Schulten (Springer, Berlin, 1995).

[6] C. Van den Broeck, Acta Phys. Pol. B **25**, 903 (1994).

[7] A. Engel, Int. J. Mod. Phys. (to be published).

[8] T. Grossman, R. Meir, and E. Domany, in *Neural Information Processing Systems 1,* edited by D. S. Touretzky (Morgan Kaufmann, San Mateo, CA, 1989), p. 73.

[9] P. Rujan, J. Phys. I France **3**, 277 (1993).

[10] M. Griniasty and H. Gutfreund, J. Phys. A **24**, 715 (1991).

[11] K. Y. M. Wong and D. Sherrington, J. Phys. A **23**, 4659 (1990).

[12] J. Anlauf and M. Biehl, Europhys. Lett. **10**, 587 (1989).

[13] M. Opper and D. Haussler, Phys. Rev. Lett. **66**, 2677 (1991).

[14] V. N. Vapnik, *Estimation of Dependences of Empirical Data* (Springer, Berlin, 1982).

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

[16] M. Anthony and N. Biggs, *Computational Learning Theory* (Cambridge University Press, Cambridge, England, 1992).

[17] M. Griniasty, Phys. Rev. E **47**, 4496 (1993).

[18] P. Majer, A. Engel, and A. Zippelius, J. Phys. A **26**, 7405 (1993).

[19] K. Y. M. Wong and D. Sherrington, Phys. Rev. E **47**, 4465 (1993).

[20] F. Gerl, Untersuchungen an Neuronalen Netzwerken mit der Cavity-Methode, Ph.D. Thesis, University of Regensburg, 1994 (unpublished).

[21] K. Y. M. Wong, Europhys. Lett. **30**, 245 (1995).

[22] J. R. L. de Almeida and D. J. Thouless, J. Phys. A **11**, 271 (1978).

[23] M. Bouten, J. Phys. A **27**, 6021 (1994).

[24] R. Meir and J. F. Fontanari, Phys. Rev. A **45**, 8874 (1992).

[25] F. Vallet and J. C. Cailton, Phys. Rev. A **41**, 3059 (1990).

[26] M. Opper, W. Kinzel, J. Kleinz, and R. Nehl, J. Phys. A **23**, L581 (1990).

[27] T. L. H. Watkin, Europhys. Lett. **21**, 871 (1993).

[28] A. Engel and C. Van den Broeck, Phys. Rev. Lett. **71**, 1772 (1993); Physica A **200**, 636 (1993).

[29] T. L. H. Watkin and J. P. Nadal, J. Phys. A **27**, 1899 (1994).