Statistical mechanics of learning with soft margin classifiers

Sebastian Risau-Gusman[1,2] and Mirta B. Gordon[1,*]
[1]*Département de Recherche Fondamentale sur la Matière Condensée CEA–Grenoble, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France*
[2]*Zentrum für Interdisziplinäre Forschung Wellenberg 1, D-33615 Bielefeld, Germany*
(Received 17 February 2001; published 29 August 2001)

We study the typical learning properties of the recently introduced soft margin classifiers (SMCs), learning realizable and unrealizable tasks, with the tools of statistical mechanics. We derive analytically the behavior of the learning curves in the regime of very large training sets. We obtain exponential and power laws for the decay of the generalization error towards the asymptotic value, depending on the task and on general characteristics of the distribution of stabilities of the patterns to be learned. The optimal learning curves of the SMCs, which give the minimal generalization error, are obtained by tuning the coefficient controlling the trade-off between the error and the regularization terms in the cost function. If the task is realizable by the SMC, the optimal performance is better than that of a hard margin support vector machine and is very close to that of a Bayesian classifier.

DOI: 10.1103/PhysRevE.64.031907 PACS number(s): 87.10.+e, 02.50.−r, 05.20.−y

I. INTRODUCTION

Neural networks are models of learning systems composed of interconnected units that, besides their biological relevance, have been shown to be very useful for classification tasks. The weights of the connections are adjusted through a process called *learning* using a set of M examples. It is assumed that these are labeled following an underlying rule, usually called *teacher*. The purpose of learning is not only to classify correctly the examples of the training set, but also to *generalize* correctly on new inputs. To this aim, the network has to infer the teacher's rule. The quality of this inference is measured through the generalization error ϵ_g, which is the probability of misclassification of a new, randomly selected, input pattern. As ϵ_g is not a quantity available for the training process, learning is usually performed through the minimization of a function of the training patterns. The tools of statistical mechanics allow us to study the properties of such learning systems, providing a deep understanding of their typical behavior [1–5]. In particular, it has been shown that the minimization of the training error, that is, the fraction of training patterns misclassified by the network, does not necessarily provide the best generalizer [6–8]. This is why other cost functions, based on geometrical properties such as the distance of the patterns to the discriminating surface, or on probabilistic error measures such as the likelihood, are used for training.

The simplest instance of a neural network, the perceptron, is a single binary unit whose output is the sign of the weighted sum of its inputs. It can only perform linear separations of the patterns. If the classification task requires more complex discriminating surfaces, these may be implemented using feedforward networks with a layer of hidden units whose number is *a priori* unknown. The cost functions used to tackle this problem usually have several minima, and determining the lowest one is one of the main difficulties in learning with multilayer neural networks. This is also a problem for the theoretical analysis, as the typical properties of such networks depend crucially on the structure of the minima in the weights' space.

Recently, a new learning scheme has been proposed, which strives to get rid of the problem raised by the multiple minima. The obtained classifiers are called *support vector machines* (SVMs) [9,10]. Instead of directly looking for a complicated discriminating surface in input space, the patterns are first mapped to a high-dimensional *feature space*, where the rule to be learned is (hopefully) linearly separable. If this is the case, a simple perceptron can be trained to find the separation in feature space. Denoting the weights by $\mathbf{w} \in \mathrm{Re}^N$, the perceptron's output to an input $\mathbf{x} \in \mathrm{Re}^N$ is given by $\sigma = \mathrm{sgn}(\mathbf{w}\cdot\mathbf{x} + b)$, where b is a bias and the dot represents the inner product in Re^N. Thus, the patterns belonging to different classes are separated by a hyperplane orthogonal to $\mathbf{w}$ at distance $|b|/\|\mathbf{w}\|$ from the origin, with $\|\mathbf{w}\| = \sqrt{\mathbf{w}\cdot\mathbf{w}}$. The SVM's solution is the *maximal stability perceptron* (MSP) [11] in feature space, also called maximal margin hyperplane. This is the hyperplane at maximal distance κ_{max} from the closest patterns in the training set. Two different formulations of this problem in terms of cost functions have been proposed in the literature. In the first one [11], the cost function counts not only the number of misclassified patterns, but also the number of correctly classified ones that lie at a distance smaller than κ from the separating hyperplane

$$E_{\mathrm{MSP}}(\kappa) = \sum_{\mu=1}^{M} \Theta(\kappa\|\mathbf{w}\| - h_\mu), \tag{1}$$

where Θ is the Heaviside function, and

$$h_\mu \equiv \tau_\mu(\mathbf{w}\cdot\mathbf{x}_\mu + b), \tag{2}$$

is called *aligned field* of the training pattern $\mathbf{x}_\mu$, $\tau_\mu \in \{-1,1\}$ being its class. If the M N-dimensional patterns are correctly classified, the aligned fields are all positive. The SVM solution has $\mathbf{w}$ and b corresponding to κ_{max}, the larg-

*Also with Centre National de la Recherche Scientifique.

est possible value of κ such that $E_{\mathrm{MSP}}(\kappa_{max})=0$. If the training set is not linearly separable, κ_{max} becomes negative. Notice that there are no constraints on the norm of $\mathbf{w}$, that can be freely chosen.

If the norm of the weight vector is chosen so that the aligned field of the closest pattern be 1, this leads to an equivalent formulation of the problem [9,10], in which the function to be minimized is

$$E_{\mathrm{SVM}}=\tfrac{1}{2}\mathbf{w}\cdot\mathbf{w}, \tag{3}$$

subject to the conditions

$$h_\mu \geqslant 1, \quad \mu=1,\ldots,M. \tag{4}$$

Clearly, the constraints (4) can only be satisfied if it is possible to classify correctly all the examples. In that case, there are no training patterns in a strip of width $1/\|\mathbf{w}\|$ on both sides of the hyperplane, meaning that in the error-free regime $1/\|\mathbf{w}\|\equiv\kappa_{max}$. An interesting property of the SVM solution is that the weight vector and the bias can be written as a linear combination of a subset of training patterns, the *support vectors*, having $h_\mu=1$.

The minimization of Eq. (1) with $\kappa=\kappa_{max}$ is equivalent to that of Eq. (3) with condition (4) *only if the training set is linearly separable*. If errors cannot be avoided, the equivalence breaks down, as in one hand Eq. (1) has either negative κ_{max}, or several minima if $\kappa_{max}\geqslant 0$ is imposed, and on the other hand the constraints (4) cannot be satisfied. This is why the second formulation has been generalized [10] through the introduction of a new set of variables $\zeta_\mu\geqslant 0$, called *slacks*, which are a measure of the ''amount of violation'' of the constraints. An increasing function of these is included in the cost function (3) and the hard margin conditions (4) are modified to allow some patterns to be closer to the hyperplane than $1/\|\mathbf{w}\|$. The new problem amounts to minimizing

$$E_{C,k}=\frac{1}{2}\mathbf{w}\cdot\mathbf{w}+C\sum_{\mu=1}^{M}\zeta_\mu^k, \tag{5}$$

subject to the following conditions for $\mu=1,\ldots,M$,

$$h_\mu \geqslant 1-\zeta_\mu, \tag{6a}$$

$$\zeta_\mu \geqslant 0. \tag{6b}$$

The coefficient C in Eq. (5) is a hyperparameter that allows to control the trade-off between the error term, defined by the slacks, and the regularization term, proportional to the squared weights. As will be shown in Sec. IV, it may be selected to optimize the generalization performance. The exponent k in Eq. (5) modulates the relative cost of errors, depending on their distance to the hyperplane. Patterns in a strip of width $1/\|\mathbf{w}\|$ at each side of the hyperplane, whether correctly or incorrectly classified, as well as those incorrectly classified outside of this strip, have $\zeta_\mu>0$. $1/\|\mathbf{w}\|$ is called *soft margin*, and the resulting classifier *soft margin* SVM or *soft margin classifier* (SMC).

As the cost (5) is a quadratic function for $k=1$ and $k=2$, and the domain of minimization defined by Eqs. (6a) and (6b) is convex, the minimum is *unique* [12]. This remarkable property makes the new formulation attractive for applications, as it allows to get rid of the multiple minima appearing in other learning schemes. Like in the hard margin formulation, the solution $\{\mathbf{w},b\}$ can be expressed as a linear combination of the *support vectors*, which now include the patterns with positive slacks. The corresponding coefficients may be obtained by solving the *dual* problem (see for example [13]) which, for $k=1$ or $k=2$ has a particularly simple expression [10]. Several efficient methods are known for solving this kind of problems, and this is one of the reasons why these classifiers are so widely used lately.

In this paper we study the typical properties of the SMCs obtained by solving equation (5) subject to the conditions (6a) and (6b), with the methods of statistical mechanics, using the replica approach. It has been shown [14,15] that the statistical properties of SVMs in high-dimensional feature spaces [16] can be well approximated by considering a simple perceptron learning anisotropically distributed patterns. The amount of anisotropy depends on the normalization of the mapping from the input to the feature space. In this paper we only consider the case of an isotropic pattern distribution, which corresponds to a non-normalized mapping.

The learning properties of a perceptron learning an isotropic input pattern distribution have been extensively studied [17], mainly for linearly separable, i.e., realizable, tasks. In this case the hypothesis of replica symmetry is generally correct, allowing for a full analytical statistical mechanics calculation. In particular, the behavior of the generalization error ϵ_g in the limit of very large $\alpha\equiv M/N$ has a universal power law decay $\epsilon_g\approx\alpha^{-\nu}$ with $\nu=1$. Its prefactor allows to characterize the convergence to perfect learning of different learning algorithms. If the rule to be inferred cannot be generalized without errors, the task is called *unrealizable*. In this case the replica symmetric solution, although generally unstable, is believed to provide a good approximation of some learning properties. However, in the case of a linearly separable rule learned with noisy training patterns, which is thus unrealizable, the replica symmetric approximation gives an exponent $\nu=1/2$ [2] whereas one step of replica symmetry breaking shows [18] that this exponent is modified to $\nu=2/3$. As this is but an approximation to the full replica symmetry breaking scheme [19] at zero temperature, it is not clear whether this exponent is correct. The same exponent has been found in the case of a quadratic hard margin SVM learning a linearly separable task, that is, a rule simpler than those implementable with the student's architecture [16]. Another case of interest is that of *inconsistent learning* [6], which refers to realizable tasks learned with algorithms that do not strive to minimize the number of training errors. In this case, the exponent within the replica symmetric approximation was found to be $\nu=1/2$ [6].

As the soft margin problem has a unique minimum for $k=1$ and $k=2$, even if the task is unrealizable, the replica symmetry hypothesis should be always correct, providing a framework for the study of complex classification tasks even when the mismatch between the student and the teacher hinders error-free learning.

In this paper we present the statistical properties of SMCs learning several kinds of realizable and unrealizable rules. The model and the statistical mechanics approach are presented in Sec. II. The theoretical properties of SMCs with exponents $k=1$ and $k=2$ in the cost function (5) are obtained as a function of the training set size $\alpha \equiv M/N$ in the thermodynamic limit $N,M \to \infty$. Several teacher rules are considered in Sec. III. One of our most striking results is that the generalization error for large α exhibits a very rich variety of asymptotic behaviors, depending on the type of rule to be inferred. In particular, even if the task is realizable, the soft margin algorithm is inconsistent unless $C \to \infty$. For finite C, we find that the fraction of training errors at finite α is finite, and the generalization error vanishes asymptotically with α following a $\nu=2/3$ power law. In the unrealizable tasks considered, ϵ_g converges to an asymptotic finite value either exponentially or with a power law with $\nu=1/2$. The usual exponent $\nu=1$ only arises for error-free learning of a realizable task. In Sec. IV we derive the best generalization performances of SMCs through the determination of the value $C_{opt}(\alpha)$ that minimizes the generalization error. Finally we present a summary of our results in Sec. V, together with some open questions. Most details of the proofs are left to the Appendix.

II. STATISTICAL MECHANICS APPROACH

We consider a student perceptron of weight vector $\mathbf{w}=(w_1,\ldots,w_N)$, *without threshold*. That is, we set $b=0$ in Eq. (6a). Given any N-dimensional input vector $\mathbf{x}$, the classifier's output is $\sigma=\mathrm{sgn}(\mathbf{w}\cdot\mathbf{x})$: all the points lying on the same side of a hyperplane orthogonal to $\mathbf{w}$ containing the origin are given the same class. We assume that the perceptron learns the classification with the soft margin algorithm, using a set $\mathcal{L}_M=\{(\mathbf{x}_\mu,\tau_\mu)\}_{\mu=1,\ldots,M}$ of M examples or training patterns. These consist of input vectors $\mathbf{x}_\mu$ drawn from an isotropic Gaussian distribution,

$$P(\mathbf{x})=\frac{e^{-N\mathbf{x}^2/2}}{(2\pi/N)^{N/2}}, \tag{7}$$

and labels $\tau_\mu \in \{-1,1\}$ that represent the corresponding classes. The classification tasks considered in this paper are given by the following teacher's rule:

$$\tau=\mathrm{sgn}[\mathcal{P}(\mathbf{w}_0\cdot\mathbf{x})], \tag{8}$$

where $\mathbf{w}_0$ is referred to as the teacher's vector hereafter, and $\mathcal{P}(z)$ is a polynomial of z. Each of its zeros z_i [20] defines a discriminating hyperplane at a distance $|z_i|/\|\mathbf{w}_0\|$ from the origin. Rules of the kind (8) partition the input space into as many different regions as the number of zeros of the polynomial plus one, separated by parallel hyperplanes normal to the teacher's vector $\mathbf{w}_0$. Patterns in successive regions belong alternatively to class $+1$ or -1. As only the zeros of the function $\mathcal{P}(z)$ matter, there is no loss of generality in our assumption that $\mathcal{P}(z)$ is a polynomial. We assume $\|\mathbf{w}_0\|=\sqrt{N}$, which is equivalent to imposing the unit of distance. Notice that the only rule realizable for the student perceptron considered in this paper is that of the linear teacher $\mathcal{P}(z)=z$.

In the following we study the properties of the solution to the soft margin problem using the by now standard tools of statistical mechanics [1,2]. That is, we assume that the ensemble of classifiers follows a Gibbs distribution defined by the energy function (5), at a fictitious temperature $1/\beta$, and we take the zero temperature limit. The constraints (6a) and (6b) play the role of infinite potential walls. Notice that the phase space in the present case has dimension $N+M$, as not only the weights $\mathbf{w}$ but also the slacks $\{\zeta_\mu\}_{\mu=1,\ldots,M}$, have to be learned. The partition function is

$$Z_{C,k}(\beta;\mathcal{L}_M,\mathcal{P})=\int \exp[-\beta E_{C,k}(\mathbf{w},\{\zeta_\mu\})] \times \prod_{\mu=1}^{M} \Theta(\tau_\mu \mathbf{w}\cdot\mathbf{x}_\mu - (1-\zeta_\mu))\Theta(\zeta_\mu)\, d\mathbf{w}\, d\zeta_\mu. \tag{9}$$

The inverse temperature β has obviously no physical meaning whatsoever; it is only introduced in order to study the properties of the SMC which, being the single minimum of the energy function, is selected in the limit $\beta \to \infty$. We assume that the number of training examples scales with the input space dimension, $M=\alpha N$, and take the thermodynamic limit $N \to \infty$, $M \to \infty$ with $\alpha \equiv M/N$ constant. The free energy per input space dimension averaged over all the possible training sets of M patterns, $f_{C,k}(\beta;\mathcal{P})$, is calculated with the replica method, that uses the identity

$$f_{C,k}(\beta;\mathcal{P})=-\lim_{N\to\infty}\frac{1}{N\beta}\overline{\ln Z_{C,k}(\beta;\mathcal{L}_M,\mathcal{P})} =-\lim_{N\to\infty}\frac{1}{N\beta}\lim_{n\to 0}\frac{\ln\overline{Z^n_{C,k}(\beta;\mathcal{L}_M,\mathcal{P})}}{n}, \tag{10}$$

where the overline represents the average over the pattern distribution (7), with labels given by Eq. (8). Z^n is the partition function of n independent replicas of the problem, that become coupled after taking the average. The typical properties of the classifier are obtained by taking the limit $\beta \to \infty$. The free energy (10) turns out to be a function of the following order parameters:

$$Q_a=\frac{\overline{\langle \mathbf{w}_a\cdot\mathbf{w}_a\rangle}}{N}, \tag{11a}$$

$$q_{ab}=\frac{\overline{\langle \mathbf{w}_a\cdot\mathbf{w}_b\rangle}}{N}, \tag{11b}$$

$$\tilde{R}_a=\frac{\overline{\langle \mathbf{w}_a\cdot\mathbf{w}_0\rangle}}{N}, \tag{11c}$$

where the brackets represent the phase space average and a and b are replica indices. The norm of the perceptron's weight vector Q_a is one of the order parameters because in the soft margin problem the weights are not normalized as

usually. q_{ab} is the overlap between two different weight vectors at temperature β^{-1}, and $\tilde{R}_a$ is the overlap of the perceptron's solution and the teacher's vector.

As for $k=1$ and $k=2$ the energy in Eq. (5) is a quadratic function in a convex domain, it has a single minimum [21], irrespective of the kind of rule that is being learned. Therefore, we may safely assume that all the replicas are equivalent, even in the case of learning unrealizable rules. We obtain thus the typical properties for cases where, using other more usual cost functions like the number of training errors, full replica symmetry breaking would be required [19]. The excellent agreement of the theoretical predictions and the numerical simulations presented in the following section is a further justification of our hypothesis of *replica symmetry*. Thus, we set $Q_a \equiv Q$, $q_{ab} \equiv q$, and $\tilde{R}_a \equiv \tilde{R}$, and we define the normalized overlap $R \equiv \tilde{R}/\sqrt{Q}$, that only depends on the angle between $\mathbf{w}$ and $\mathbf{w}_0$.

Due to the unicity of the soft margin solution, only one point in phase space has nonvanishing probability in the limit $\beta \to \infty$, so that $q \to Q$. It is convenient to introduce a new parameter, $x \equiv \beta(Q-q)$, which reflects how fast the fluctuations around the minimum of Eq. (5) vanish as $\beta \to \infty$. In this limit we obtain the typical free energy of the SMC learning a rule defined by the polynomial $\mathcal{P}$,

$$f_{C,k}(\mathcal{P}) = -\,\mathrm{extr}_{\{Q,R,x\}}\,[G_0(Q,R,x) - \alpha G_{C,k}(Q,R,x;\mathcal{P})], \tag{12}$$

where

$$G_0(Q,R,x) = \frac{Q}{2x}(1-R^2-x), \tag{13}$$

is an entropic term. The dependence on the rule to be learned is embodied in the second term of Eq. (12) through $\mathcal{P}(z)$, and on the learning algorithm through k and C. Integrating out the slack variables in the limit $\beta \to \infty$ using the saddle point method, we get

$$G_{C,k}(Q,R,x;\mathcal{P}) = \int_{-\infty}^{\infty} Dy \int_{\phi(y;Q,R,\mathcal{P})}^{\infty} Dt \times \min_{\zeta} W(\zeta;y,t,Q,R,x,\mathcal{P}), \tag{14}$$

where $Dt \equiv dt\,\exp(-t^2/2)/\sqrt{2\pi}$,

$$\phi(y;Q,R,\mathcal{P}) = \frac{yR\ \ \mathrm{sgn}[\mathcal{P}(y)] - 1/\sqrt{Q}}{\sqrt{1-R^2}}, \tag{15}$$

and

$$W(\zeta;y,t,Q,R,x,\mathcal{P}) = C\,\zeta^k + \frac{\{\zeta - \sqrt{Q(1-R^2)}\,[t-\phi(y;Q,R,\mathcal{P})]\}^2}{2x}. \tag{16}$$

In (14), according to the saddle point method, $W(\zeta;y,t,Q,R,x,\mathcal{P})$ has to be taken at its minimum $\zeta(t,y) \in [0, \sqrt{Q(1-R^2}]\,\phi(y;Q,R,\mathcal{P})$ for each couple (y,t). It is easy to see that there is a unique local minimum inside this interval for $k>1$. For $k=1$, W is a quadratic function of ζ, whose global minimum falls inside the allowed interval only for a finite range of values of t. Outside this range, the minimum lies at the boundary $\zeta=0$. As a consequence, for $k=1$ the inner integral in $G_{C,k}$ splits into two parts. The results for $k=1$ and $k=2$ are, respectively,

$$G_{C,1}(Q,R,x;\mathcal{P}) = \int_{(-1)/\sqrt{Q}}^{(xC-1)/\sqrt{Q}} Dt\,\frac{(t\sqrt{Q}+1)^2}{2x}\,g(t;R,\mathcal{P}) + \int_{(xC-1)/\sqrt{Q}}^{\infty} Dt\,C\left(t\sqrt{Q}+1-\frac{xC}{2}\right) \times g(t;R,\mathcal{P}), \tag{17}$$

$$G_{C,2}(Q,R,x;\mathcal{P}) = \int_{-1/\sqrt{Q}}^{\infty} Dt \frac{C(t\sqrt{Q}+1)^2}{1+2xC}\,g(t;R,\mathcal{P}), \tag{18}$$

with

$$g(t;R,\mathcal{P}) = \int \frac{dy}{\sqrt{2\pi(1-R^2)}} \exp\left(-\frac{(y\ \ \mathrm{sgn}(P(y)) + tR)^2}{2(1-R^2)}\right). \tag{19}$$

Deriving the free energy (12) with respect to Q, R, and x gives three coupled equations for the order parameters. These in turn determine the properties of the SMC. The explicit expression of the saddle point equations for $k=1$ and $k=2$ is left to the Appendix, where we also derive some general properties of the learning curves described in the next sections.

The generalization error ϵ_g, which is the probability of misclassification of any pattern drawn with probability (7), is a geometric property that depends only on R and the rule to be learned. In the case of rules of type (8), it is straightforward to obtain

$$\epsilon_g = \int Dt\,H\left(\frac{tR\ \ \mathrm{sgn}[\mathcal{P}(t)]}{\sqrt{1-R^2}}\right), \tag{20}$$

where $H(x) = \int_x^{\infty} Dt$. In the particular case of a linearly separable rule $\mathcal{P}(z)=z$, Eq. (20) reduces to the usual expression $\epsilon_g = \arccos(R)/\pi$.

The distribution of stabilities $\gamma_\mu \equiv h_\mu/\|\mathbf{w}\|$ of the training patterns $\rho(\gamma)$ is given by

$$\rho(\gamma) = \delta\left(\gamma - \frac{1}{\sqrt{Q}}\right) \int_{-1/\sqrt{Q}}^{(-1+xC)/\sqrt{Q}} Dt\,g(t;R,\mathcal{P}) + \Theta\left(\gamma - \frac{1}{\sqrt{Q}}\right) \frac{e^{-\gamma^2/2}}{\sqrt{2\pi}}\,g(-\gamma;R,\mathcal{P}) + \Theta\left(\frac{1}{\sqrt{Q}} - \gamma\right) \frac{\exp[-(\gamma - xC/\sqrt{Q})^2/2]}{\sqrt{2\pi}} \times g\left(-\gamma + \frac{xC}{\sqrt{Q}}; R, \mathcal{P}\right) \tag{21}$$

for the case $k=1$, where $\delta(y)$ is the Dirac delta, and

$$\rho(\gamma)=\Theta\left(\frac{1}{\sqrt{Q}}-\gamma\right)\frac{\left\{\exp-\frac{1}{2}[\gamma(1+2xC)-2xC/\sqrt{Q}]^2\right\}}{\sqrt{2\pi}}$$
$$\times g\left(-\gamma(1+2xC)+\frac{2xC}{\sqrt{Q}};R,\mathcal{P}\right)(1+2xC)$$
$$+\Theta\left(\gamma-\frac{1}{\sqrt{Q}}\right)\frac{e^{-\gamma^2/2}}{\sqrt{2\pi}}g(-\gamma;R,\mathcal{P}) \tag{22}$$

for the case $k=2$. The Dirac delta present in Eq. (21) implies that in the thermodynamic limit there is a macroscopic fraction of the examples that are at a distance of exactly $1/\sqrt{Q}$ from the hyperplane of the student, for the case $k=1$, which is not the case if $k=2$. This is a consequence of the different structure of the support vectors in both cases, which can be obtained by analyzing the dual problem: if $k=1$, all the vectors at the distance $1/\sqrt{Q}$ are support vectors, whereas for $k=2$ they are not.

The training error ϵ_t is the average fraction of incorrectly classified training patterns. Integrating Eqs. (21) and (22) over the negative stabilities we obtain,

$$\epsilon_t=\int Dt\,H\left(\frac{tR\ \ \mathrm{sgn}[\mathcal{P}(t)]+kxC/\sqrt{Q}}{\sqrt{1-R^2}}\right). \tag{23}$$

As expected, the training error is always strictly smaller than the generalization error. Both converge to the same limit for $\alpha\to\infty$.

III. LEARNING CURVES

In this section we present the learning curves, namely, the training error $\epsilon_t(\alpha)$ and the generalization error $\epsilon_g(\alpha)$ of the SMCs for different teacher rules. The results of computer simulations drawn on the same figures have been obtained by solving numerically the dual problem [13] using the Quadratic Optimizer for Pattern Recognition program [22], that we adapted to the case without threshold treated in this paper. The average has been taken over as many training sets as necessary (typically $\sim$500 for small α and $\sim$50 for big α) to ensure that the error bars are smaller than the symbols. These simulations are in excellent agreement with the theoretical predictions.

A. The linear rule

Introducing the expression $\mathcal{P}(z)=z$ corresponding to a linearly separable teacher's rule in Eq. (19), we obtain

$$g(t;R,\mathcal{P})=2\,H\left(\frac{Rt}{\sqrt{1-R^2}}\right). \tag{24}$$

The training and generalization errors, obtained after solving the extremum equations for different values of the hyperparameter C, are plotted against α on Figs. 1 and 2 for

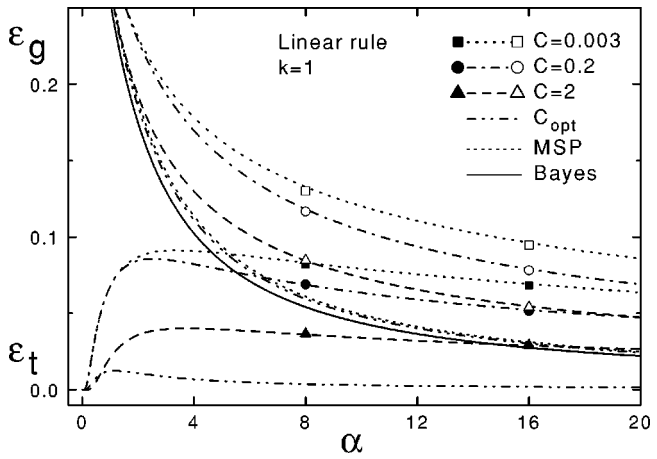

FIG. 1. Linearly separable rule. SMC's learning curves (ϵ_t below, ϵ_g above) corresponding to an exponent $k=1$ in the cost function, for different values of the hyperparameter C. The generalization errors of the MSP and the optimal (Bayesian) generalizer, are included for comparison. The learning curves of the optimal SMC, discussed in Sec. IV, are also represented. Symbols, ϵ_t in black, ϵ_g in white, correspond to results of computer simulations with $N=100$. Error bars are smaller than the symbols.

$k=1$ and $k=2$, respectively. The generalization error of the hard margin classifier, solution of Eq. (3) with conditions (4), and that of the optimal Bayesian generalizer [23], both of which are error-free solutions, are included in the figures for comparison. Despite the fact that the task is realizable by the student perceptron, the training error for finite C is positive. It goes through a maximum and vanishes asymptotically in the limit $\alpha\to\infty$. As expected, both for $k=1$ and $k=2$ at any α, ϵ_t is larger the smaller the value of C, which controls the relative importance of the error term in the cost function (5). We can also see from the figures that, given C, the machine with $k=2$ performs better than the one with $k=1$. This can be understood from the fact that, according to Eq. (6a) the examples that are errors have $\zeta_\mu>1$ and those that are not errors satisfy $0\leqslant\zeta_\mu<1$. Thus, the second term in Eq. (5), which is proportional to ζ_μ^k, penalizes the errors more heavily in the case $k=2$, forcing the machine to classify better than in the case $k=1$.

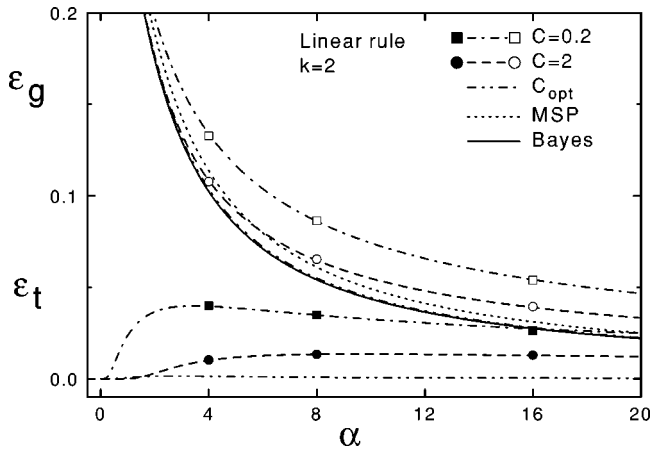

FIG. 2. Linearly separable rule. Same as the preceding figure, with an exponent $k=2$ in the cost function.

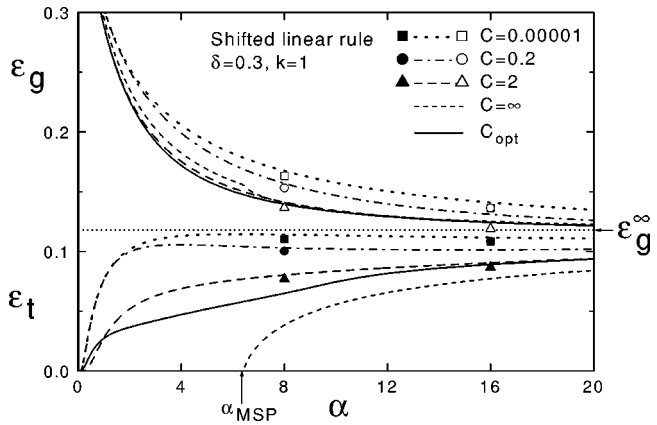

FIG. 3. Shifted linear rule. SMC's learning curves corresponding to an exponent $k=1$ in the cost function, for different values of the hyperparameter C. Symbols correspond to results of computer simulations with $N=50$. Error bars are smaller than the symbols. Asymptotically, $\epsilon_g^\infty=0.1179$.

On increasing C, the learning curves approach those of the MSP. In fact, by taking the limit $C\to\infty$ in our saddle point equations we get exactly the equations of the MSP for every value of α, independently of the power k. This is not surprising, as in this limit the error term dominates completely the soft margin cost function (5), which can only be minimized if all the slack variables, and consequently the training error, vanish. This is possible because the rule is realizable. It is well known that the generalization error of the MSP is larger than that of the Bayesian generalizer even asymptotically, as for $\alpha\to\infty$ both algorithms have $\epsilon_g\sim a/\alpha$, but $a=0.5005$ in the case of the MSP [24], whereas $a=0.442$ for the Bayesian perceptron [23].

The obtained behavior of the learning curves at finite C is reminiscent of that arising with other learning algorithms having a hyperparameter. In the inconsistent algorithms studied by Meir and Fontanari [6], patterns closer to the hyperplane than a finite imposed distance $\kappa>\kappa_{max}$ contribute to the cost, linearly in the case of the perceptron algorithm and quadratically in the case of the relaxation one. In the algorithm Minimerror [24] the hyperparameter is equivalent to a learning temperature. By training with these algorithms, as well as with the SMC studied here, the generalization error can be made smaller than that of the MSP by choosing appropriate values for the hyperparameters, at the price of learning with errors. The reason is that, in contrast with the MSP, the Bayesian solution presents a finite fraction of training patterns at any distance of the hyperplane [8]. Thus, solutions with a small controlled fraction of training errors may be closer to the optimal bayesian hyperplane than the MSP, which has no patterns at distances smaller than κ_{max}.

Unlike the generalization error of the inconsistent learning algorithms, that vanishes asymptotically like $\epsilon_g\sim 1/\sqrt{\alpha}$ [Ref. [6]], SMCs with finite C present a faster power law decay:

$$\epsilon_g\simeq\frac{\epsilon_0}{C^{1/6}\alpha^{2/3}}, \tag{25}$$

where the constant ϵ_0, is independent of C and is larger for $k=1$ than for $k=2$. In the limit $C\to\infty$ Eq. (25) no longer holds, and the well-known decay $\epsilon_g\approx\alpha^{-1}$ characteristic of error-free trained perceptrons learning realizable tasks is recovered.

Independently of the value of C, both the regularization term, proportional to Q, and the slacks term diverge like $\sim\alpha^{2/3}$ for $\alpha\to\infty$. In fact, this divergence arises because we divided the free energy in Eq. (10) by N, instead of dividing by $N(1+\alpha)$, which gives the energy per degree of freedom. In the large α limit, this converges to 0 as it should, like $\alpha^{-1/3}$. The separable case is the only one where the error term in the cost function presents the same asymptotic behavior as the regularization term. In this limit, the soft margin $1/\sqrt{Q}$ vanishes like $\alpha^{-1/3}$, in contrast with the hard margin behavior, $\kappa_{max}\approx\alpha^{-1}$ [Ref. [24]].

B. The shifted linear rule

Next we analyze the case of a linear teacher with a bias $\delta>0$. The corresponding polynomial has a single root: $\mathcal{P}(z)=z-\delta$. This teacher separates linearly the examples with a hyperplane at a distance $\delta/\sqrt{N}$ from the origin. As the student perceptron has no bias ($b=0$), zero generalization error cannot be achieved: this rule is unrealizable. The lowest value of ϵ_g, obtained by taking the asymptotic limit $R\to 1$ in Eq. (20), is $\epsilon_g^\infty=0.5-H(\delta)$.

The function g defined by Eq. (19) is

$$g(t;R,\mathcal{P})=H\left(\frac{Rt+\delta}{\sqrt{1-R^2}}\right)+H\left(\frac{Rt-\delta}{\sqrt{1-R^2}}\right). \tag{26}$$

Learning curves for different values of C are represented as a function of α in Fig. 3, for the particular value $\delta=0.3$.

If we take the limit $C\to\infty$ in our equations, we get those corresponding to the MSP only for $\alpha<\alpha_{MSP}$. At α_{MSP}, the training error of the SMC starts increasing (discontinuously if $k=2$) and the generalization error curve detaches down from that of the MSP, both through a second order phase transition. The learning curves obtained in the limit $C\to\infty$ are different for $k=1$ and $k=2$, in contrast with the realizable rule considered before, in which they converge to that of the MSP irrespective of the value of k. The same features are present for all the unrealizable rules studied in this paper. For $\alpha>\alpha_{MSP}$ the exact curve for the MSP is unknown, as in this region the symmetry of the replicas is broken.

For the shifted linear rules, α_{MSP} is a decreasing function of δ. It diverges at $\delta=0$, as the problem becomes separable, and tends to the perceptron's capacity $\alpha_c=2$ in the infinite δ limit. α_{MSP} cannot be smaller than α_c since in the thermodynamic limit any training set can be learned without errors for $\alpha<\alpha_c$ [Ref. [25]].

For finite values of C the transition at α_{MSP} becomes a crossover both for ϵ_t and ϵ_g, at values of $\alpha<\alpha_{MSP}$ that decrease on decreasing C. The training error for all α is larger than that for infinite C, both for $k=1$ and $k=2$. The generalization errors for different values of C cross each other as a function of α. The envelope of the curves $\epsilon_g(\alpha)$

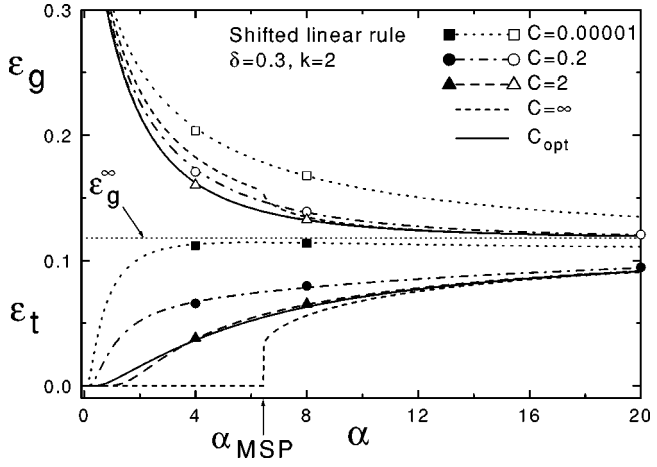

FIG. 4. Shifted linear rule. Same as the preceding figure, with an exponent $k=2$ in the cost function. Simulation results correspond to $N=100$.

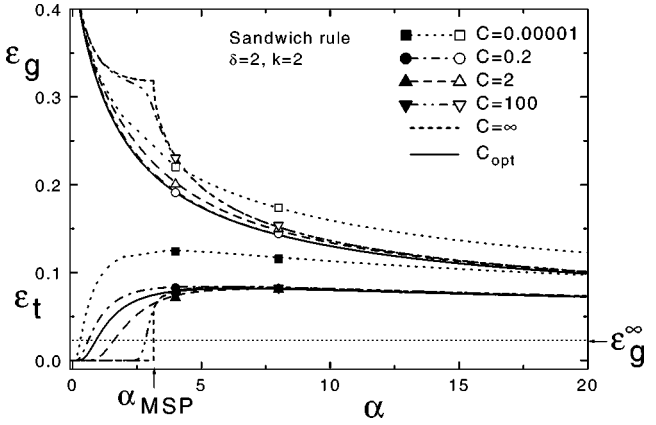

FIG. 5. Sandwich rule. SMC's learning curves corresponding to an exponent $k=2$ in the cost function, for different values of the hyperparameter C. Symbols correspond to results of computer simulations with $N=100$. Asymptotically, $\epsilon_g^\infty=0.023$.

corresponds to the lowest possible value of ϵ_g reachable by the corresponding SMCs. It depends on the exponent k.

The convergence of the generalization error to its asymptotic limit, for all values of C, is exponentially fast with α:

$$\epsilon_g-\epsilon_g^\infty \simeq \exp\left(-\frac{\alpha}{a_k}\right) \tag{27}$$

The decay constant a_k does not depend on C. A stronger exponential drop of the generalization error, with α^2 in the exponent, has been found [16] for SVMs learning "easy" teacher rules. These not only are realizable, but present a gap in the patterns distribution close to the discriminating surface. In contrast, here the student's hyperplane is surrounded by unlearnable patterns. The student cannot get rid of the errors by decreasing the soft margin, like with the linear rule. On increasing α, Q converges to a constant that depends on k and δ while the error term in Eq. (5) increases with α. For large enough α, the cost function is mainly dominated by the error term, and then C only plays the role of an irrelevant multiplicative constant. This is why the convergence rate to the asymptotic value of the generalization error does not depend on C.

Similar results are obtained for $k=2$, as is shown in Fig. 4.

C. Sandwich rule

Consider now rules of the form $\mathcal{P}(z)=z(z-\delta)$, where the polynomial defining the teacher's output has two roots. The corresponding discriminating surfaces are two parallel hyperplanes, one containing the origin and the other at a distance $\delta/\sqrt{N}$ of it. The patterns lying between the hyperplanes belong to class $+1$, the others to class -1. Thus, not only these are unrealizable rules, but the classification errors will necessarily correspond to patterns at a large distance of the student's hyperplane.

Here α_{MSP} is an increasing function of δ, starting at $\alpha_{MSP}=2$ for $\delta=0$, which corresponds to the most difficult learning task and diverging for $\delta\to\infty$.

The properties of the SMC are obtained by replacing

$$g(t;R,\mathcal{P})=2H\left(\frac{Rt}{\sqrt{1-R^2}}\right)+H\left(\frac{\delta-Rt}{\sqrt{1-R^2}}\right)-H\left(\frac{Rt+\delta}{\sqrt{1-R^2}}\right) \tag{28}$$

in the saddle point equations (A1)–(A3) of the Appendix.

The learning curves for $k=2$, for different values of the hyperparameter C, corresponding to a width $\delta=2$, are represented in Fig. 5. Even though the corresponding curves for $k=1$ are qualitatively similar, for large enough α the training error curves $\epsilon_t(\alpha)$ for $k=1$ are below those for $k=2$, given C. This is so because the unavoidable errors, which are very far from the hyperplane, are more heavily penalized if $k=2$. Thus, the SMC tries to learn these examples even if this increases the overall number of errors. As a result, learnable patterns close to the hyperplane, that have small slacks, are incorrectly classified. This can be checked up by taking a look at the distribution of stabilities, Fig. 6.

Like with the previous shifted linear rule, the norm of the student's weight vector Q tends to a constant value and therefore, the error term dominates the cost function in the asymptotic limit $\alpha\to\infty$. However, instead of the exponential convergence, the generalization error decays asymptotically to $\epsilon_g^\infty=H(\delta)$ like $\alpha^{-1/2}$. The reason for this difference is discussed in Sec. V.

D. The reversed wedge

Teachers defined by third order polynomials like $\mathcal{P}(z)=z(z-\delta)(z+\delta)$ with $\delta>0$, correspond to the so called reversed wedge [26] rules. Patterns $\mathbf{x}_\mu$ with $\mathbf{w}_0\cdot\mathbf{x}_\mu\in(-\infty,-\delta)\cup(0,\delta)$ belong to class -1, those outside this subspace to class $+1$. The generalization properties of a perceptron learning a reversed wedge teacher have been addressed in [26], and within the on-line paradigm, using Hebb's learning rule in [27].

In this case, α_{MSP} diverges both in the limits of vanishing and infinite wedge width δ, for which the problem becomes

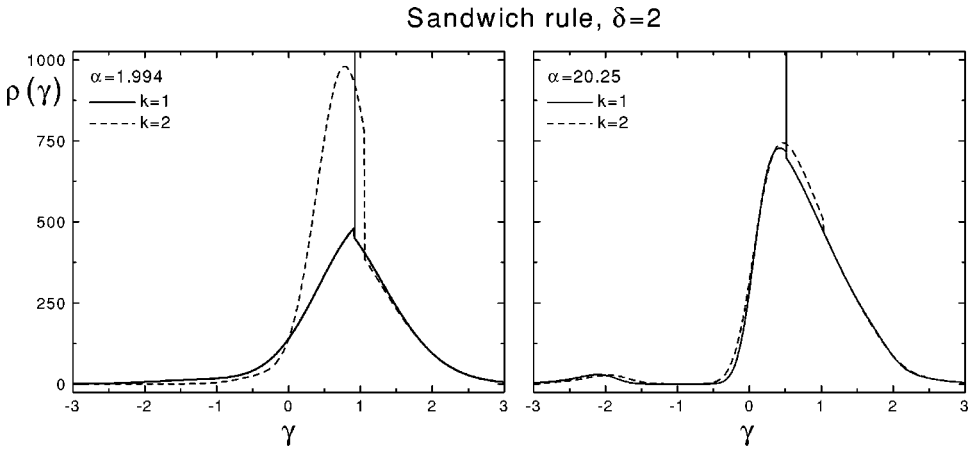

FIG. 6. Sandwich rule. Distribution of stabilities of the SMC for two different training set sizes α, obtained with $C=2$ in the cost function. The vertical lines give the position of the deltas, in the case $k=1$. The deltas contain 32.4 and 4.6% of the training patterns, for $\alpha=1.99$ and $\alpha=20.25$, respectively.

separable, and has a minimum at $\delta_c=\sqrt{2\ln 2}$ [Ref. [26]]. At this value of δ the patterns' stability distribution along the teacher's weight $\mathbf{w}_0$ has zero mean. Thus, for δ_c, $R=0$ for every value of α, and $\alpha_{MSP}=2$, equal to the perceptron's capacity.

The properties of the SMCs are deduced after insertion of

$$g(t;R,\mathcal{P})=2H\left(\frac{Rt-\delta}{\sqrt{1-R^2}}\right)+H\left(\frac{Rt+\delta}{\sqrt{1-R^2}}\right)-H\left(\frac{Rt}{\sqrt{1-R^2}}\right) \tag{29}$$

into the saddle point equations.

In contrast with the problems considered before, the generalization error of a perceptron learning the reversed wedge rule is a monotonic function of R *only if* $\delta=0$ or $\delta>\delta_c$ [Ref. [27]]. The different behaviors of ϵ_g are represented in Fig. 7.

For the values of k investigated, R has two distinct behaviors as a function of α, depending on the wedge's width δ. If $\delta<\delta_c$, the teacher's average stability is positive, and $R(\alpha)$ is a monotonic continuous function growing from 0 to its asymptotic value $+1$. In this range of small wedges, the soft margin learning algorithm does not converge to the minimal value of the generalization error in the limit of infinite α, as is the case in the other tasks considered before. In fact it "overshoots," in the sense that $R(\alpha)$ continues to grow beyond the value that optimizes the generalization performance. Correspondingly, $\epsilon_g(\alpha)$ goes through a minimum at finite α but, as R increases with α, it converges to a larger value, $\epsilon_g^\infty\equiv\epsilon_g(R=1)$. The learning curves of Fig. 8 are an example of this behavior. Notice that for $0.8086<\delta<\delta_c$ this value of ϵ_g^∞ corresponds to the *largest* value of the student's generalization error. Moreover, for $0.67\,449<\delta<\delta_c$ the asymptotic behavior is even worse than a random guess, because $\epsilon_g(R=1)>0.5$.

At $\delta=\delta_c$ there is an abrupt change of the learning behavior, as beyond this wedge's width the average teacher's stability is negative, and R becomes a decreasing function of α. Correspondingly, the soft margin solution converges to the optimal generalizer in the limit $\alpha\to\infty$. This corresponds to $R=-1$, because for large δ, most of the patterns lie inside the reversed wedge, so that the student's weight vector tends to orient *antiparallel* with the teacher's vector $\mathbf{w}_0$, in order to classify correctly most of the examples. Learning curves for $\delta=2>\delta_c$ obtained with exponent $k=1$ for the slacks exponent in the cost function are represented in Fig. 9.

As for the sandwich rule, the generalization error decays, independently of C, as $\alpha^{-1/2}$ to the corresponding asymptotic

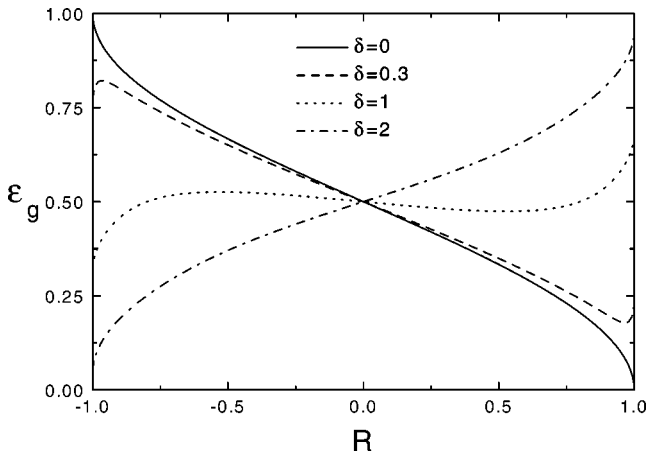

FIG. 7. Reversed wedge rule. Generalization error as a function of the normalized overlap R between the teacher's and the student's weight vectors, for different wedge widths δ.

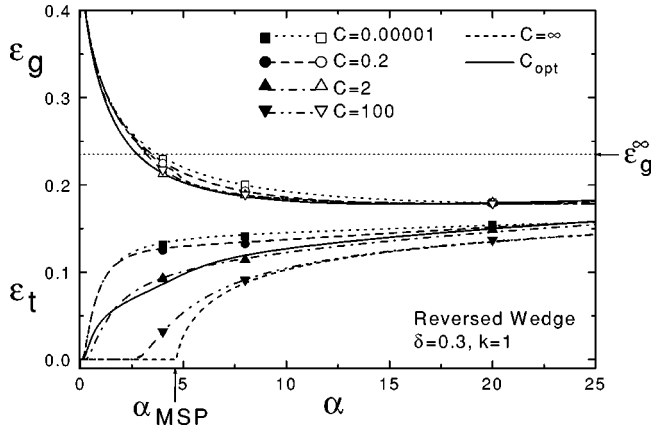

FIG. 8. Reversed wedge rule with $\delta=0.3$. SMC's learning curves corresponding to an exponent $k=1$ in the cost function. The optimal value of the generalization error is $\epsilon_g^{opt}=0.178$, but the SMC converges asymptotically to $\epsilon_g^\infty=0.235$. Simulation results correspond to $N=100$.

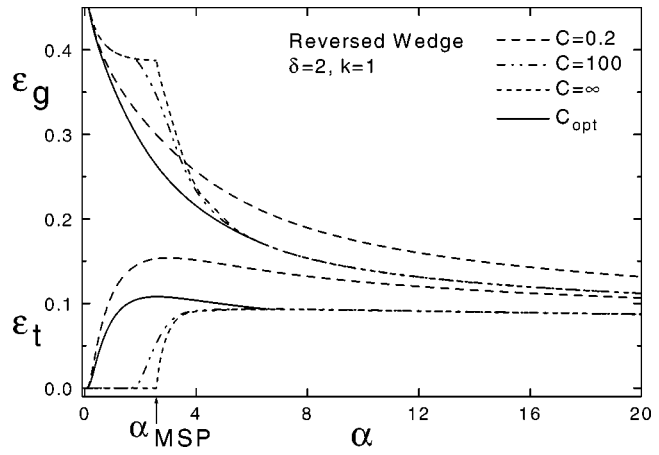

FIG. 9. Reversed wedge rule with $\delta=2$. Learning curves obtained with different values of the hyperparameter C, with $k=1$ in the cost function. Asymptotically, $\epsilon_g^\infty=0.0455$.

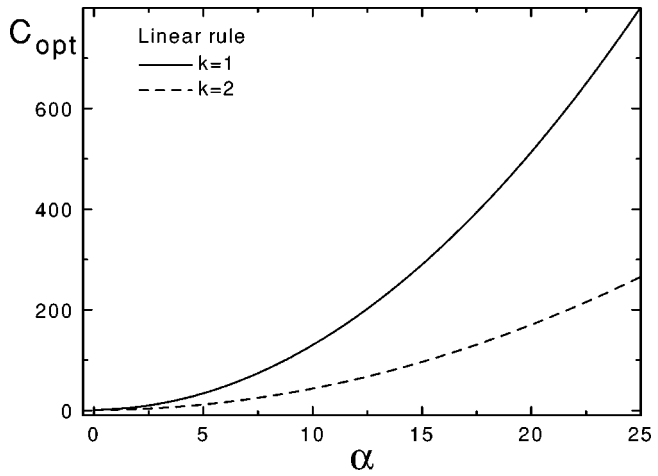

FIG. 10. Linear rule. Optimal values of the hyperparameter C_{opt}.

values, $\epsilon_g^\infty=1-2H(\delta)$ for $\delta<\delta_c$, and $\epsilon_g^\infty=2H(\delta)$ for $\delta>\delta_c$. The same asymptotic behaviors for ϵ_g and R, but with different prefactors, were obtained by Inoue *et al.* [27] for the online Hebbian learning scenario.

The asymptotic value of Q tends to zero as δ tends to δ_c. In the two limiting cases $\delta\to\infty$ and $\delta\to 0$, the task becomes linearly separable and correspondingly $Q\to\infty$.

For the particular case of $\delta=\delta_c$, the only solution of the saddle point equations is $R=0$ for every value of α. This ''no learning'' regime is discussed in Sec. V.

IV. OPTIMIZATION OF THE HYPERPARAMETER

The figures of the preceding section show that the behavior of the generalization error of the SMC is not monotonic with C. It can be seen that there is an optimal value $C_{opt}(\alpha)$ that allows to obtain the minimum generalization error for each α. Obviously, C_{opt} cannot be calculated using the training examples alone, so that in the applications it can only be estimated. Several methods for doing this have been proposed recently [28,29]. We have determined the statistical properties of the optimal SMC by finding $C_{opt}(\alpha)$ for all the rules, thus providing reference curves against which results obtained using the different estimators may be tested.

The optimal generalization curves for the different rules considered in this paper are represented on the figures of the preceding section. Notice that for $\alpha<\alpha_{MSP}$, the MSP is not optimal for any value of α, as it is obtained in the limit $C\to\infty$. In the case of the realizable linear separation, the optimal generalization error of the SMC vanishes asymptotically as $0.488\alpha^{-1}$ for $k=1$, and as $0.449\alpha^{-1}$ for $k=2$. The latter is very close to that of the Bayesian perceptron, $0.442\alpha^{-1}$, but the curves are also very close for finite values of α, as can be seen in Fig. 2. Notice that, even for $k=1$, the asymptotic decay of ϵ_g for the SMC is faster than that of the MSP, which is $\epsilon_g\sim 0.5005\alpha^{-1}$. This is an interesting result, as it shows that, even when a hard margin solution exists, learning with a soft margin machine allows to obtain better classifiers.

For the nonseparable cases, even if C_{opt} allows one to obtain the best performances at finite α, all the learning curves, including the optimal one, behave asymptotically in exactly the same way, as shown in the corresponding sections.

The evolution of C_{opt} with α can be seen in Figs. 10 and 11. The behavior of the curves is qualitatively similar for the shifted linear rule and the reversed wedge with small δ on one hand, and for the sandwich rule and the reversed wedge with large δ on the other. The divergences of C_{opt} are related to the presence of errors with unbounded slack values. For α beyond the divergence, $C_{opt}=\infty$.

V. DISCUSSION

In the preceding sections we presented the learning curves of a SMC learning a variety of rules, characterized by an anisotropy axis parallel to the teacher's vector $\mathbf{w}_0$. Some of the obtained results, and in particular the asymptotic behavior in the $\alpha\to\infty$ limit, can be generalized to other teacher rules (proofs are detailed in the Appendix). As shown by Reimann and Van den Broeck [30], it is useful to characterize the teacher rules by the average patterns' stability of a perceptron aligned with the teacher's vector,

$$\langle\gamma\rangle=\int d\gamma\,\rho(\gamma)=\int Dz\,z\,\mathrm{sgn}[\mathcal{P}(z)], \tag{30}$$

where the second equality in Eq. (30) stems from our assumption (7) that the patterns' distribution is a Gaussian.

In the Appendix we show that in the limit $\alpha\to\infty$, both for $k=1$ and $k=2$, R converges asymptotically either to 1 or to -1, that is, the student perceptron gets either completely aligned or completely antialigned with the teacher's vector. Furthermore, for nonseparable rules, $1-R^2\sim 1/\alpha$. In this limit of $R\to\pm 1$ we find $\epsilon_g\to\epsilon_g^\infty=\int Dz\,\theta(\mp z\,\mathcal{P}(z))$, irrespective of the teacher's rule. The convergence law for this asymptotic value depends on whether or not the polynomial $\mathcal{P}(z)$ defining the rule in Eq. (8) has a root $z_i=0$. If 0 is *not* a root of $\mathcal{P}(z)$, $\mathcal{P}(0)\neq 0$ and $\epsilon_g-\epsilon_g^\infty\sim\exp[-\varepsilon/(1-R^2)]$ with ε a constant, whereas if 0 is a root, then the decay follows the law $\epsilon_g-\epsilon_g^\infty\sim\sqrt{1-R^2}$.

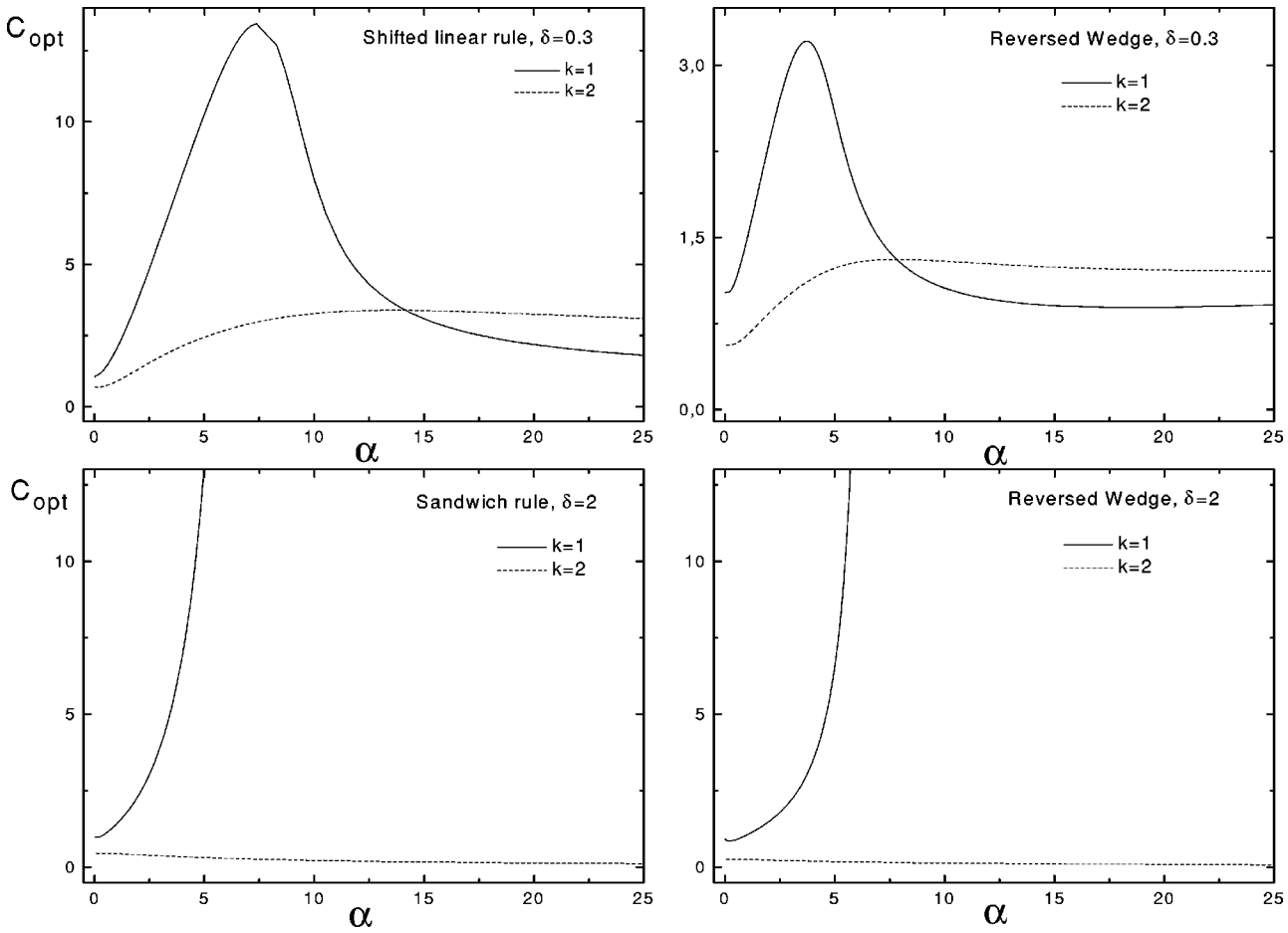

FIG. 11. Optimal values of the hyperparameter C_{opt} for unrealizable rules.

Thus, for the unrealizable rules that have 0 as one of the roots of $\mathcal{P}$, the generalization error decays to the asymptotic value as $\epsilon_g - \epsilon_g^\infty \sim \alpha^{-1/2}$. A similar result has been obtained by Amari *et al.* [31] within the annealed approximation for the case of a deterministic machine learning a noisy teacher (which is unrealizable), and by other authors for Hebbian learning of unrealizable tasks [27,4]. The same power law has been obtained by Meir and Fontanari [6] for a realizable problem learned with inconsistent algorithms, within the approximation of replica symmetry, which is probably not valid for large values of α. Indeed, the soft margin algorithm with finite C is also inconsistent when the rule is the linear separation considered in Sec. III A, and in that case we obtain a different power law decay.

In the case of a linearly separable rule, the SMC with C_{opt} has $\epsilon_g \approx 1/\alpha$, like the MSP, which corresponds to $C=\infty$. However, at fixed finite values of C the decay is slower, like $\sim 1/\alpha^{2/3}$. The same exponent has been obtained for a perceptron learning a separable rule using noisy examples with one step of replica symmetry breaking [18]. Within the replica symmetric approximation to the same problem the exponent is 1/2 instead of 2/3 [2].

In the cases where 0 is not a root of $\mathcal{P}(z)$, like for the shifted linear rule, the decay is exponential, $\epsilon_g - \epsilon_g^\infty \sim \exp(-\varepsilon\alpha)$.

The presence or the absence of a root $z_i = 0$ induces different asymptotic behaviors because if 0 is a root, then a student perceptron aligned with the teacher has $|R|=1$ and can perfectly separate the patterns closest to the hyperplane. In that case, any small misalignment modifies the classification induced by the student, thus strongly modifying the error term in the cost function. On the other hand, if 0 is not a root, the student's hyperplane is immersed in a sea of patterns of the same class. Small tilts of the hyperplane do not change significantly the classification nor the slacks term in the cost.

It is interesting to notice that the figures of the learning curves as well as those of C_{opt} show an analogy between the behavior for the SMCs with bounded slacks, like in the case of the shifted linear rule and that of the reversed wedge when $\delta < \delta_c$, and between those with unbounded slacks, as is the case with the sandwich rule and the reversed wedge when $\delta > \delta_c$. For this last type of rules, C_{opt} diverges beyond some finite α.

Consider now the small α limit. It can be shown that $R \sim \langle \gamma \rangle \sqrt{\alpha}/\sqrt{2\pi}$ and so, $\epsilon_g \sim 1/2 - \langle \gamma \rangle^2 \sqrt{\alpha}/2\pi$. Thus, irrespective of the rule considered, when the fraction of training examples is small, the SMC generalizes better than by random guessing. This is not necessarily the case for larger values of α. The power law found for R in this limit is common

to many learning algorithms of the perceptron, the MSP [24] and Hebb's rule among them.

If we put $R=0$ in the equations, and solve for α, the only possible solution when $\langle\gamma\rangle\neq 0$ is $\alpha=0$. Thus, $R\neq 0$ for all α and has the sign of $\langle\gamma\rangle$ unless it has discontinuous changes of sign. Notice that, given the asymptotic behaviors just mentioned, if R is discontinuous it can only have an even number of changes of sign. A similar result has already been obtained in a broader frame [30]. From the behavior of R in the small α limit, it can be seen that the problem gets very difficult to learn for rules with $\langle\gamma\rangle$ close to 0. In fact, in the very special case of $\langle\gamma\rangle=0$, $R=0$ is a solution of the saddle point equations for every value of α. If this is the only solution, the machine cannot learn at all, as is the case for the reversed wedge rule when $\delta=\delta_c$. This behavior is similar to the one of retarded learning, found in problems of unsupervised learning with quadratic cost functions [30]. In that case, it has been shown that learning is still possible, provided that the cost function is capable of extracting the information about the anisotropy of the distribution of stabilities, contained in its higher order moments [32]. Notice that this is not the case for the cost functions for the SMCs considered in this paper.

VI. CONCLUSION

The properties of the recently proposed support vector machines have been previously studied [16] in two situations of interest, namely, for the cases where the student has either the same structure as the teacher, or it is more complex than it. In both situations the rule to be learned is realizable, and interesting properties of hard margin SVMs, like the existence of hierarchical generalization, could be analyzed within the replica symmetry hypothesis.

In the present paper we addressed the situation where the task is more complex than the learning machine. In this case the cost function for the SVMs is modified. It allows one to obtain a soft margin classifier that results from a trade-off, controlled by a single parameter C, between increasing the margin and minimizing the number of training errors. As the cost function is quadratic and the domain of solutions is convex, we obtain the typical learning curves for a variety of unrealizable tasks using the replica symmetry hypothesis. We considered problems characterized by a single symmetry-breaking direction $\mathbf{w}_0$, along which the patterns have alternating positive or negative class label. We have shown that the convergence of the corresponding learning curves to the asymptotic value follows either a power law or an exponential, depending on the position of the singularities of the teacher's rule.

Even if the student is well adapted to the task's complexity, the SMC may generalize better than the error-free hard margin SVM, provided the hyperparameter C in the cost function is correctly tuned. It can even attain almost Bayesian performance.

We have studied the case of a rule given by a function $\mathcal{P}(z)$, which has a finite number of zeroes. It would be interesting to study the case of a function with an infinite number of zeroes, as, for exemple, $\mathcal{P}(z)=\sin(1/z)$, which is not included in the class of functions we have analyzed.

We showed that the prefactors of the different asymptotic behaviors are proportional to the average stability of the teacher's rule, $\langle\gamma\rangle$. When this vanishes, the SMC with cost function (5) cannot learn, and the overlap between the student and the teacher directions is $R=0$. We considered two exponents for the error term in the cost function, $k=1$ and $k=2$. It would be interesting to study the properties of SMCs trained using exponents $k>2$ in the cost function, as we expect that these should detect the difference of the odd moments of the patterns' distribution in the directions parallel and orthogonal to $\mathbf{w}_0$.

Another interesting question is whether the hierarchical learning of hard margin SVMs exists also with SMCs. To tackle this question, pattern distributions with two different anisotropies have to be considered.

ACKNOWLEDGMENTS

We thank Alex Smola for providing us with the Quadratic Optimizer for Pattern Recognition program [22]. SR-G acknowledges economic support from the EU-research contract ARG/B7-3011/94/97. It is a pleasure to acknowledge support from the Zentrum für Interdisziplinäre Forschung in Bielefeld, where this work was finished in the framework of the Research Group ''The Sciences of Complexity: From Mathematics to Technology to a Sustainable World.'' We are grateful to the Max Planck Institute für Komplexer Systeme in Dresden, where the final corrections to this work were completed in the framework of the seminar. ''Statistical Mechanics of Information Processing in Cooperative Systems'' (March 2001).

APPENDIX

The saddle point equations for the cases $k=1$ and $k=2$ are

$$1-R^2-x=\alpha I_1(xC,\sqrt{Q},R;k), \tag{A1}$$

$$R=-\alpha I_2(xC,\sqrt{Q},R;k), \tag{A2}$$

$$1-R^2=\alpha I_3(xC,\sqrt{Q},R;k), \tag{A3}$$

with, for the case $k=1$,

$$I_1(xC,Q,R;1)=\int_{-1/\sqrt{Q}}^{(xC-1)/\sqrt{Q}}Dt\,t\left(t+\frac{1}{\sqrt{Q}}\right)g(t;R,\mathcal{P}) + \int_{(xC-1)/\sqrt{Q}}^{\infty}Dt\,\frac{txC}{\sqrt{Q}}g(t;R,\mathcal{P}), \tag{A4}$$

$$I_2(xC,Q,R;1)=\int_{-1/\sqrt{Q}}^{(xC-1)/\sqrt{Q}}Dt\frac{1}{2}\left(t+\frac{1}{\sqrt{Q}}\right)^2\frac{\partial g(t;R,\mathcal{P})}{\partial R} + \int_{(xC-1)/\sqrt{Q}}^{\infty}Dt\,\frac{xC}{\sqrt{Q}}\left(t+\frac{2-xC}{2\sqrt{Q}}\right) \times\frac{\partial g(t;R,\mathcal{P})}{\partial R}, \tag{A5}$$

$$I_3(xC,Q,R;1)=\int_{-1/\sqrt{Q}}^{(xC-1)/\sqrt{Q}} Dt\left(t+\frac{1}{\sqrt{Q}}\right)^2 g(t;R,\mathcal{P}) + \int_{(xC-1)/\sqrt{Q}}^{\infty} Dt\frac{(xC)^2}{Q} g(t;R,\mathcal{P}) \quad \text{(A6)}$$

and, for the case $k=2$,

$$I_1(xC,Q,R;2)=\int_{-1/\sqrt{Q}}^{\infty} Dt\frac{2xCt}{1+2xC}\left(t+\frac{1}{\sqrt{Q}}\right) g(t;R,\mathcal{P}), \quad \text{(A7)}$$

$$I_2(xC,Q,R;2)=\int_{-1/\sqrt{Q}}^{\infty} Dt\frac{xC}{1+2xC}\left(t+\frac{1}{\sqrt{Q}}\right)^2 \frac{\partial g(t;R,\mathcal{P})}{\partial R}, \quad \text{(A8)}$$

$$I_3(xC,Q,R;2)=\int_{-1/\sqrt{Q}}^{\infty} Dt\frac{(2xC)^2}{(1+2xC)^2}\left(t+\frac{1}{\sqrt{Q}}\right)^2 g(t;R,\mathcal{P}). \quad \text{(A9)}$$

From Eq. (A9) it can be seen that, for $k=2$, x must vanish in the infinite α limit in order to make $I3$ vanish. Notice that the function $g(t;R,\mathcal{P})$ in Eq. (19) is always non-negative. For the case $k=1$ the analysis of Eq. (A6) shows that x must either vanish or tend to a positive constant with q tending to infinity. This last case can be ruled out by noticing that it is inconsistent with the vanishing of $I2$ [notice that Eq. (A5), as well as Eq. (A8), can be solved analytically].

To show that R can only tend to 1 or -1 in the infinite α limit, it is useful to rewrite $I1$ and $I2$, which in the case $k=1$ are

$$I_1(xC,Q,R;1)=\int_{-1/\sqrt{Q}}^{(xC-1)/\sqrt{Q}} Dt\, g(t;R,\mathcal{P}) + R\, I_2(xC,Q,R;2), \quad \text{(A10)}$$

$$I_2(xC,Q,R;1) = \sum_{i=1}^{Z} \tau(z_i^+)\frac{e^{-z_i^2}}{\sqrt{2\pi}} \times\left\{ \int_{(-1/\sqrt{Q}-z_iR)/\sqrt{1-R^2}}^{[(xC-1)/\sqrt{Q}-z_iR]/\sqrt{1-R^2}} Dt\left(t\sqrt{1-R^2}+\frac{1}{\sqrt{Q}}+z_iR\right) + \frac{xC}{\sqrt{Q}}\int_{[(xC-1)/\sqrt{Q}-z_iR]/\sqrt{1-R^2}}^{\infty} Dt + (z_i\leftrightarrow -z_i),\right\} \quad \text{(A11)}$$

where the z_i, $i=1,\ldots,Z$, are the zeros of the polynomial $\mathcal{P}(z)$ and $\tau(z_i^+)=\mathrm{sgn}[(z_i+z_{i+1})/2]$ for $k=2$,

$$I_1(xC,Q,R;2)=\frac{2xC}{1+2xC}\int_{-1/\sqrt{Q}}^{\infty} Dt\, g(t;R,\mathcal{P}) + R\, I_2(xC,Q,R;2), \quad \text{(A12)}$$

$$I_2(xC,Q,R;2) = \frac{-2xC}{1+2xC}\sum_{i=1}^{Z} \tau(z_i^+)\frac{e^{-z_i^2}}{\sqrt{2\pi}} \times\left\{ \int_{(-1/\sqrt{Q}-z_iR)/\sqrt{1-R^2}}^{\infty} Dt\left(t\sqrt{1-R^2}+\frac{1}{\sqrt{Q}}+z_iR\right) + (z_i\leftrightarrow -z_i)\right\}. \quad \text{(A13)}$$

Let us suppose that R tends to a constant different from 1 and -1 as α tends to infinity. It can be seen that in that case $I1$, $I2$, and $I3$ must vanish at the *same* rate. If we consider teachers with at least one positive root, i.e., *unrealizable* teachers, it can be seen that the integral in $I3$ (the second one for the case $k=1$) never vanishes. Thus, $I3$ must vanish as $(x/\sqrt{Q})^2$ for $k=1$ and as x^2 for $k=2$, if Q tends to a constant or to infinity. But equations (A10) and (A12) show that $I1$ and $I2$ cannot vanish at the same rate as $I3$ because the first term on the right-hand side vanishes as $x/\sqrt{Q}$ for $k=1$ and as x for $k=2$. If Q tends to 0 then I3 must vanish as $(x/\sqrt{Q})^2$ for both cases. But then $I2$ cannot vanish at the same rate, because equations (A11) and (A13) show that $I2$ must vanish as $xC\langle\gamma\rangle/\sqrt{Q}$, unless $\langle\gamma\rangle=0$ (this case will be analyzed below). Therefore, R tends either to 1 or to -1 for all teachers with $\langle\gamma\rangle\neq 0$.

By putting $R=0$ in the equations one can easily [notice that $g(t;0,\mathcal{P})\equiv 1$] see that if $\langle\gamma\rangle\neq 0$, it can only be a solution for $\alpha=0$. On the other hand, for $\langle\gamma\rangle=0$, $R=0$ is a solution for *every* value of α, i.e., learning is impossible for this kind of teacher.

It is also possible to find the condition that makes R go to each one of its limiting values (1 or -1). From what has been said before regarding $I3$ it can be seen that it vanishes as x^2, and so, $1-R^2\sim\alpha x^2$. Using this, and equation (A1) it is evident that $I1$ must vanish faster than x. But, in the infinite α limit, $I1$ is written, to first order, as

$$I_1(xC,Q,R;1)\sim\frac{-x}{\sqrt{Q}}\left\{\mathrm{sgn}(R)\langle\gamma\rangle + \int_{-\infty}^{-1/\sqrt{Q}} Dt\, t\, g(t;\pm 1,\mathcal{P})\right\}, \quad \text{(A14)}$$

$$I_1(xC,Q,R;2)\sim -x\left\{\mathrm{sgn}(R)\frac{\langle\gamma\rangle}{\sqrt{Q}}-\int_{-1/\sqrt{Q}}^{\infty} Dt g(t;\pm 1,\mathcal{P})\right\} - \mathrm{sgn}(R)\sum_{i/|z_i|>1/\sqrt{Q}} \tau(z_i^+) \times\left(|z_i|-\frac{1}{\sqrt{Q}}\right)\frac{e^{-z_i^2}}{\sqrt{2\pi}}. \quad \text{(A15)}$$

Thus, the term within brackets must vanish. For Eq. (A14)

it is evident that this can only happen if $R \to \mathrm{sgn}(\langle \gamma \rangle)$. The same can be shown for Eq. (A15), with a bit of algebra. The asymptotic value of Q can be obtained by imposing the vanishing of the above-mentioned terms.

To see the rate of decay of $1-R^2$, notice that, from Eq. (A3) and from the fact (shown above) that $I3 \sim x^2$, one gets that $1-R^2 \sim \alpha x^2$. But, using the fact that $I1$ must decay faster than x, equations (A11) and (A13) impose that $I2 \sim x$. This, together with Eq. (A2), gives that $x \sim 1/\alpha$. Therefore, $1-R^2 \sim 1/\alpha$.

[1] E. Gardner, Europhys. Lett. **4**, 481 (1987).

[2] G. Györgyi and N. Tishby, in *Neural Networks and Spin Glasses*, edited by W. K. Theumann and R. Köberle (World Scientific, Singapore, 1990), p. 3.

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

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

[5] P. Peretto, *An Introduction to the Modeling of Neural Networks* (Cambridge University Press, Cambridge, UK, 1992).

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

[7] O. Kinouchi and N. Caticha, Phys. Rev. E **54**, R54 (1996).

[8] A. Buhot, J. M. Torres-Moreno, and M. B. Gordon, Phys. Rev. E **55**, 7434 (1997).

[9] V. Vapnik, *The Nature of Statistical Learning Theory* (Springer Verlag, New York, 1995).

[10] C. Cortes and V. N. Vapnik, Machine Learning **20**, 273 (1995).

[11] W. Krauth and M. Mezard, J. Phys. A **20**, L745 (1987).

[12] C. J. C. Burges and D. J. Crisp, in *Advances in Neural Information Processing Systems 12*, edited by S. A. Solla, T. K. Leen, and K-R. Muller (MIT Press, Cambridge, MA, 2000), p. 223.

[13] B. Martos, *Nonlinear Programming Theory and Methods* (North-Holland, Amsterdam, 1975).

[14] S. Risau-Gusman and M. B. Gordon, in *Advances in Neural Information Processing Systems 12* (Ref. [12]), p. 321.

[15] S. Risau-Gusman and M. B. Gordon, Phys. Rev. E **62**, 7092 (2000).

[16] R. Dietrich, M. Opper, and H. Sompolinsky, Phys. Rev. Lett. **82**, 2975 (1999).

[17] M. Opper and W. Kinzel, in *Models of Neural Networks III*, edited by E. Domany, J. L. van Hemmen, and K. Schulten (Springer-Verlag, Heidelberg, 1995), p. 151.

[18] T. Uezu and Y. Kabashima, J. Phys. A **29**, L55 (1996).

[19] G. Györgyi and P. Reimann, Phys. Rev. Lett. **79**, 2746 (1997).

[20] Here, as in the rest of the paper, we call zeros the points where the function changes sign.

[21] This is true only if the patterns are in a general position (that is, if all the patterns in every subset of N patterns are linearly independent) [12]. For the Gaussian pattern distribution we are considering, the probability that M points are not in a general position vanishes in the thermodynamic limit.

[22] Program available upon request to http://svm.first.gmd.de.

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

[24] M. B. Gordon and D. R. Grempel, Europhys. Lett. **29**, 57 (1995).

[25] T. Cover, IEEE Trans. Electron. Comput. **14**, 326 (1965).

[26] T. L. H. Watkin and A. Rau, Phys. Rev. A **45**, 4102 (1992).

[27] J.-I. Inoue, H. Nishimori, and Y. Kabashima, J. Phys. A **30**, 3795 (1997).

[28] P. Sollich, in *Advances in Neural Information Processing Systems 12* (Ref. [12]), p. 349.

[29] M. Seeger, in *Advances in Neural Information Processing Systems 12* (Ref. [12]), p. 603.

[30] P. Reimann and C. Van den Broeck, Phys. Rev. E **53**, 3989 (1996).

[31] S.-I. Amari, N. Fujita, and S. Shinomoto, Neural Comput. **4**, 605 (1992).

[32] A. Buhot and M. B. Gordon, Phys. Rev. E **57**, 3326 (1998).