Learning by examples from a nonuniform distribution

P. Reimann* and C. Van den Broeck
Limburgs Universitair Centrum, 3590 Diepenbeek, Belgium

(Received 28 August 1995)

We present a general replica calculation for learning from examples generated by a nonuniform pattern distribution with a single symmetry-breaking orientation. Our results cover the three main learning scenarios: storage of patterns with random classifications by a perceptron, supervised learning from a teacher, and unsupervised learning. We show that for a perceptron the critical storage capacity $\alpha_c=2$ is completely independent of the pattern distribution provided it is point symmetric or provided the classification as ±1 is unbiased. In a particular model for supervised learning we find that an ideal (Bayes) student learns most from a few examples if they are easy and from a large number if they are difficult. Learning based on the minimization of a specific class of (quadratic) cost functions is solved completely for all three scenarios.

PACS number(s): 87.10.+e, 05.20.−y, 02.50.−r

I. INTRODUCTION

Over the last few years, several simple models describing the process of "learning from examples" have been analyzed using powerful techniques borrowed from the statistical mechanics of spin glasses [1]. One distinguishes the storage problem [2], in which one searches for a perceptron that reproduces the correct classification for these patterns, supervised learning [3], where a "teacher" provides the classification for these patterns, and unsupervised learning [4,5], where one tries to discover the characteristics of the underlying nonuniform distribution that generates the patterns. The purpose of this paper is to present a general replica calculation which, within the validity of replica symmetry, covers all three situations and which includes most learning algorithms of interest such as Bayes, Gibbs, maximal *a posteriori* probability, and the minimization of *ad hoc* cost functions. Our only restriction is that we consider the case of a single symmetry-breaking orientation $\boldsymbol{B}$ along which the pattern distribution is nonuniform and which at the same time plays the role of the teacher in the case of supervised learning.

In Sec. II the general framework of unsupervised learning considered here is introduced. It is shown how both supervised learning and the capacity problem with nonuniformly distributed examples can be transformed into an unsupervised problem. In Sec. III we demonstrate that all quantities of interest in the different learning scenarios can be reduced to the study of a single partition function. Our central result for the associated free energy based on a standard replica calculation follows from Eq. (13) below under the assumption of replica symmetry. The corresponding local stability condition is given in (18) and turns out to be (at least marginally) fulfilled for all examples studied in Secs. IV–VI. Various expressions for the free energy and the stability condition derived previously in the literature are recovered as special cases.

As a first application of the general framework we investigate in Sec. IV the capacity problem for an arbitrary nonuniform distribution of the patterns. We prove that the maximal storage capacity $\alpha_c=2$ is independent of the details of this distribution provided it is invariant under $\boldsymbol{B}\mapsto-\boldsymbol{B}$ or provided that the classification of the patterns as ±1 is unbiased. Moreover, we show that $\alpha_c>2$ in all other generic cases. As a further application we study in Sec. V a particular model of supervised learning of a teacher perceptron with a Gaussian nonuniform distribution of the examples. We find that in the best possible case, where the student follows the Bayes learning rule, the generalization ability from a few examples is optimal if most of them are easy (far from the decision plane of the teacher). On the other hand, if a lot of examples are taught to the student, then it is favorable to choose difficult ones. Analogous results have been found in Ref. [6] for a related supervised problem (which also fits into our general framework).

In Sec. VI we present a solution for all three learning scenarios with arbitrary pattern distributions if the learning strategy is based on the minimization of a cost function with a quadratic *ad hoc* potential, including maximal variance, Hebb, and adaline learning rules as special cases. In the final Sec. VII, implications of our findings, in particular with respect to the phenomenon of "retarded classification" [5], are discussed and their relation is pointed out to results that will be presented in detail elsewhere.

II. LEARNING FROM EXAMPLES

We first formulate the general setup for *unsupervised learning* considered in this paper. We assume that a set of patterns $\{\xi^\mu\}^p:=\{\xi^1,\xi^2,\dots,\xi^p\}$ is generated by p independent samplings from a nonuniform probability distribution $P^*(\xi|\boldsymbol{B})$, where $\boldsymbol{B}$ represents a symmetry-

*Present address: Eötvös University, Puskin utca 5-7, H-1088 Budapest, Hungary.

breaking orientation. Under the further assumption of cylindrical symmetry around the $\boldsymbol{B}$ axis, one can always write this so-called *a priori* probability in the following form:

$$P^*(\xi|\boldsymbol{B}) \propto \exp\{-V^*(\lambda)\}\delta(\xi^2-N) , \tag{1}$$

where $\lambda := \xi\cdot\boldsymbol{B}/\sqrt{N}$ is the overlap, N stands for the dimension of the space, and the proportionality $\propto$ accounts for a missing normalization constant on the right-hand side (RHS). Both patterns ξ and other N-dimensional vectors such as $\boldsymbol{B}$ are henceforth supposed to be normalized according to $\xi^2=\boldsymbol{B}^2=N$.

The corresponding distribution $P^*(\lambda)$ of the overlaps λ is easily found to be

$$\begin{aligned} P^*(\lambda) &= \int \delta(\lambda-\xi\cdot\boldsymbol{B}/\sqrt{N})P^*(\xi|\boldsymbol{B})d\xi \\ &\propto \exp\{-\lambda^2/2-V^*(\lambda)\} \end{aligned} \tag{2}$$

(here and in the following integration limits $\pm\infty$ are dropped). In particular, for a uniform distribution of patterns on the sphere one has that $V^*(\lambda)\equiv 0$ and hence λ is a normal random variable. In this case, there is no preferential direction and the choice of $\boldsymbol{B}$ is irrelevant. Note that the cosine of the angle between ξ and $\boldsymbol{B}$ is given by $\lambda/\sqrt{N}$ and the patterns ξ are thus concentrated for large N in a very small belt about the "equator" plane perpendicular to $\boldsymbol{B}$. Therefore we will sometimes call λ a "microscopic" overlap. The distribution of these overlaps (2) will play a central role in our subsequent calculations.

The general task in unsupervised learning is to infer a structure or rule from the available examples $\{\xi^\mu\}^p$ [4,5]. This is only possible on the basis of additional *a priori* knowledge, so that one can make a reasonable guess about the structure or rule to be inferred with not too many free parameters to be fitted. In our case, we will take for granted that the rule behind the examples is of the general form (1) and our task is to make a guess $\boldsymbol{J}$ about the unknown symmetry-breaking orientation $\boldsymbol{B}$.

If the form of the *a priori* probability (1) that generates the patterns is known exactly (but not the vector $\boldsymbol{B}$), one can obtain the so-called *a posteriori* probability that the unknown $\boldsymbol{B}$ coincides with a particular hypothesis vector $\boldsymbol{J}$ by using the familiar Bayes rule. This rule exploits the fact that the joint *a priori* probability $P(\boldsymbol{J},\{\xi^\mu\}^p)$ is proportional to both $P(\boldsymbol{J}|\{\xi^\mu\}^p)P(\{\xi^\mu\}^p)$ and $P^*(\{\xi^\mu\}^p|\boldsymbol{J})P(\boldsymbol{J})$ with the uniform *a priori* probabilities $P(\{\xi^\mu\}^p)\propto\prod_{\mu=1}^p\delta((\xi^\mu)^2-N)$ and $P(\boldsymbol{J})\propto\delta(\boldsymbol{J}^2-N)$. One thus finds

$$P(\boldsymbol{J}|\{\xi^\mu\}^p) \propto \exp\left\{-\sum_{\mu=1}^{p} V^*(\lambda^\mu)\right\}\delta(\boldsymbol{J}^2-N) \tag{3}$$

for this so-called *a posteriori* probability, where $\lambda^\mu := \xi^\mu\cdot\mathbf{J}/\sqrt{N}$.

Several options to select a particular $\boldsymbol{J}$ vector on the basis of this result are now open. First, one can sample at random a vector $\boldsymbol{J}_G$ from the probability distribution given in Eq. (3). This scenario will be called *Gibbs* or *Boltzmann learning*. Second, one can take the center of mass of the $\boldsymbol{J}$ vectors (with a properly normalized length)

$$\boldsymbol{J}_B \propto \int d\boldsymbol{J}\,\boldsymbol{J}P(\boldsymbol{J}|\{\xi^\mu\}^p) , \tag{4}$$

corresponding to *Bayes* or *optimal learning* [5]. Finally, one can sample a $\boldsymbol{J}$ vector $\boldsymbol{J}_M$ from the set that maximizes the probability given in Eq. (3), known as *maximum a posteriori probability* or *maximal likelihood learning*.

By exploiting results from Ref. [5] it is shown in Appendix A that the Bayes vector $\boldsymbol{J}_B$ makes the smallest angle with the unknown $\boldsymbol{B}$ among all hypotheses $\boldsymbol{J}$ that can be inferred from a given set of examples $\{\xi^\mu\}^p$. Furthermore, there exists a simple relation between the cosine of this angle $R_B(\alpha)=\boldsymbol{J}_B\cdot\boldsymbol{B}/N$ and the one following from Gibbs learning $R_G(\alpha)=\boldsymbol{J}_G\cdot\boldsymbol{B}/N$:

$$R_B(\alpha)=\sqrt{R_G(\alpha)} . \tag{5}$$

These results are valid in the limit $N\rightarrow\infty$ with $\alpha := p/N \geq 0$ fixed, under some weak conditions specified in Appendix A. In particular, the relation $R_B(\alpha)\geq R(\alpha)$ applies to *any* neural network of *any* architecture for which the overlap $R(\alpha)$ between the hypothesis $\boldsymbol{J}$ and $\boldsymbol{B}$ is self-averaging.

In some cases, the *a priori* potential $V^*(\lambda)$ in (1) is only roughly known, so that the distribution (3) cannot be constructed. A simple and often used procedure is then to choose the properly normalized $\boldsymbol{J}$ vector that minimizes a cost function E of the following form:

$$E(\boldsymbol{J})=\sum_{\mu=1}^{p} V(\lambda^\mu) \tag{6}$$

with $\lambda^\mu := \xi^\mu\cdot\boldsymbol{J}/\sqrt{N}$. The potential $V(\lambda)$ is chosen in an *ad hoc* fashion, in the hope that it captures some of the structure of the true pattern distribution [4]. Note that the $\boldsymbol{J}$ vector which minimizes the *ad hoc* cost function (6) can also be obtained as the vector that maximizes the probability distribution (3), provided one chooses $V^*(\lambda)=V(\lambda)$.

We now turn to the case of *supervised learning*. In addition to the patterns $\{\xi^\mu\}^p$ that are generated by independent samplings from the *a priori* distribution (1), a teacher provides the classifications ξ_0^μ, $\mu=1,\ldots,p$, for each of them. We restrict ourselves to the simplest case of binary classification, $\xi_0^\mu=+1$ or -1. Within the context of a problem with a single symmetry-breaking orientation, we assume that the $\boldsymbol{B}$ vector in the pattern distribution (1) also controls the classification:

$$\xi_0=\mathrm{sgn}[f(\lambda=\boldsymbol{B}\cdot\xi/\sqrt{N})] , \tag{7}$$

where $f(\lambda)$ is a general function of its argument. The aim in supervised learning is to construct a student vector $\boldsymbol{J}$ that can reproduce the classification rule (7) as well as possible. To make the connection with unsupervised learning we assume that $f(\lambda)$ is known to be an odd function of its argument. Consequently, the classification of a pattern ξ as ξ_0 automatically implies that the pattern $-\xi$ is classified as $-\xi_0$. One can thus concentrate on the set of patterns $\xi^\mu\xi_0^\mu$, $\mu=1,\ldots,p$, all of which are classified as $+1$. The overlap distribution $P^{**}(\lambda)$ of these "aligned" patterns readily follows from the original one (2) as

$$P^{**}(\lambda)=[P^*(\lambda)+P^*(-\lambda)]\Theta(f(\lambda)) \quad \text{(supervised)} , \tag{8}$$

where the step function $\Theta(x)$ is 1 for $x>0$ and 0 for $x\leq 0$. The corresponding potential $V^{**}(\lambda)$ is given (up to a free additive constant) by $-\ln P^{**}(\lambda)-\lambda^2/2$, cf. (2). In particular, we have that $V^{**}(\lambda)=\infty$ for $f(\lambda)\leq 0$. The teacher vector $\boldsymbol{B}$ can now be identified as the symmetry-breaking orientation for the patterns with positive classification and the student $\boldsymbol{J}$ with the hypothesis vector. In this way, *we have transformed the supervised problem into an unsupervised one.* For example, in the case of a teacher perceptron $\Theta(f(\lambda))=\Theta(\lambda)$ with uniformly distributed patterns $V^*(\lambda)\equiv 0$, the equivalent unsupervised problem is characterized by a uniform distribution of patterns but restricted to the upper hemisphere with $\boldsymbol{B}$ as north pole, $V^{**}(\lambda)=0$ for $\lambda\geq 0$ and $V^{**}(\lambda)=\infty$ for $\lambda<0$.

Finally we turn to the *capacity problem.* In this case, the classifications ξ_0^μ of the training patterns are supposed to be random and one searches for perceptrons with $\boldsymbol{J}$ vectors that reproduce this classification. Since the condition $\operatorname{sgn}(\boldsymbol{\xi}\cdot\boldsymbol{J}/\sqrt{N})=\xi_0$ is equivalent to the condition $\tilde{\lambda}:=\boldsymbol{\xi}\xi_0\cdot\boldsymbol{J}/\sqrt{N}\geq 0$, it is again convenient to absorb the classification in the patterns with a resulting distribution $P^{**}(\lambda)$ of the transformed patterns along $\boldsymbol{B}$

$$P^{**}(\lambda)=\frac{1+m}{2}P^*(\lambda)+\frac{1-m}{2}P^*(-\lambda)\ \text{(capacity)}\ , \tag{9}$$

where $(1\pm m)/2$, $m\in[-1,1]$, is the probability for a classification $\xi_0=\pm 1$. The most interesting case of unbiased classifications corresponds to $m=0$. The condition that $\tilde{\lambda}\geq 0$ for any pattern of the training set can be incorporated by the following simple choice of the *ad hoc* potential in (6):

$$V(\lambda)=\begin{cases}\infty & \text{for } \lambda<0\\ 0 & \text{for } \lambda\geq 0\ .\end{cases} \tag{10}$$

So *the capacity problem has also been transformed into an unsupervised problem* although one characterized by a specific *ad hoc* potential. The extension to other potentials $V(\lambda)$ that have been studied in the literature is obvious (see, e.g., [7], and further references therein). As a further generalization one may include a λ dependence in the probability $[1\pm m(\lambda)]/2$ for a classification $\xi_0=\pm 1$ to study the effect of correlations between the patterns and their classifications [8]. Note that $1\pm m$ in (9) then becomes $1\pm m(\pm\lambda)$ and thus supervised learning (8) is recovered as special case $m(\lambda)=\operatorname{sgn}[f(\lambda)]$. So an appropriately chosen $m(\lambda)$ can also be considered to represent the case of a "noisy teacher" [3].

We end with a technical note concerning symmetric potentials $V^*(-\lambda)=V^*(\lambda)$. In this case, one finds that the right-hand side in Eq. (4) vanishes. Similar problems arise for a symmetric $V(\lambda)$ in (6). In the following, we adopt the usual solution to this problem by treating symmetric potentials $V^*(\lambda)$ or $V(\lambda)$ as limiting cases of appropriate asymmetric approximants.

III. REPLICA CALCULATION

As we shall see, all the cases discussed in the previous section can be reduced to the study of the partition function

$$Z=\int d\boldsymbol{J}\exp\left\{-\beta\sum_{\mu=1}^{p}V(\lambda^\mu)\right\}\delta(\boldsymbol{J}^2-N) \tag{11}$$

with appropriate choices of $V(\lambda)$ and the "inverse temperature" β. The associated free energy F is expected to be an extensive quantity in N, self-averaging with respect to the pattern distribution $P^*(\boldsymbol{\xi}|\boldsymbol{B})$ from (1), and independent of $\boldsymbol{B}$:

$$-\beta F=\ln Z=\langle \ln Z\rangle^*\ . \tag{12}$$

The average $\langle\ \rangle^*$ over the pattern distribution can be performed using the replica trick [9]. Assuming replica symmetry one finds by means of a standard calculation that [10]

$$-\beta F=N\operatorname*{extr}_{q,R}G(q,R)\ , \tag{13}$$

$$G(q,R):=\frac{\ln(1-q)}{2}+\frac{1-R^2}{2(1-q)}+\alpha\int D^*t\int Dz\,\ln\left\{\int D\tau\, e^{-\beta V(\sigma)}\right\}, \tag{14}$$

$$\sigma:=tR+z\sqrt{q-R^2}+\tau\sqrt{1-q}\ , \tag{15}$$

where $\alpha:=p/N$,

$$Dz:=dz\exp\{-z^2/2\}/\sqrt{2\pi},$$

$$D\tau:=d\tau\exp\{-\tau^2/2\}/\sqrt{2\pi},$$

and $D^*t:=dtP^*(t)$, cf. Eq. (2). Generically, the extremization procedure with respect to q and R has a unique solution $q(\alpha)$ and $R(\alpha)$. These so-called order parameters have the usual meaning: $q(\alpha)=\boldsymbol{J}^a\cdot\boldsymbol{J}^b/N$ is the self-overlap between two typical vectors $\boldsymbol{J}$ from different replicas (Edward-Anderson parameter [1]) and $R(\alpha)=\boldsymbol{J}\cdot\boldsymbol{B}/N$ is the overlap between a typical vector $\boldsymbol{J}$ and the vector $\boldsymbol{B}$. The word typical refers to those vectors that give the exponentially dominant contribution to F for large N and is motivated by the fact that these overlaps are expected to be self-averaging. By closer inspection one can see that for finite α and β and under very weak conditions on $V(\lambda)$ and $V^*(\lambda)$ these extremizing $q=q(\alpha)$ and $R=R(\alpha)$ in (13) satisfy $-1<R(\alpha)<1$ and $[R(\alpha)]^2<q(\alpha)<1$ and thus they can be determined by studying the zeros of $\partial G(q,R)/\partial q$ and $\partial G(q,R)/\partial R$ without investigating separately $G(q,R)$ at the boundaries of the allowed q-R regime. It is only in the zero-temperature limit $\beta\to\infty$ that $q(\alpha)$ *may* tend to 1, typically with

$$x(\alpha):=\beta[1-q(\alpha)] \tag{16}$$

converging to a value satisfying $0<x(\alpha)<\infty$. Then the extremization problem (13) readily simplifies to

$$-F=N\operatorname*{extr}_{x,R}\left\{\frac{1-R^2}{2x}-\alpha\int D^*t\int Dz\,\min_{\lambda}\left[V(\lambda)+\frac{(\lambda-s)^2}{2x}\right]\right\}, \tag{17}$$

where $s:=tR+z\sqrt{1-R^2}$. This "streamlined" zero-temperature version (17) of the extremization problem (13) is expected to be valid [i.e., $x(\alpha)$ from (16) stays finite for $\beta\rightarrow\infty$] if the cost function (6) has a unique (quadratic) absolute minimum. For more details we refer to [7,11–13].

All of the above results rely on the assumption of replica symmetry. Following an adaptation of the usual arguments [2,3,14–16], one finds the following condition for the local stability of the replica symmetric solution (13):

$$1>\alpha\int D^*t\int Dz[\rho(t,z)]^2\ , \tag{18}$$

$$\rho(t,z):=\frac{\int D\tau\int D\tau'[1-(\tau-\tau')^2/2]e^{-\beta[V(\sigma)+V(\sigma')]}}{\int D\tau\int D\tau' e^{-\beta[V(\sigma)+V(\sigma')]}}\ , \tag{19}$$

where σ' is defined as σ in (15) but with τ' instead of τ and the values of q and R entering σ and σ' are those that extremize Eq. (14), i.e., $q=q(\alpha)$, $R=R(\alpha)$. Within the validity of the "streamlined zero-temperature formalism" (17), the condition (18) simplifies to

$$1>\alpha\int D^*t\int Dz\left[\frac{\partial\lambda_0(s=s(\alpha),x(\alpha))}{\partial s}-1\right]^2\ , \tag{20}$$

where $\lambda_0(s,x)$ denotes the minimizing λ in (17) and

$$s(\alpha):=tR(\alpha)+z\sqrt{1-[R(\alpha)]^2}\ .$$

We henceforth will restrict ourselves to situations for which the replica symmetric solution (13) is valid and we tacitly will take this for granted if (18) is satisfied.

Next we discuss how our central result (13) relates to the different learning scenarios introduced in Sec. II. We first address the various learning rules for unsupervised learning. In this case, the value of $R=R(\alpha)$ is of basic interest. To find this value $R_G(\alpha)$ for Gibbs learning, one has to set $V(\lambda)=V^*(\lambda)$ and $\beta=1$ in Eq. (14). Indeed, with this choice, one samples from the *a posteriori* probability, as is clear from a comparison of Eq. (11) with Eq. (3). One can further simplify calculations [3,5] by noting that the symmetry-breaking orientation $\boldsymbol{B}$ and (the different replicas of) the hypothesis vector $\boldsymbol{J}$ play an equivalent role in the free energy (12) and thus one can focus on $q=R\in[0,1]$ in the extremization problem (13) [17]. Once $R_G(\alpha)$ is known, the overlap $R_B(\alpha)$ for Bayes learning follows immediately from (5). To obtain $R_M(\alpha)$ corresponding to maximum *a posteriori* probability learning, one has to set $V(\lambda)=V^*(\lambda)$ and $\beta\rightarrow\infty$ in (14). Indeed, in this way one obviously selects the $\boldsymbol{J}$ vectors that maximize the *a posteriori* probability (3). Finally, in view of (11) the overlap $R(\alpha)$ corresponding to the minimization of the cost function (6) with an *ad hoc* potential $V(\lambda)$ follows from (13) by letting $\beta\rightarrow\infty$ in (14).

The free energies (13) or (17) as well as the stability condition (18) or (20) that are obtained in all these variants of unsupervised learning reduce to those mentioned in the literature by filling in the specific form of the nonuniform distribution $P^*(\lambda)$ that was used; see, e.g., [4] for an example with an *ad hoc* potential and [5] for the application of the Gibbs, Bayes, and maximum *a posteriori* probability rules to the same model.

We now turn to the other two learning scenarios. As discussed in the previous section, the capacity problem for a perceptron can be reduced to that of unsupervised learning by working with the transformed pattern distribution $P^{**}(\lambda)$ from (9) instead of $P^*(\lambda)$ and with the specific choice (10) for the potential $V(\lambda)$. Here, a particularly interesting problem [2,15] is to locate the α value for which only one $\boldsymbol{J}$ vector remains that reproduces correctly the classification of all examples. This so-called critical storage capacity α_c can be identified with the α value for which $q(\alpha)\rightarrow 1$.

The case of supervised learning is transformed into an unsupervised problem by choosing $P^{**}(\lambda)$ according to (8). From there on one can repeat the discussion given above concerning the various learning rules. It is easily verified that the corresponding free energy from Eq. (13) or (17) and the stability condition (18) or (20) reduce for specific choices of $P^*(\lambda)$ and $f(\lambda)$ to the results given in the literature; see, e.g., [3,6,13,18–20]. Unlike in the capacity problem, the aim in supervised learning [3] is not merely the correct memorization of the examples by the student but also the correct application of the learned rule to new "test" examples. The generalization ability of the student with the hypothesis vector $\boldsymbol{J}$ is usually quantified through the generalization error $\epsilon(\alpha)$, defined as the probability for disagreement with the teacher on a new example. This probability can be readily obtained from the overlap $R(\alpha)$ as follows [3,6,21]:

$$\epsilon(\alpha)=\int d\lambda\, P^*_{\text{test}}(\lambda)\times\int Dz\,\Theta[-f(\lambda)f(\lambda R(\alpha)+z\sqrt{1-[R(\alpha)]^2})]\ , \tag{21}$$

where $P^*_{\text{test}}(\lambda)$ represents a distribution of the test examples of the same general form (2) as the training patterns but not necessarily identical.

IV. CAPACITY PROBLEM FOR NONUNIFORMLY DISTRIBUTED PATTERNS

As a first application of the replica calculation presented in the previous section, we consider the capacity problem for a perceptron. A set of patterns is generated according to an arbitrary *a priori* probability (1). Each one receives a random binary classification valued ±1 with probability $(1\pm m)/2$, $m\in[-1,1]$. The corresponding distribution $P^{**}(\lambda)$ follows from Eq. (9) and the potential $V(\lambda)$ is given in Eq. (10). We ask for the critical storage capacity α_c above which no perceptron can be found that performs these classifications without error. This capacity α_c can be identified with the α value for which $q(\alpha)\rightarrow 1$.

To this end we have to consider the extremization problem (13) and (14) but with $D^{**}t=dt\,P^{**}(t)$ instead of D^*t. With (10) the condition $\partial G(q,R)/\partial q=0$ then yields

$$q-R^2=\alpha\int D^{**}t\int Dz\left[tR+z\frac{1-R^2}{\sqrt{q-R^2}}\right]\times\frac{\sqrt{1-q}\,e^{-u^2/2}}{\int_{-u}^{\infty}d\tau\, e^{-\tau^2/2}}\,, \qquad (22)$$

where $u:=(tR+z\sqrt{q-R^2})/\sqrt{1-q}$. For q approaching 1 from below we have by definition that $\alpha\rightarrow\alpha_c$ and the integrand in (22) can be simplified by means of de l'Hôpital's rule, resulting in

$$1-R^2=\alpha_c\int D^{**}t\int_{-\infty}^{-v}Dz[tR+z\sqrt{1-R^2}]^2\,, \qquad (23)$$

where $v:=tR/\sqrt{1-R^2}$. Similarly, the condition $\partial G(q,R)/\partial R=0$ implies for $q\rightarrow 1$ that

$$R=\alpha_c\int D^{**}t\int_{-\infty}^{-v}Dz[tR+z\sqrt{1-R^2}]\times[-t+zR/\sqrt{1-R^2}]\,. \qquad (24)$$

The trivial solution $\alpha_c=0$ and thus $R=\pm 1$ can be excluded by means of the extremization condition (13) and (14) and we henceforth can assume that $\alpha_c>0$ and $|R|<1$. Subtraction of R times Eq. (23) from $1-R^2$ times Eq. (24) then yields

$$0=\int D^{**}t[te^{-v^2/2}\sqrt{(1-R^2)/2\pi}-t^2R\int_{-\infty}^{-v}Dz]\,, \qquad (25)$$

where we exploited that $\int_{-\infty}^{-v}Dz\,z=-e^{-v^2/2}/\sqrt{2\pi}$.

In the limit of a symmetric pattern distribution *or* unbiased classifications ($m\rightarrow 0$) the distribution $P^{**}(\lambda)$ becomes symmetric, cf. (9). Consequently, the first summand in (25) vanishes and we can infer that $R=R(\alpha_c)=0$ and with (23) that $\alpha_c=2$. This result is completely independent of the pattern distribution $P^*(\lambda)$ (provided it is symmetric or $m\rightarrow 0$). It agrees with the well known result [2] for uniformly distributed patterns, and is compatible with Cover's theorem for patterns in a "general" configuration [22]. A somewhat similar result for nonuniformly distributed patterns has also been obtained in Ref. [23]. Using $R(\alpha_c)=0$, one can infer that the stability condition (18) is identical to that for uniformly distributed patterns and is thus marginally satisfied.

In Appendix B a more general version of this result is derived, namely,

$$\alpha_c=2 \quad \text{for } \bar{\lambda}=0 \qquad (26)$$

and

$$\alpha_c>2 \quad \text{for } \bar{\lambda}\neq 0\,, \qquad (27)$$

where we introduced

$$\bar{\lambda}:=\int d\lambda\,\lambda P^{**}(\lambda)\,. \qquad (28)$$

It is also demonstrated that

$$R(\alpha_c)=0 \quad \text{if and only if } \bar{\lambda}=0\,, \qquad (29)$$

which will lead to an interesting connection between the capacity problem and the phenomenon of retarded classification, cf. the discussion in Sec. VII. As before, the stability condition (18) turns out to be always marginally satisfied.

If in Eq. (9) the pattern distribution $P^*(\lambda)$ is asymmetric *and* $m\neq 0$, then generically $\bar{\lambda}$ from (28) will be nonzero and the solution of Eq. (25) will be no longer at $R=0$, and hence the α_c following from (23) will be larger than 2 according to (27) and (29). As a simple explicit example, we mention $P^*(\lambda)\propto\Theta(\lambda)\exp\{-\lambda^2/2\}$ and $m=1$. The corresponding unsupervised problem (9) and (10) becomes identical to the supervised student-teacher perceptron scenario [$\Theta(f(\lambda))=\Theta(\lambda)$ in (7), $P^*(\lambda)\propto\exp\{-\lambda^2/2\}$ in (8), and $V(\lambda)$ as in (10)]. Furthermore, one readily sees that we are dealing with Gibbs learning in the equivalent unsupervised problem with the property $q(\alpha)=R(\alpha)<1$ for all $\alpha<\infty$ and thus $\alpha_c=\infty$.

The fact that the storage capacity α_c is at least equal to 2 can be easily understood as follows. Since $\boldsymbol{B}$ is the only symmetry-breaking direction, the pattern distribution is uniform in the subspace orthogonal to $\boldsymbol{B}$. The capacity problem in this subspace is of the usual Gardner type [2,15], and a solution $\boldsymbol{J}$ orthogonal to $\boldsymbol{B}$ exists up to $\alpha=2$. Note that this argument also applies in the presence of a finite number of symmetry-breaking orientations. The argument breaks down if the patterns are not drawn independent from each other and a capacity smaller than 2 can result; see, e.g., [8].

We finally mention that the capacity problem with "biased" patterns ξ^μ *and* classifications ξ_0^μ studied in Ref. [2] fits into the general framework of this section only in certain limiting cases and then indeed leads to identical results.

V. SUPERVISED LEARNING FROM GAUSSIAN DISTRIBUTED EXAMPLES

In this section the application of the general framework from Secs. II and III is exemplified for supervised learning: A teacher, which for simplicity is assumed to coincide with the symmetry-breaking orientation $\boldsymbol{B}$ of the patterns, provides a classification (7) for each of them. Our aim is to find a student $\boldsymbol{J}$ that extracts this hidden rule (7) from the training patterns and as measure of success we take the overlaps $R(\alpha)$ with $\boldsymbol{B}$. We concentrate on the best possible student following the Bayes learning rule and consider the case of a teacher perceptron, i.e., $\Theta(f(\lambda))=\Theta(\lambda)$ in (7). We further restrict ourselves to symmetric Gaussian pattern distributions in (1) and (2),

$$V^*(\lambda)=\frac{a}{2}\lambda^2\,, \qquad (30)$$

where a is a parameter satisfying $-1<a<\infty$. The appeal of this distribution is that, while being nonuniform for $a\neq 0$, the analytic calculations are not more difficult than those for the uniform $a=0$ distribution. Furthermore, the parameter a provides an interesting control parameter. For $a>0$ the patterns are more densely distributed around the "equator" orthogonal to $\boldsymbol{B}$, while for $a<0$ this zone is less densely populated. Put differently, $a>0$ corresponds to a teacher who preferably presents difficult examples to the student, $a=0$ represents the practical man who explains to his fellow whatever happens to occur, and the $a<0$ trainer tries to bring out the "essential point" of his knowledge by means of simple examples. In Ref. [6] a related problem has been studied, namely, the case that $V^*(\lambda)$ vanishes for $|\lambda| \geq K$ and is infinity otherwise, where $K \geq 0$ is a parameter.

In order to find $R_B(\alpha)$ we will first determine the overlap $R_G(\alpha)$ for Gibbs learning, cf. Eq. (5), which can be obtained from the extremization problem (13) and (14) by setting $\beta=1$, $q=R$, and $V(\lambda)=V^*(\lambda)$ and replacing D^*t by $D^{**}t=P^{**}(t)dt$ according to (8). By means of straightforward standard manipulations [3,13] the condition $dG(q=R,R)/dR=0$ then yields the following implicit equation for $R=R_G(\alpha)$:

$$1+a(1-R)=\alpha\left[\frac{a^2(1-R)}{1+a}+\frac{2}{R}\int Dz\, U_2(\gamma z)\right], \tag{31}$$

$$\gamma:=-\left[\frac{R}{(1-R)(1+a)}\right]^{1/2}, \quad U_i(z):=\frac{[e^{-z^2/2}]^i}{(2\pi)^{i/2}\int_z^\infty Dt}. \tag{32}$$

For unstructured patterns, $a=0$, the result from Ref. [3] is recovered. In Fig. 1, we plot $R_B(\alpha)$ for Bayes learning with different parameters a following from (31) and (5). Finally, the stability condition (18) can be rewritten in the form $\Gamma(R_G(\alpha))>0$, where the stability parameter is defined as

$$\Gamma(R):=1-\frac{R}{1+a(1-R)} \times\frac{a^2(1-R)^2+2\int Dz\,\hat{U}(\gamma z)}{[a^2/(1+a)]R(1-R)+2\int Dz\, U_2(\gamma z)}, \tag{33}$$

$$\hat{U}(z):=U_2(z)\{2a(1-R)+[U_1(z)-z]^2\}. \tag{34}$$

We find that the replica symmetric solution is indeed stable, cf. Fig. 2.

From (5) and (31) one finds the following asymptotic expressions:

$$R_B(\alpha)=\left[\frac{2}{\pi}\frac{\alpha}{1+a}\right]^{1/2}+O(\alpha^2), \tag{35}$$

$$R_B(\alpha)=1-\frac{1}{\alpha^2(1+a)}\frac{\pi}{4[\int Dz\, U_1(z)]^2}+O(\alpha^{-3}). \tag{36}$$

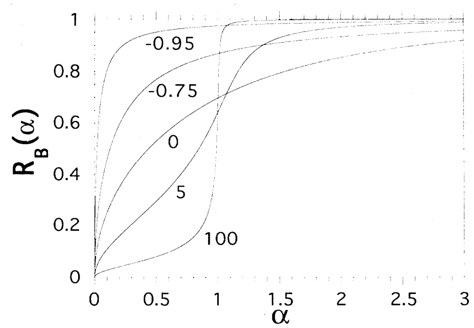

FIG. 1. The overlap $R_B(\alpha)$ following from (5) and (31) for supervised Bayes learning from Gaussian distributed examples (1) and (30) with different parameter values a between -0.95 and 100. The trainer is a teacher perceptron providing classifications of the examples according to (7) with $\Theta(f(\lambda))=\Theta(\lambda)$.

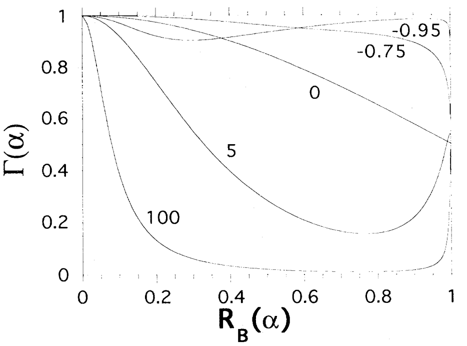

FIG. 2. The stability parameter $\Gamma(R)$ from (33) for the same values of the parameter a as in Fig. 1.

For $a=0$, one recovers the well known results for uniformly distributed patterns, while for $a\neq0$ the results are in agreement with known bounds and asymptotic results [24,25]. Furthermore, we find, as in Ref. [6], that an ideal student learns best on the basis of simple examples (negative a) as a beginner (α small) but in order to become an expert for large α a "tough" teacher (large positive a) is more favorable.

The following two limiting cases are also worth mentioning. First, one can see that

$$R_B(\alpha)=\Theta(\alpha) \quad \text{for } a\rightarrow-1. \tag{37}$$

Since in this limit the teacher only presents patterns $\{\xi^\mu\}^p$ parallel to $\boldsymbol{B}$, it is obvious that the $\boldsymbol{J}$ vector of a Hebb student, and *a fortiori* of an "ideal" student, will point in the direction of the teacher right away. Second, one has that

$$R_B(\alpha)=\Theta(\alpha-1) \quad \text{for } a\rightarrow\infty. \tag{38}$$

For $\alpha<1$ the "ideal" student, and hence any student, does not learn at all, while for $\alpha>1$ perfect learning is achieved. In this limit all examples ξ^μ lie exactly on the equator perpendicular to the unknown $\boldsymbol{B}$. Since for $p<N$ the examples are linearly independent with probability 1, they define an $(N-p)$-dimensional hypothesis subspace of possible $\boldsymbol{B}$'s. It is not difficult to see that this implies $R_B=0$ for $p<N$ and $R_B=1$ for $p\geq N$.

VI. COMPLETE SOLUTION FOR QUADRATIC *ad hoc* POTENTIALS

In this section we give a complete solution for the storage, supervised, and unsupervised problems with arbitrary pattern distributions (1) in the case that the learning strategy is based on the minimization of a cost function (6) with a quadratic *ad hoc* potential

$$V(\lambda)=\frac{c}{2}\lambda^2-d\lambda. \tag{39}$$

The quadratic potential (39) includes several commonly used learning rules, such as the adaline ($c>0$, see Sec. 2.4 in [7] and Secs. III B and III C in Ref. [13]), the Hebb rule ($c=0$, see Sec. III A in [13]) or Hopfield rule (see Sec. 4.5 in [7]), and the maximal variance principle ($c<0$, $d=0$, see Refs. [4,5]). Similar potentials have been studied previously for both unsupervised [4,5,26,27] and supervised [13,20] learning. They may also be of interest for the capacity problem above α_c [7,15,18].

We now turn to the evaluation of the free energy (13). The logarithm appearing in (14) can be evaluated explicitly, yielding an integrand that is a quadratic form in t and z. Therefore the integral involving $D^*t=dt\,P^{*(*)}(t)$, and hence the free energy, only depends on the distribution $P^{*(*)}$ *through its first two moments:*

$$\bar{\lambda}:=\int d\lambda\,\lambda P^{*(*)}(\lambda)\ , \tag{40}$$

$$A:=1-\int d\lambda(\lambda-\bar{\lambda})^2 P^{*(*)}(\lambda)\ , \tag{41}$$

where $P^{*(*)}:=P^*(\lambda)$ for unsupervised learning and $P^{*(*)}:=P^{**}(\lambda)$ for the supervised and the capacity problem, cf. (8), (9), and (28). The order parameters $q(\alpha)$ and $R(\alpha)$ are obtained from the extremization problem (13) and (14) by using (39)–(41) and letting $\beta\to\infty$. Several cases have to be distinguished.

We begin with the case $c=0$ (Hebb rule) for which one obtains $q(\alpha)=\Theta(\alpha)$ and

$$R(\alpha)=\sqrt{\alpha\bar{\lambda}^2/(1+\alpha\bar{\lambda}^2)}\ . \tag{42}$$

The same result applies for $d\to\infty$ and arbitrary but fixed c. Note that nothing at all is learned in the symmetric limit $\bar{\lambda}\to 0$.

For $c\neq 0$ and finite d the minimizing $\boldsymbol{J}$ in the cost function (6) only depends on the ratio

$$D:=d/c \tag{43}$$

and the sign of c. For $c<0$ one obtains again $q(\alpha)=\Theta(\alpha)$, while $R(\alpha)$ is given by the unique solution of

$$\alpha\left[A-\bar{\lambda}^2+\frac{\bar{\lambda}D}{R}\right]^2=\frac{1-AR^2+(D-\bar{\lambda}R)^2}{1-R^2}\ ,$$

$$0\le R\le R_0\ , \tag{44}$$

$$R_0:=\begin{cases}\dfrac{\bar{\lambda}D}{\bar{\lambda}^2-A} & \text{if } 0<\dfrac{\bar{\lambda}D}{\bar{\lambda}^2-A}\le 1\\ 1 \text{ otherwise}\ . & \end{cases} \tag{45}$$

For $c>0$ there exists a positive α_c below which $q(\alpha)$ is smaller than 1. In this regime $\alpha\le\alpha_c$ one finds that

$$R(\alpha)=\alpha\frac{\bar{\lambda}D}{1+\alpha(\bar{\lambda}^2-A)}\ , \tag{46}$$

$$q(\alpha)=R(\alpha)\frac{D}{\bar{\lambda}}\frac{1-\alpha A}{1-\alpha}\ . \tag{47}$$

The unique solution of $q(\alpha)=1$ in the domain $0\le\alpha\le 1$ readily follows from (47) and can be identified with α_c. Above α_c we have $q(\alpha)=1$ as expected and $R(\alpha)$ is again given by the unique solution of (44). At $\alpha=\alpha_c<1$, both $q(\alpha)$ and $R(\alpha)$ are continuous but nondifferentiable (see Fig. 3).

A more detailed derivation of (42)–(47) will be presented elsewhere. Before proceeding with the discussion of

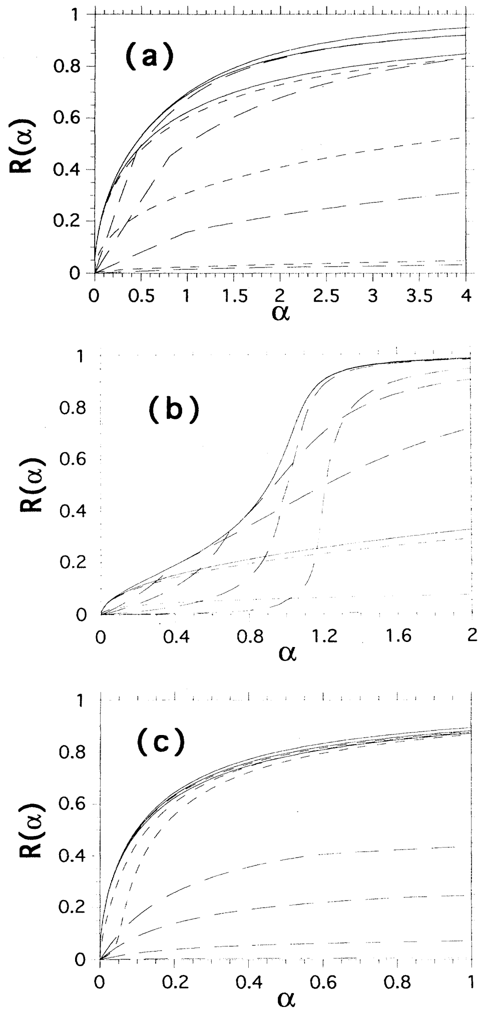

FIG. 3. The overlaps $R(\alpha)$ for supervised learning of a teacher perceptron with Gaussian distributed examples (1) and (30) with different parameter values $a=0$ (a), 10 (b), and -0.8 (c). Solid lines from top: the overlap $R_B(\alpha)$ for Bayes learning (5) and (31), $R_{\text{opt}}(\alpha)$ from (50)–(52) [hardly distinguishable from $R_B(\alpha)$ in (b)], and the performance (42) of the Hebb rule. Results $R(\alpha)$ from the minimization of the cost function (6) with quadratic potentials (39) at different parameter values $D=d/c$ are plotted as dashed lines. Long dashes from top: $D=1.3$, 0.7, 0.2, 0.02. Note the nondifferentiabilities in these lines [best visible in (a)]. Short dashes from top: $D=-10$, -0.5, -0.03. For both $D\to\infty$ and $D\to-\infty$ the Hebb rule is approached. $R_{\text{opt}}(\alpha)$ is the envelope of all the $R(\alpha)$ curves when the parameter D runs from $-\infty$ to ∞ but not every such $R(\alpha)$ curve actually touches $R_{\text{opt}}(\alpha)$. The $D=0.02$ curve in (b) and the $d=-0.03$ curve in (c) announce the occurrence of retarded classification at $D=0$, cf. (53).

these results we mention that the stability condition (18) turns out to be always fulfilled and that here and in the following we tacitly restrict ourselves to $\bar{\lambda} d \geq 0$ [the transformation $\bar{\lambda} d \mapsto -\bar{\lambda} d$ merely changes the sign of $R(\alpha)$].

For small α one can infer from (46) and (44) that $R(\alpha)$ starts off proportional to α when $c>0$ and proportional to $\sqrt{\alpha}$ when $c<0$ but still worse than for the Hebb rule $c=0$, cf. (42). For large α the overlap $R(\alpha)$ approaches 1 as $1/\alpha$ except when $c<0$ and $\bar{\lambda} D \leq \bar{\lambda}^2 - A$, in which case a convergence slower than $1/\alpha$ towards the asymptotic value $R_0 \leq 1$ is observed.

For symmetric distributions $P^{*(*)}(\lambda)$, i.e., $\bar{\lambda} \to 0$, as well as for symmetric potentials $d \to 0$ (with $c \neq 0$) we find from (44) and (46) that

$$R(\alpha)=\Theta(c[A-\bar{\lambda}^2])\Theta(\alpha-\alpha_0)\left[\frac{\alpha-\alpha_0}{\alpha-(A-\bar{\lambda}^2)^{-1}}\right]^{1/2}, \tag{48}$$

$$\alpha_0 := (1+D^2)/(A-\bar{\lambda}^2)^2 . \tag{49}$$

This phenomenon that $R(\alpha)$ remains zero below a threshold $\alpha_0 > 0$ is called "retarded classification" in [5] and will be discussed in more detail in Sec. VII. In particular, Eq. (48) implies that for $c[A-\bar{\lambda}^2] \leq 0$ nothing at all can be learned if either $\bar{\lambda} \to 0$ or $d \to 0$.

One may ask for the optimal choice of the parameters c and d in (39) for given values of α, $\bar{\lambda}$, and A, and the corresponding largest value of $R(\alpha)$. This optimal $R = R_{\text{opt}}(\alpha)$ is obtained as the unique solution of

$$\alpha\left[\bar{\lambda}^2+\frac{A^2R^2}{1-AR^2}\right]=\frac{R^2}{1-R^2}, \quad 0 \leq R \leq 1 , \tag{50}$$

and the corresponding optimal parameters $c = c_{\text{opt}}(\alpha)$ and $d = d_{\text{opt}}(\alpha)$ satisfy

$$\frac{c_{\text{opt}}(\alpha)}{d_{\text{opt}}(\alpha)}=\frac{A}{\bar{\lambda}}R_{\text{opt}}(\alpha) . \tag{51}$$

To illustrate further the above results, we finally turn to the Gaussian pattern distributions introduced in the previous section in the context of supervised learning from a teacher perceptron (see also Fig. 3). In this case one obtains from (40) and (41) that

$$\bar{\lambda}=\left[\frac{2}{\pi(1+a)}\right]^{1/2}, \quad A=1-\frac{1-2/\pi}{1+a} . \tag{52}$$

Thus $\bar{\lambda}$ is decreasing with $a \in [-1,\infty]$ from ∞ to 0, while A increases from $-\infty$ to 1. For small α a comparison of (35) with (42) and (52) shows that the Hebb rule saturates the Bayes limit. On the other hand, the large-α behavior of $R(\alpha)$ is always significantly worse than for the Bayes rule (36) and $R(\alpha)$ even stays below 1 for $\alpha \to \infty$ when $D < -\pi a/\sqrt{2(1+a)}$. The latter phenomenon has also been observed in [4]. For a symmetric potential $d \to 0$ Eq. (48) takes the form

$$R(\alpha)=\Theta(ac)\Theta(\alpha-\alpha_0)\left[\frac{\alpha-\alpha_0}{\alpha-(1+1/a)}\right]^{1/2}, \tag{53}$$

where $\alpha_0=(1+1/a)^2$. In particular, *maximal variance learning* ($c<0$) *performs quite well for* $a<0$ *and large* α, *saturates even the Bayes limit (38) for* $a \to -1$ *and arbitrary* α, *but fails completely for* $a \geq 0$ *and for* $a<0$ *when* $\alpha \leq \alpha_0$ (see also [4,13]). Note that for $a \to -1$ the Bayes limit (37) can also be saturated but now by a potential with $c<0$ and $d=0$. Some of these results for the particular case of uniformly distributed patterns $a=0$ have been obtained recently also in Ref. [13].

Finally, we turn to the optimal choice of the parameters of the *ad hoc* potential, as given by (51). The overlaps $R(\alpha)$ from (44) and (46), $R_B(\alpha)$ from (31) and (5), and $R_{\text{opt}}(\alpha)$ from (50) are shown in Fig. 3. As expected from the discussion below Eq. (52), the Bayes limit $R_B(\alpha)$ is reached by $R_{\text{opt}}(\alpha)$ for asymptotically small α with $c_{\text{opt}}(\alpha) \to 0$, whereas for large α one obtains a convergence towards 1 like $1/\alpha$ which is worse than for the Bayes rule (36). For intermediate α values the differences with the Bayes rule become remarkably small both for $a=10$ [Fig. 3(b)] and $a=-0.18$ [Fig. 3(c)].

VII. PERSPECTIVES

In the present paper we introduced a common framework for unsupervised learning, supervised learning, and the capacity problem within the context of nonuniform pattern distributions with a single symmetry-breaking orientation. In a companion publication, further applications of this formalism will be presented, including a general study of Gibbs and Bayes learning. We will show there that the overlap $R_B(\alpha)$ for Bayes learning vanishes for α values below a certain threshold $\alpha_0 > 0$ if and only if the first moment $\bar{\lambda}$ of the relevant pattern distribution [cf. (40)] is zero. The fact that nothing about the pattern structure can be inferred below a nonzero threshold α_0 has been observed previously in various special cases (see [4,5,27–29] and Sec. VI) and has been termed "retarded classification" in Ref. [5]. Note that for learning rules other than Bayes the threshold α_0 may be larger than for Bayes learning [4] and retarded classification may occur even when $\bar{\lambda} \neq 0$. For instance, we have seen in Sec. VI that a quadratic *ad hoc* potential $V(\lambda)=c\lambda^2/2$ leads to retardation for *any* pattern distribution (1) and it can be proven that the same is true for arbitrary symmetric potentials $V(-\lambda)=V(\lambda)$.

A special case of retarded classification follows from our investigation of the capacity problem in Sec. IV. In this case, we found that $R(\alpha)=0$ at $\alpha=\alpha_c$, where $q(\alpha)$ becomes 1 for the first time, if and only if $\bar{\lambda}=0$; see Eq. (29). Even though we were not able to prove this rigorously, we expect that $R(\alpha)$ vanishes not only at $\alpha=\alpha_c$ but for all $\alpha \leq \alpha_c$. In other words, we have

$$\alpha_c \leq \alpha_0 \tag{54}$$

for such a pattern distribution when the *ad hoc* potential (10) is used. The same relation (54) follows from the results of Sec. VI for arbitrary pattern distributions and quadratic *ad hoc* potentials (39) when retarded classification is present, i.e., $\bar{\lambda}=0$ or $V(-\lambda)=V(\lambda)$. Finally, retarded classification with $\alpha_c=\alpha_0=1$ was also observed for the Ising reversed wedge perceptron for a specific width of the wedge corresponding precisely to the condition $\bar{\lambda}=0$ [30]. We therefore conjecture that (54) is valid under rather general conditions within the context of *ad hoc* potentials leading to retarded classification. [Note that for Gibbs learning (54) is generally violated since $q(\alpha)=R(\alpha)$.] This implies a surprising connection between our result (26)–(29) for the capacity problem and the phenomenon of retarded classification in the corresponding unsupervised problem. Moreover, one expects that for more general potentials [2,15] $V(\lambda)=\Theta(\lambda-\kappa)$ than in (10) one will obtain similar "universal" α_c values which only depend on the "stability parameter" $\kappa\geq 0$ when $\bar{\lambda}=0$.

Throughout our investigation we tacitly assumed that the distribution of the patterns (1) was exactly known except for the symmetry-breaking orientation $\boldsymbol{B}$ itself. From the results obtained in Sec. VI we can conclude that this knowledge is indeed practically indispensable in order to design a learning strategy that provides a reasonably good hypothesis $\boldsymbol{J}$ for the unknown $\boldsymbol{B}$. However, in particular in the context of unsupervised learning, one would like to include cases with little or no *a priori* knowledge about the pattern distribution. In a subsequent work we will show how the case of a pattern distribution (1) for which *both* $\boldsymbol{B}$ and $V^*(\lambda)$ are unknown can be reduced to the one studied in the present paper. In other words, we will describe a method that allows the exact determination of $V^*(\lambda)$ from the presented patterns provided it is known that they are axially symmetric about a single (unknown) direction $\boldsymbol{B}$. The only and obviously unavoidable exceptions are pattern distributions with $\bar{\lambda}=0$ below the retardation threshold α_0 belonging to Bayes learning.

ACKNOWLEDGMENTS

Discussions with E. Lootens and M. Bouten are gratefully acknowledged. We thank the Program on Inter-University Attraction Poles, Prime Minister's Office, Belgian Government for financial support. C.V.d.B. also acknowledges support from the NFWO Belgium.

APPENDIX A

Following Ref. [5] we define the quality of a hypothesis $\boldsymbol{J}$ as

$$Q(\boldsymbol{J}):=\int d\boldsymbol{B}'P(\boldsymbol{B}'|\{\xi^\mu\}^p)h(\boldsymbol{J}\cdot\boldsymbol{B}'/N)\ , \qquad \text{(A1)}$$

where $h(x)$ is strictly monotonically increasing for $-1\leq x\leq 1$ but otherwise arbitrary. As proven in [5], for any set of patterns $\{\xi^\mu\}^p$ the Bayes hypothesis $\boldsymbol{J}_B$ from (4) maximizes the quality (A1), independently of the specific choice of $h(x)$, and the overlaps $R_B(\alpha)=\boldsymbol{J}_B\cdot\boldsymbol{B}/N$ and $R_G(\alpha)=\boldsymbol{J}_G\cdot\boldsymbol{B}/N$ of the Bayes and Gibbs hypotheses with the true (unknown) $\boldsymbol{B}$ satisfy (5). These results are valid in the limit $N\to\infty$ with $\alpha:=p/N\geq 0$ fixed. Moreover, the pattern distribution (1) must be such that the self-overlaps $q(\alpha)=\boldsymbol{J}^a\cdot\boldsymbol{J}^b/N$ of two hypotheses drawn independently from (3) are distributed with a very sharp peak. This condition is equivalent to the *assumption of replica symmetry in the Gibbs learning scenario.* Taking into account that $q(\alpha)=R_G(\alpha)$ in this scenario, one expects [31] that replica symmetry is actually never broken. The latter can even be proven rigorously for asymptotically small and large α and arbitrary pattern distributions (1), as we will show elsewhere. Moreover, it was found true for arbitrary α in all special cases so far studied.

Next we restrict ourselves to *learning algorithms for which the overlap $R(\alpha)=\boldsymbol{J}\cdot\boldsymbol{B}/N$ is self-averaging* (i.e., δ distributed) in the limit $N\to\infty$. Note that this assumption is rather weak and, in particular, does not require replica symmetry to hold. We want to prove that $\boldsymbol{J}_B$ makes the smallest angle with $\boldsymbol{B}$ among all the hypotheses $\boldsymbol{J}$ following from these learning algorithms. To this end it is sufficient to demonstrate that $Q(\boldsymbol{J}_B)-Q(\boldsymbol{J})\geq 0$ implies $R_B(\alpha)-R(\alpha)\geq 0$. By multiplying $Q(\boldsymbol{J}_B)-Q(\boldsymbol{J})\geq 0$ by $\prod_{\mu=1}^p\delta((\xi^\mu)^2-N)$, integrating over all ξ^μ, and comparing (1) with (3) one finds from (A1) that

$$\int d\boldsymbol{B}'\delta(\boldsymbol{B}'^2-N)\int\prod_{\mu=1}^{p}d\xi^\mu P(\{\xi^\mu\}^p|\boldsymbol{B}')[\boldsymbol{J}_B-\boldsymbol{J}]\cdot\boldsymbol{B}'/N \geq 0\ , \qquad \text{(A2)}$$

where we made the particular choice $h(x)=x$ in (A1). Though $\boldsymbol{J}_B$ and $\boldsymbol{J}$ depend on $\{\xi^\mu\}^p$ and the learning algorithm used, they are independent of the integration variable $\boldsymbol{B}'$ in (A2) and we thus can interpret any fixed $\boldsymbol{B}'$ in the integral $\int\prod_{\mu=1}^p d\xi^\mu\cdots$ as the real (unknown) symmetry-breaking orientation $\boldsymbol{B}$. Using the self-averaging assumption for the overlaps, the integral $\int\prod_{\mu=1}^p d\xi^\mu\cdots$ becomes $R_B(\alpha)-R(\alpha)$ independently of $\boldsymbol{B}'$ and thus $R_B(\alpha)\geq R(\alpha)$.

APPENDIX B

Equation (25) can be rewritten in the form $\psi(r)=0$, where $r:=R/\sqrt{1-R^2}$ and

$$\psi(x):=\int D^{**}t\left[te^{-x^2t^2}/\sqrt{2\pi}-xt^2\int_{xt}^{\infty}Dz\right]. \qquad \text{(B1)}$$

One readily sees that $\psi(0)=\bar{\lambda}/\sqrt{2\pi}$ [cf. (28)] and $\psi'(x)<x$ for all x. It follows that the solution of (25) is unique. Moreover, we can conclude that $r=0$ if $\bar{\lambda}=0$, $r>0$ if $\bar{\lambda}>0$, and $r<0$ if $\bar{\lambda}<0$. Using (25), a partial integration of (23) yields $\alpha_c\phi(r)=1$, where $\phi(x):=\int D^{**}t\int_{xt}^{\infty}Dz$. Next we observe that $\phi(0)=\frac{1}{2}$ and

$$\phi'(x)=-\psi(x)-x\int D^{**}tt^2\int_{xt}^{\infty}Dz$$

and consequently

$$\frac{1}{2}-\int_0^r dx\left[\psi(x)+x\int D^{**}tt^2\int_{xt}^{\infty}Dz\right]=\frac{1}{\alpha_c}\ . \qquad \text{(B2)}$$

For $\bar{\lambda}=0$ we have seen that $r=0$ and thus $\alpha_c=2$. For $\bar{\lambda}>0$ we found that $r>0$ and that $\psi(x)$ decreases from $\bar{\lambda}/\sqrt{2\pi}$ to 0 when x increases from 0 to r. Consequently, we obtain $\alpha_c>2$ for $\bar{\lambda}>0$ and similarly for $\bar{\lambda}<0$.

[1] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987); K. Binder and A. P. Young, Rev. Mod. Phys. **58**, 801 (1986).

[2] E. Gardner, J. Phys. A **21**, 247 (1988).

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

[4] M. Biehl and A. Mietzner, Europhys. Lett. **24**, 421 (1993); J. Phys. A **27**, 1885 (1994).

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

[6] I. Derényi, T. Geszti, and G. Györgyi, Phys. Rev. E **50**, 3192 (1994).

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

[8] B. López, M. Schröder, and M. Opper, J. Phys. A **28**, L447 (1995).

[9] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Reading, MA, 1991).

[10] Terms that vanish for $N \rightarrow \infty$ as well as an additive constant have been omitted in (14). We also recall that "extr" in (13) means that one has to perform *first* a minimization with respect to $q \in [R^2,1]$ for any fixed $R \in [-1,1]$ and then to determine the maximum with respect to $R \in [-1,1]$ (see, for instance, [6] and further references therein).

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

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

[13] M. Bouten, J. Schietse, and C. Van den Broeck, Phys. Rev. E **52**, 1958 (1995).

[14] E. Gardner, J. Phys. A **20**, 3453 (1987).

[15] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).

[16] J. R. L. de Almeida and D. J. Thouless, J. Phys. A **11**, 983 (1978).

[17] One actually has to *maximize* with respect to $q = r \in [0,1]$ since Gibbs learning requires a limit $n \rightarrow 1$ at a certain point of the replica calculation (n is the number of replicas), while in the usual limit $n \rightarrow 0$ the maximization over 9 turns into a minimization; see also [10].

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

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

[20] M. B. Gordon and D. R. Grempel, Europhys. Lett. **29**, 257 (1995).

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

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

[23] R. Monasson, J. Phys. A **25**, 3701 (1992).

[24] N. V. Vapnik and A. Y. Chernovenkis, Theory Probab. Appl. **16**, 264 (1971); J. M. R. Parrondo and C. Van den Broeck, J. Phys. A **26**, 2211 (1993).

[25] C. Van den Broeck and J. M. R. Parrondo, Phys. Rev. Lett. **71**, 2355 (1993).

[26] A. Mietzner, M. Opper, and W. Kinzel, J. Phys. A **28**, 2785 (1995).

[27] E. Lootens and C. Van den Broeck, Europhys. Lett. **30**, 381 (1995).

[28] D. Hansel, G. Mato, C. Meunier, Europhys. Lett. **20**, 471 (1992).

[29] L. Reimers and A. Engel (unpublished).

[30] G.-J. Bex, R. Serneels, and C. Van den Broeck, Phys. Rev. E **51**, 6309 (1995).

[31] A. Engel and L. Reimers, Europhys. Lett. **28**, 531 (1994).