A Gaussian scenario for unsupervised learning

1996 J. Phys. A: Math. Gen. 29 3521

(http://iopscience.iop.org/0305-4470/29/13/021)

J. Phys. A: Math. Gen. **29** (1996) 3521–3535. Printed in the UK

A Gaussian scenario for unsupervised learning

P Reimann†, C Van den Broeck‡ and G J Bex‡

† Eötvös University, Puskin-u 5-7, H-1088 Budapest, Hungary

‡ Limburgs Universitair Centrum, B-3590 Diepenbeek, Belgium

Received 4 December 1995

Abstract. We consider random patterns on the N-sphere which are uniformly distributed with the exception of a single symmetry-breaking orientation, along which they are Gaussian distributed. The unsupervised recognition of this orientation by different learning rules is studied in the large-N limit using the replica method. The model is simple enough to be analytically tractable and rich enough to exhibit most of the phenomena observed with other pattern distributions. A learning algorithm based on the minimization of a cost function is identified which reaches the upper theoretical limit imposed by the optimal (Bayes-) learning scenario. An implementation of this algorithm is proposed and tested numerically.

1. Introduction

The main objective in unsupervised learning is the detection of structure in a given set of data, typically a set of p N-dimensional vectors $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$ [1–6]. This can only be achieved provided a certain amount of *a priori* knowledge about the form of the probability distribution that generates the patterns $\boldsymbol{\xi}^\mu$ is available. In this paper we investigate a 'Gaussian scenario', in which this distribution has a Gaussian form. This simple case allows for a fully analytical treatment, while it surprisingly retains many of the characteristic features and phenomena observed for more complicated distributions [2, 3, 6]. More precisely, we assume that there exists a single (unknown) direction $\boldsymbol{B}$ such that those components of a pattern $\boldsymbol{\xi}^\mu$ orthogonal to $\boldsymbol{B}$ are independent normal Gaussian random variables, while for the component $\lambda = \boldsymbol{\xi}^\mu \cdot \boldsymbol{B}/||\boldsymbol{B}||$ parallel to $\boldsymbol{B}$ the Gaussian distribution has a non-zero mean and a modified dispersion:

$$P^*(\lambda) \propto \mathrm{e}^{-\lambda^2/2 - V^*(\lambda)} \tag{1}$$

$$V^*(\lambda) = a\,\lambda^2/2 - b\,\lambda \tag{2}$$

where proportionality $\propto$ accounts for an omitted normalization constant and the parameters a and b are assumed to satisfy $a > -1$ and $b \geqslant 0$. In other words, each data point $\boldsymbol{\xi}^\mu$ carries a single information-rich scalar λ, which, however, is buried in a large amount of random signals. We assume to know the values of a and b and our goal is to infer the unknown symmetry-breaking orientation $\boldsymbol{B}$ from the set of patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$.

The paper is organized as follows. In section 2, we briefly review the basic strategies to infer a hypothesis for $\boldsymbol{B}$ from the patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$ and their resulting performance based on a replica calculation. The so-called Gibbs and Bayes learning algorithms are discussed in detail in section 3, while learning based on the minimization of a specific (quadratic) cost function is worked out in section 4. In particular, a cost function is identified, the performance of which reaches the upper theoretical limit imposed by the optimal (Bayes-)

0305-4470/96/133521+15$19.50

learning scenario. A practical algorithm to minimize this cost function is discussed within our conclusions in section 5.

2. General framework

A widely used strategy to select a hypothesis $\boldsymbol{J}$ for $\boldsymbol{B}$ is to introduce a scalar cost function E that describes how well a specific vector $\boldsymbol{J}$ incorporates the information of the training set $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$. In the present problem, we note that each pattern $\boldsymbol{\xi}^\mu$ has, as a result of the central limit theorem, a length approaching $\sqrt{N}$ for large N. It is therefore convenient to normalize the length of the other N-dimensional vectors, such as $\boldsymbol{B}$ and $\boldsymbol{J}$, to be equal to $\sqrt{N}$. Because the training examples are assumed to be sampled independently of one another, it turns out that an additive cost function of the following form covers most interesting learning scenarios [2–6]:

$$E(\boldsymbol{J}) = \sum_{\mu=1}^{p} V(\boldsymbol{\xi}^\mu \cdot \boldsymbol{J}/\sqrt{N}) \tag{3}$$

under the side condition $\boldsymbol{J}^2 = N$ and with a properly chosen *ad hoc* potential $V(\lambda)$. Within the context of our 'Gaussian scenario' (cf equation (2)) we will concentrate on the following class of quadratic *ad hoc* potentials:

$$V(\lambda) = c\,\lambda^2/2 - d\lambda \tag{4}$$

where c and d are parameters. The adaline rule† is recovered for $c = 1$, maximal variance learning (principal component analysis) [2, 3] for $c = -2$, $d = 0$, and the Hebb (or Hopfield) rule for $c = 0$, $d = 1$ (see section 3.1 in [9] and section 4.5 in [8]).

In order to quantify the quality of a hypothesis $\boldsymbol{J}$ constructed on the basis of the cost function (3), (4) we consider the distribution of its overlap $R = \boldsymbol{J} \cdot \boldsymbol{B}/N$ with the unknown 'true' $\boldsymbol{B}$, assuming for the moment that $\boldsymbol{J}$ is selected from the Boltzmann-like ensemble with cost function E and 'inverse temperature' β:

$$\rho(R) \propto \int \mathrm{d}\boldsymbol{J}\, \mathrm{e}^{-\beta E(\boldsymbol{J})}\, \delta(\boldsymbol{J} \cdot \boldsymbol{B}/N - R)\, \delta(\boldsymbol{J}^2 - N)\,. \tag{5}$$

The description in terms of such an ensemble also includes several other learning algorithms to be discussed below. Assuming replica symmetry, a standard calculation [6] shows that in the limit $N \to \infty$ with $\alpha := p/N$ fixed, the distribution $\rho(R)$ is self-averaging with respect to the pattern distribution and approaches $\delta(R - R(\alpha))$. The location $R(\alpha)$ of the δ-peak follows from the extremization‡ of

$$G(q, R) = \frac{\ln x}{2} + \frac{\beta(1-R^2)}{2x} - \frac{\alpha \ln(1+cx)}{2} - \frac{\alpha\,\beta/2}{1+cx}\left[c - \left(d^2 + \frac{c^2}{\beta^2}\right)x - 2\,B\,d\,R + c(B^2 - A)\,R^2\right] \tag{6}$$

where we introduced

$$x := \beta(1-q) \qquad A := a/(1+a) \qquad B := b/(1+a)\,. \tag{7}$$

† Unlike in the original adaline algorithm [7], in our case the minimization of the cost function (3) is performed under the extra condition that $\boldsymbol{J}$ must be properly normalized, see also section 2.4 in [8] and sections 3.2, 3.3 in [9].

‡ By 'extremization' we mean a minimization with respect to $q \in [R^2, 1]$ under the side condition $1 + cx > 0$ for any $R \in [-1, 1]$, followed by a maximization with respect to $R \in [-1, 1]$.

Note that $A < 1$ and $B \geqslant 0$ due to $a > -1$ and $b \geqslant 0$. As usual, the extremizing $R = R(\alpha)$ is the typical overlap of $\boldsymbol{J}$ and $\boldsymbol{B}$, whereas the extremizing $q = q(\alpha)$ represents the typical self-overlap of $\boldsymbol{J}$-vectors from two different replicas. The local stability condition [10] for the replica-symmetric solution (6) takes the simple form [6]

$$1 > \alpha \left(\frac{c\, x(\alpha)}{1 + c\, x(\alpha)} \right)^2 \tag{8}$$

where $x(\alpha) := \beta(1 - q(\alpha))$. Note that the pattern distribution and the value of d only enter indirectly through $x(\alpha)$.

Following the general discussion presented in [6], we briefly review the various learning rules that can be obtained from (6). It is clear from (5) that the performance $R(\alpha)$ of a hypothesis vector $\boldsymbol{J}$ that minimizes the cost function (3), (4) follows by extremization of (6) in the limit $\beta \to \infty$. Further interesting 'learning rules' that can be studied by means of (6) are found by observing that, given the patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$, the *a posteriori* probability for the unknown $\boldsymbol{B}$ to coincide with a hypothesis $\boldsymbol{J}$ is given by†

$$P(\boldsymbol{J}|\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p) \propto \exp\left\{ -\sum_{\mu=1}^{p} V^*(\boldsymbol{\xi}^\mu \cdot \boldsymbol{J}/\sqrt{N}) \right\} \delta(\boldsymbol{J}^2 - N)\,. \tag{9}$$

By comparison of (9) with (3), (5) the performance $R_{\mathrm{M}}(\alpha)$ of the most probable *a posteriori* (or maximal likelihood) hypothesis $\boldsymbol{J}$ follows again by extremization of (6) in the limit $\beta \to \infty$ provided we choose $V(\lambda) = V^*(\lambda)$, i.e. $c = a$ and $d = b$. Next we note that the choice $\beta = 1$ and $V(\lambda) = V^*(\lambda)$ in (3), (5) means sampling at random a vector $\boldsymbol{J}$ according to the *a posteriori* probability (9). This strategy is known as Gibbs (or Boltzmann) learning and the corresponding performance $R_{\mathrm{G}}(\alpha)$ is again recovered by extremization of (6) with $c = a$, $d = b$ and $\beta = 1$. In this case, $\boldsymbol{B}$ and the different replicas of $\boldsymbol{J}$ play an equivalent role so that one finds $q = R$ and one is left with a maximization of $G(q = R, R)$ with respect to $R \in [0, 1]$ in order to find $R_{\mathrm{G}}(\alpha)$. Finally, the performance $R_{\mathrm{B}}(\alpha)$ of the so-called Bayes rule, corresponding to the best hypothesis that possibly can be inferred from the given patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$, is related to the Gibbs overlap $R_{\mathrm{G}}(\alpha)$ through

$$R_{\mathrm{B}}(\alpha) = \sqrt{R_{\mathrm{G}}(\alpha)} \tag{10}$$

according to a general argument given in [3]. Hence, this upper theoretical limit $R_{\mathrm{B}}(\alpha)$ for the performance of *any* learning rule can be determined within our general framework as well.

3. Gibbs and Bayes learning

As explained in the previous section, the overlap $R_{\mathrm{G}}(\alpha)$ for Gibbs learning follows from (6) by maximization over $R \in [0, 1]$ with $\beta = 1$, $V(\lambda) = V^*(\lambda)$, and $q = R$. One finds the following explicit result (cf appendix A):

$$R_{\mathrm{G}}(\alpha) = \frac{1 + \alpha[B^2(1 + A) - A^2] - \sqrt{Q_{\mathrm{G}}}}{2\,A[1 + \alpha(B^2 - A)]} \tag{11}$$

$$Q_{\mathrm{G}} := \left(1 - \alpha[B^2(1 - A) + A^2]\right)^2 + 4\alpha B^2(1 - A)\,. \tag{12}$$

† This relation (9) is a straightforward consequence of the so-called Bayes rule applied to the conditional distribution of the patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$ given $\boldsymbol{B}$ with a uniform prior on $\boldsymbol{B}$, see e.g. [3, 6].

For asymptotically small and large α this yields

$$R_{\mathrm{G}}(\alpha) = \alpha\, B^2 + \mathrm{O}(\alpha^2) \tag{13}$$

$$R_{\mathrm{G}}(\alpha) = 1 - \frac{1}{\alpha}\,\frac{(1+a)^2}{b^2 + a^2(1+a)} + \mathrm{O}(\alpha^{-2})\,. \tag{14}$$

We also mention the following particular cases and limits:

$$R_{\mathrm{G}}(\alpha) = \Theta(\alpha) \qquad \text{for} \quad a \to -1 \quad \text{or} \quad b \to \infty \tag{15}$$

$$R_{\mathrm{G}}(\alpha) = \frac{\alpha\, b^2}{1 + \alpha\, b^2} \qquad \text{for} \quad a = 0 \tag{16}$$

$$R_{\mathrm{G}}(\alpha) = \Theta(\alpha - \alpha_0)\,\frac{\alpha - \alpha_0}{\alpha - 1/A} \qquad \alpha_0 := 1/A^2 \quad \text{for} \quad b = 0 \tag{17}$$

$$R_{\mathrm{G}}(\alpha) = \begin{cases} \dfrac{\alpha B^2}{1 - \alpha + \alpha B^2} & \text{for } \alpha \leqslant 1 \\ 1 & \text{for } \alpha \geqslant 1 \end{cases} \qquad \text{for} \quad a \to \infty \quad b/a \to B = \text{constant}\,. \tag{18}$$

In the latter case, all patterns lie on a cone with fixed overlap $\lambda = B$ with the symmetry-breaking orientation $\boldsymbol{B}$. The Heaviside function is defined as $\Theta(x) = 1$ for $x > 0$ and $= 0$ for $x \leqslant 0$. The local stability condition (8) turns out to be always satisfied with the exception of $b = 0$ and $\alpha = \alpha_0$, where a marginally stable situation is encountered.

The overlap $R_{\mathrm{B}}(\alpha)$ for Bayes learning follows from (10), (11) and is shown for a few representative values of a and b in figure 1. It starts off like $\sqrt{\alpha}$ for small α, provided there is a bias in the pattern distribution, $b > 0$. For $b = 0$ the phenomenon of 'retarded classification' [3] is observed: there exists a threshold value α_0 (cf equation (17)), below which the structure underlying the patterns $\{\boldsymbol{\xi}^{\mu}\}_{\mu=1}^{p}$ cannot be recognized by the Bayes and

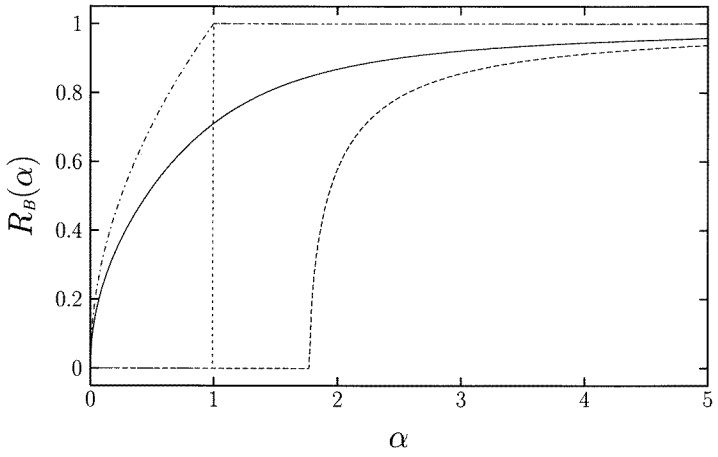

Figure 1. The overlap $R_{\mathrm{B}}(\alpha)$ for Bayes learning according to (10), (11) at four different parameter values a and b. The full curve ($a = b = 3$) illustrates the 'typical' behaviour, the broken curve ($a = 3$, $b = 0$) exemplifies retarded classification ending with a second-order transition at $\alpha = \alpha_0 \simeq 1.78$. The chain curve represents a case where all patterns lie on a cone ($a \to \infty$ with $b/a = 1$) leading to perfect learning for $\alpha \geqslant 1$. When all patterns are perpendicular to $\boldsymbol{B}$ (dotted line, $a \to \infty$, $b/a \to 0$) retarded classification goes over to perfect learning at $\alpha = 1$ through a first-order transition.

thus by any learning rule, $R_B(\alpha) = 0$ for $\alpha \leqslant \alpha_0$, see figure 1. The same phenomenon has also been observed for more complicated non-uniform distributions, using *ad hoc* potentials [2, 4], or the Gibbs and Bayes rules [3]. The closely related effect of 'retarded generalization' in supervised learning has been discussed in [11].

Once $R_B(\alpha)$ is non-zero it stays monotonically increasing, and typically approaches 1 like α^{-1} for large α. Furthermore, $R_B(\alpha)$ is monotonically increasing as a function of b for any fixed value of a and $\alpha > 0$. Thus, at least in the optimal case of Bayes learning, the learnability of the symmetry-breaking orientation increases with the number of presented patterns as well as with their bias. For $b = 0$ one finds that $\forall \alpha > \alpha_0$ the overlap $R_B(\alpha)$ increases with increasing parameter a when $a \geqslant 0$ and decreases with increasing a when $-1 < a < 0$. In other words, without bias ($b = 0$) the Bayes rule learns the fastest for $a \to -1$ and $a \to \infty$, namely as $R_B(\alpha) = \Theta(\alpha)$ and $R_B(\alpha) = \Theta(\alpha - 1)$, respectively, and performs worse as a approaches 0 from both the negative and positive side, with the obvious worst case $R_B(\alpha) \equiv 0$ for uniformly distributed patterns ($a = b = 0$).

The result $R_B(\alpha) = \Theta(\alpha)$, found for $a \to -1$ or $b \to \infty$, can be understood from the fact that the typical overlaps of the patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$ with the symmetry-breaking orientation $\boldsymbol{B}$ become very large in these limits. It is therefore plausible that $\boldsymbol{B}$ can be inferred from a number $p = o(N)$ of patterns, see equation (15). The result $R_B(\alpha) = \Theta(\alpha - 1)$ for $a \to \infty$, $b/a \to 0$ follows from the fact that, in this limit, all patterns $\boldsymbol{\xi}^\mu$ lie exactly on the big circle perpendicular to the unknown $\boldsymbol{B}$. For $p < N$, they are linearly independent with probability 1 and define a $(N - p)$-dimensional subspace of equally probable hypotheses $\boldsymbol{J}$. It is not difficult to see that this implies $R_B = 0$ for $p < N$ and $R_B = 1$ for $p \geqslant N$. By similar arguments one can understand the more general result (18).

4. Quadratic *ad hoc* potentials

We recall that the overlap $R(\alpha)$ corresponding to the minimization of the cost function (3), (4) follows by letting $\beta \to \infty$ in (6). Obviously, we can restrict ourselves to $d \geqslant 0$ since $d \mapsto -d$ merely changes the sign of $R(\alpha)$. We first reproduce the bare results, with the identification of several subcases, and afterwards turn to a more detailed discussion.

4.1. Results

For $c = 0$ (and $d > 0$) the extremizing overlaps in (6) are readily found to be

$$q(\alpha) = \Theta(\alpha) \tag{19}$$

$$R(\alpha) = \sqrt{\alpha\, B^2/(1 + \alpha\, B^2)}\,. \tag{20}$$

Note that the Hebb rule ($d = 1$) follows as a special case and that the value of d is actually irrelevant (apart from the trivial case $d = 0$).

For $c \neq 0$, the results do not depend on c and d separately, but only on their ratio

$$D := d/c \tag{21}$$

and the sign of c. One has to distinguish two cases, corresponding to whether $q = q(\alpha)$ in (6) stays below 1 or converges to 1 for $\beta \to \infty$. The details of the calculations are given in appendix B. For $c > 0$ and $\alpha < \alpha_c < 1$, with

$$\alpha_c = \frac{1 + D^2 + A - B^2 - \sqrt{Q_c}}{2\,[A D^2 + A - B^2]} \tag{22}$$

$$Q_c := [1 - D^2 - A + B^2]^2 + 4\, D^2 (1 - A) \tag{23}$$

it is found that $q = q(\alpha)$ converges to a value < 1 and one obtains

$$q(\alpha) = R(\alpha)\,\frac{D}{B}\,\frac{1-\alpha A}{1-\alpha} \tag{24}$$

$$R(\alpha) = \alpha\,\frac{B\,D}{1+\alpha(B^2-A)}\,. \tag{25}$$

The fact that $q(\alpha) < 1$ below α_c can be understood as follows. For $c > 0$ minimization of the cost function (3), (4) is equivalent to that of $\sum_{\mu=1}^{p}(\boldsymbol{J}\cdot\boldsymbol{\xi}^{\mu}/\sqrt{N} - d/c)^2$. For $d = 0$ and $p < N$ there is a whole set of properly normalized $\boldsymbol{J}$'s for which this sum takes its absolute minimal value 0 and therefore $q(\alpha) < 1$ for $\alpha \in [0, 1[$. When $d \neq 0$ the size of this set is expected to be smaller, and a solution $q(\alpha) < 1$ will appear in a smaller α-interval.

In the remaining cases, i.e. $\alpha > 0$ when $c < 0$ or $\alpha \geqslant \alpha_c$ when $c > 0$, one finds that $q(\alpha) \to 1$ for $\beta \to \infty$, and $R(\alpha)$ follows as the unique solution of the fourth-order equation:

$$\alpha\left(A - B^2 + \frac{B\,D}{R}\right)^2 = \frac{F(R)}{1-R^2} \qquad 0 \leqslant R \leqslant R_0 \tag{26}$$

$$F(R) := 1 - A\,R^2 + (D - B\,R)^2 \tag{27}$$

$$R_0 := \begin{cases} \dfrac{B\,D}{B^2 - A} & \text{if } 0 < \frac{B\,D}{B^2-A} < 1 \\ 1 & \text{otherwise}\,. \end{cases} \tag{28}$$

4.2. Discussion

We now turn to the discussion of the results from the previous subsection. For illustrations see figure 2. We first note that $R(\alpha)$ and $q(\alpha)$ are non-decreasing functions of α and $R(\alpha)$ is even strictly increasing apart from specific cases of α-domains with a constant $R(\alpha)$. Further, $q(\alpha)$ and $R(\alpha)$ depend smoothly on α with the exception of $\alpha = \alpha_c$ when $c > 0$, where they are continuous but non-differentiable (see figures 2(*a*), (*b*), and of $\alpha = 0$ when $c \leqslant 0$, where $q(\alpha)$ jumps from 0 to 1 (see figures 2(*c*), (*d*). A surprising phenomenon may occur for $c > 0$ in the domain $\alpha < \alpha_c$ which, to the best of our knowledge, is observed here for the first time: *for $B > D$ and sufficiently small α, equation (24) implies that the overlap $R(\alpha)$ exceeds the self-overlap $q(\alpha)$!* (However, $q(\alpha) \geqslant R(\alpha)^2$ is still fulfilled.) An example can be seen in figure 2(*b*). This somewhat counterintuitive phenomenon does not seem to call for far reaching physical consequences. Rather it appears to be yet another curiosity that may occur in a space of high dimensionality N. Though it has never been noticed previously we expect it to also occur beyond the 'Gaussian scenario' considered here.

For small α the overlap $R(\alpha)$ starts off proportional to α when $c > 0$, cf (25), and proportional to $\sqrt{\alpha}$ when $c \leqslant 0$. The explicit behaviour for $c < 0$ is obtained from (26):

$$R(\alpha) = \sqrt{\alpha\,\frac{B^2 D^2}{1+D^2}} + \mathrm{O}(\alpha)\,. \tag{29}$$

The coefficients in these asymptotic expressions for small α are non-zero only for $B\,D \neq 0$ ($B\,d \neq 0$ if $c = 0$). For $B\,D = 0$, including symmetric pattern distributions ($b = 0$) [3] or symmetric potentials ($d = 0$), but also in the limits $a \to \infty$ or $c \to \pm\infty$, retarded

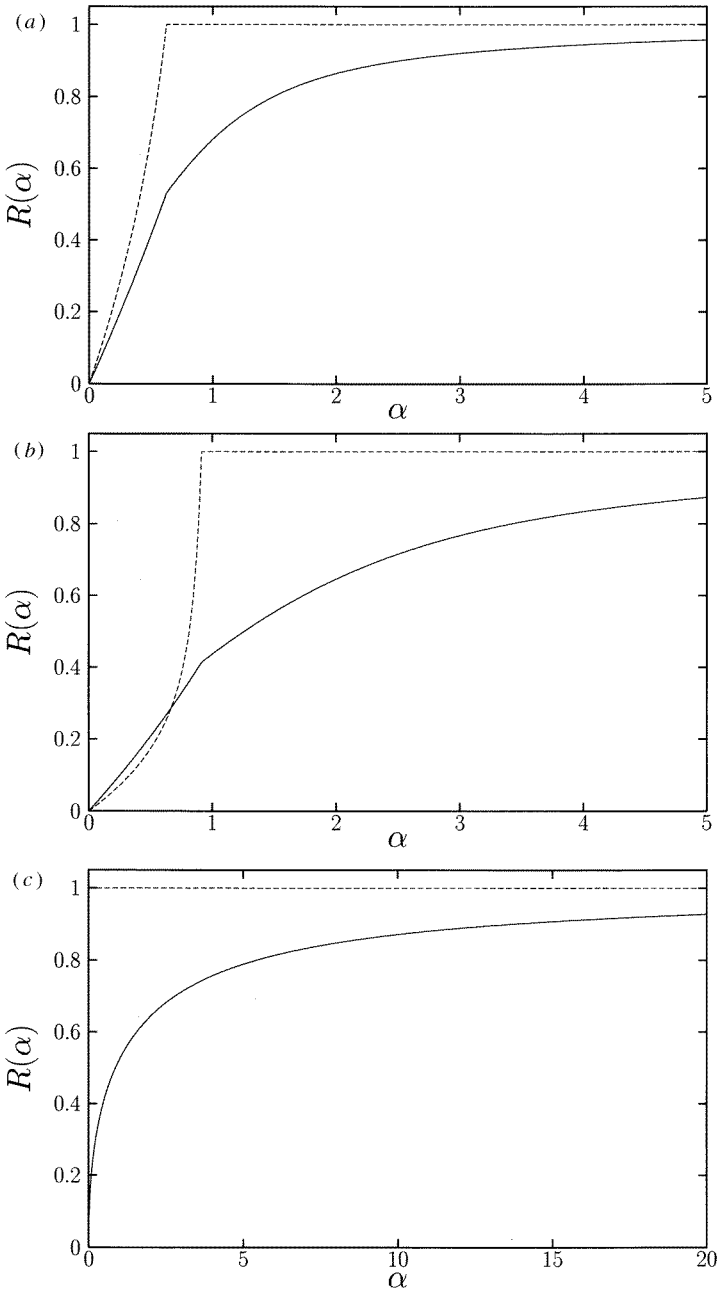

Figure 2. The overlap $R(\alpha)$ (full) and the self-overlap $q(\alpha)$ (broken) for unsupervised learning of Gaussian patterns (1), (2) by minimization of the quadratic cost function (3), (4). (*a*) Typical example with $c > 0$ ($a = b = 3$, $c = d = 1$) showing a characteristic singularity of the overlaps at α_c. (*b*) The same but with $R(\alpha) > q(\alpha)$ for small α ($a = b = 3$, $c = 2$, $d = 1$). (*c*) Typical example with $c < 0$ ($a = b = 3$, $c = -1$, $d = 3$) exhibiting the characteristic $\sqrt{\alpha}$-behaviour of $R(\alpha)$ for small α and the trivial self-overlap $q(\alpha) = \Theta(\alpha)$. (*d*) The same but with $R(\alpha) < 1$ for $\alpha \to \infty$ ($a = b = 3$, $c = -1$, $d = 0.2$). (*e*) An example for retarded classification due to a symmetric pattern distribution ($a = 3$, $b = 0$, $c = d = 1$). (*f*) Retarded classification induced by a symmetric *ad hoc* potential ($a = 3$, $b = 1$, $c = 1$, $d = 0$).

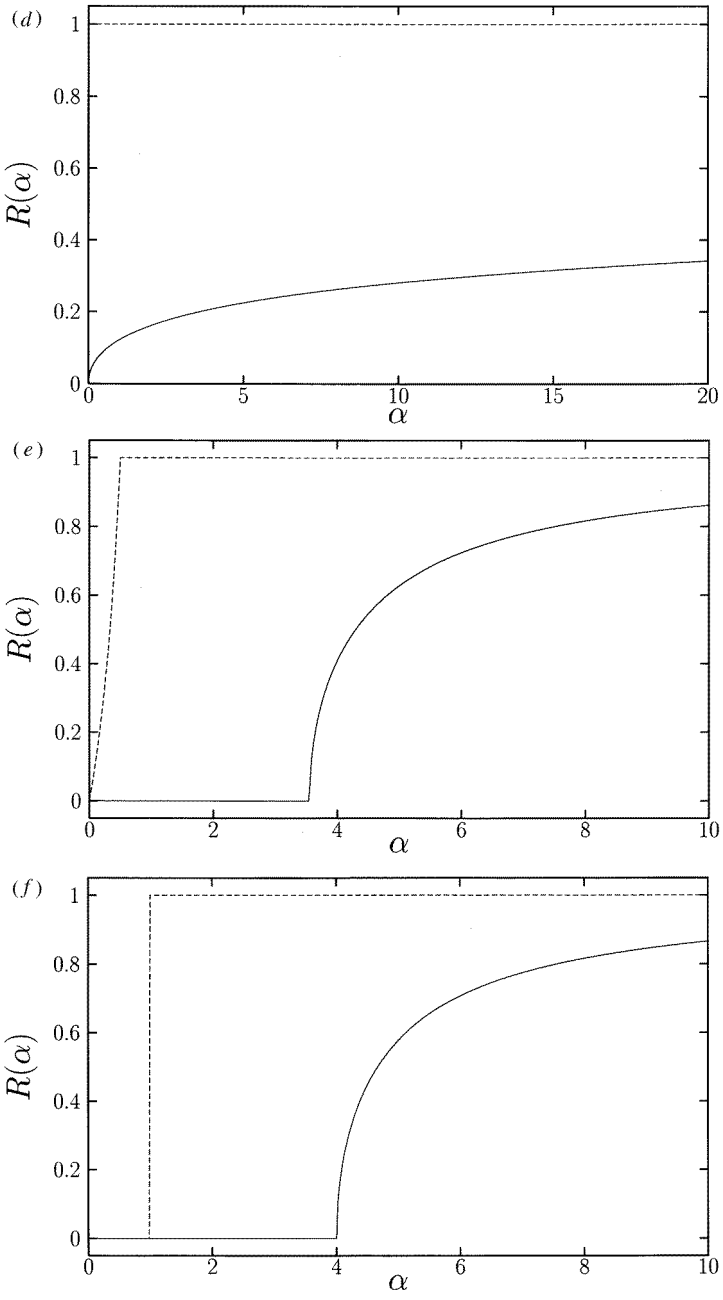

Figure 2. (Continued.)

classification occurs (see figures 2(*e*), (*f*)). One then finds for arbitrary α that

$$R(\alpha) = \Theta(c[A - B^2])\,\Theta(\alpha - \alpha_0)\sqrt{\frac{\alpha - \alpha_0}{\alpha - (A - B^2)^{-1}}} \tag{30}$$

$$\alpha_0 := [1 + D^2]/[A - B^2]^2 \tag{31}$$

if $c \neq 0$ and $R(\alpha) \equiv 0$ if $c = 0$. In particular, for $c\,[A - B^2] \leqslant 0$ nothing at all can be learned if either $B \to 0$ or $D \to 0$. For $c < 0$, $d = 0$ (maximal variance learning) and $c > 0$, $d = 0$ (corresponding to 'minimal variance learning') it is, in fact, rather obvious that the symmetry-breaking orientation $\boldsymbol{B}$ will be orthogonal to the direction with maximal or minimal variance of the patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$ if $A - B^2$ is positive or negative, respectively, since the variance of the examples in the $\boldsymbol{B}$-direction is just $1 - [A - B^2]$ and 1 in any direction orthogonal to $\boldsymbol{B}$. A similar phenomenon has been observed in [2]. Note also that the retardation threshold α_0, in general, depends on the specific learning rule under consideration, but is obviously bounded from below by the one corresponding to Bayes learning (17). Finally, for $B\,D = 0$ one has $q(\alpha) = \Theta(\alpha)$ for $c < 0$, while for $c > 0$ the self-overlap exhibits the non-trivial behaviour $q(\alpha) = \alpha\,D^2/(1 - \alpha)$ in the domain $\alpha < \alpha_c = 1/(1 + D^2)$, see figures 2(*e*), (*f*). In agreement with the general conjecture made in [6], the retardation threshold α_0 is always larger or equal to α_c. The spin-glass-type phase with $q(\alpha) > 0$ but $R(\alpha) = 0$ has been termed 'phase of confusion' in [3] since the hypotheses $\boldsymbol{J}$ become oriented but not correlated with the 'true' $\boldsymbol{B}$.

The large-α asymptotics for the Hebb rule ($c = 0$) and for $B\,D = 0$ is obvious from (20) and (30), respectively, whereas in any other case one finds that

$$R(\alpha) = \begin{cases} R_0 - \sqrt{\dfrac{1}{\alpha}\,\dfrac{F(R_0)\,R_0^4}{(1 - R_0^2)\,B^2\,D^2}} & \text{if} \quad 0 < R_0 < 1 \\ 1 - \sqrt[3]{\dfrac{1}{\alpha}\,\dfrac{F(1)}{2\,B^2\,D^2}} & \text{if} \quad BD = B^2 - A \\ 1 - \dfrac{1}{\alpha}\,\dfrac{F(1)}{2(A - B^2 + BD)^2} & \text{otherwise}\,. \end{cases} \tag{32}$$

In particular, for $0 < R_0 < 1$ the symmetry-breaking direction $\boldsymbol{B}$ can never be perfectly inferred even from infinitely many examples, cf figure 2(*d*).

4.3. Saturation of the Bayes limit

From equations (10), (16) and (20) one readily sees that the overlap $R(\alpha)$ corresponding to the minimization of the cost function (3), (4) coincides with the Bayes result $R_{\mathrm{B}}(\alpha)$ for any pattern distribution (1) with $a = 0$ provided one chooses $c = 0$ (Hebb rule). As pointed out in section 4.1, the particular value of $d > 0$ does not matter and one can take, for instance, $d = b$. In other words, for $a = 0$ the overlap $R_{\mathrm{M}}(\alpha)$ for the maximum *a posteriori* learning rule (minimization of the cost function (3) with $V(\lambda) = V^*(\lambda)$, cf section 2) reaches the upper theoretical limit imposed by the Bayes algorithm. The same result $R_{\mathrm{M}}(\alpha) = R_{\mathrm{B}}(\alpha)$ is recovered also for symmetric pattern distributions $b = 0$ by comparison of (17) and (30), while for a and b both non-vanishing one can show that $R_{\mathrm{M}}(\alpha) < R_{\mathrm{B}}(\alpha)$ (except for $\alpha = 0$ and $\alpha \to \infty$). In the latter case one might wonder whether one still can find a potential $V(\lambda)$ different from $V^*(\lambda)$ that reaches the Bayes limit. In the appendix C it is proved that this is indeed possible, namely by choosing a quadratic potential (4) with parameter values c and d that satisfy

$$\frac{c}{d} = \frac{a}{b}\,R_{\mathrm{B}}(\alpha) \tag{33}$$

where $b > 0$ and $d > 0$ is tacitly assumed as usual. For examples see figure 3. The mismatch between this optimal choice of the potential $V(\lambda)$ and the maximal *a posteriori* learning rule $V(\lambda) = V^*(\lambda)$ can be understood [3] by the fact that, given a set of patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$, the corresponding most probable hypothesis $\boldsymbol{J}$ from (9) may be quite different

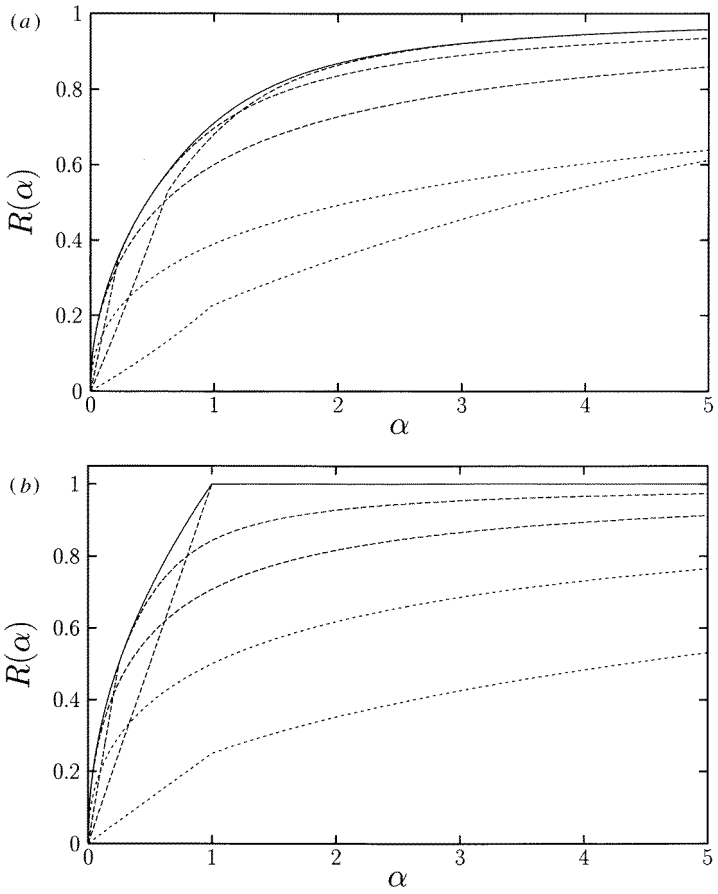

Figure 3. The overlap $R_{\mathrm{B}}(\alpha)$ for Bayes learning (full curve) according to (10) and (11) at parameter values $a = b = 3$ in (*a*) and $a \to \infty$ with $b/a = 1$ fixed in (*b*). The broken curves are the overlaps $R(\alpha)$ corresponding to minimizing the quadratic cost function (3), (4) with parameter values $c = 0$, $d = 1$ (Hebb rule, touching $R_{\mathrm{B}}(\alpha)$ for small α), $2c = d = 1$ and $c = d = 1$ (maximal *a posteriori* probability rule, touching $R_{\mathrm{B}}(\alpha)$ for large α). The dotted curves represent examples ($c = 4$, $d = 1$ and $c = -1$, $d = 1$) which never touch $R_{\mathrm{B}}(\alpha)$ for finite α.

from a suitably defined 'typical' one. Exceptions are $a = 0$ or $b = 0$, as already seen, as well as asymptotically large α (according to (33) with $R_{\mathrm{B}}(\infty) = 1$) or $a \to \infty$ with b/a fixed and $\alpha > 1$, see equation (18) and figure 3(*b*). In the latter case, where all patterns lie on a cone about $\boldsymbol{B}$, the choice $d/c = b/a$ leads to a perfect guess for $\boldsymbol{B}$ after $p = N$ patterns have been seen, while any other choice of d/c yields $R(\alpha) < 1$ for all $\alpha < \infty$.

In conclusion, we see that a potential saturating the Bayes limit can always be found but that this optimal choice typically depends on α and the details of the pattern distribution. Exceptions are asymptotically small α for which the Hebb rule is always optimal. We finally note that according to (33), for given non-vanishing values of a and b the function $R_{\mathrm{B}}(\alpha)$ is recovered as the envelope of all the learning curves $R(\alpha)$ generated by different choices of c and d. However, only those with $0 \leqslant d \leqslant bc/a$ and c of the same sign as a actually touch $R_{\mathrm{B}}(\alpha)$, see figure 3. Although for $a > 0$ all these $R(\alpha)$ curves start off proportional to α, the envelope $R_{\mathrm{B}}(\alpha)$ increases like $\sqrt{\alpha}$. These features are reminiscent of those reported in [12].

5. Outlook

We studied unsupervised learning from examples governed by a distribution (1) with a single symmetry-breaking orientation $\boldsymbol{B}$ by means of replica methods. Our results include the performance of hypotheses $\boldsymbol{J}$ for this unknown direction $\boldsymbol{B}$ based on Bayes, Gibbs and maximal *a posteriori* learning algorithms as well as the minimization of a cost function (3) with quadratic *ad hoc* potentials (4). By restricting ourselves to Gaussian pattern distributions (2), a complete analytical solution of the problem was possible. In particular, we were able to determine the relevant *global* extremum of the replica-symmetric free energy (6) and to verify the local stability condition (8) analytically in all cases.

In spite of its simplicity, our 'Gaussian scenario' exhibits most of the features observed previously for more complicated pattern distributions as well as various novel phenomena. We only mention here the effect of retarded classification connected with a first- or second-order transition in the overlap $R(\alpha)$, the saturation of the Bayes limit by *ad hoc* potentials, the possibility of perfect learning for $\alpha < \infty$ and of imperfect learning even in the limit $\alpha \to \infty$. In fact, the general results obtained in [13] indicate that our 'Gaussian scenario' is indeed representative in practically all respects, at least for smooth pattern distributions (1). For instance, one can always find an *ad hoc* potential saturating the Bayes limit. Typically, it depends on α, coinciding with the Hebb rule for asymptotically small α and with the most probable *a posteriori* learning strategy for asymptotically large α. Also our finding that a symmetric pattern distribution or a symmetric potential imply retarded classification carries over unchanged to the general case.

The saturation of the Bayes limit by using *ad hoc* potentials is not only of principal but also of practical interest since the implementation of the Bayes algorithm itself is numerically extremely expensive. Therefore some remarks regarding the numerical minimization of the quadratic cost function (3), (4) with side condition $\boldsymbol{J}^2 = N$ may be in place. For a linear potential ($c = 0$) this minimization becomes equivalent to Hebbian learning and is thus trivial to implement. The case $d = 0$ and $c < 0$ corresponds to maximal variance learning and can be efficiently realized, e.g. by Oja's rule [14]. For $d = 0$ and $c > 0$ one can show that the cost function has a unique minimum on the N-sphere (at least beyond the retardation threshold α_0, cf section 4.2, and apart from the trivial symmetry $\boldsymbol{J} \mapsto -\boldsymbol{J}$) which thus can be found by gradient descent or similar more efficient procedures [14]. If both c and d are non-zero we cannot exclude local minima of the cost function and algorithms which might get stuck in one of them should not be used. In order to reach the Bayes limit we propose the following novel learning algorithm: one starts with Hebbian learning of a few examples. Then $\boldsymbol{J}$ is updated sequentially in two steps by adding one or a few new examples each time: in a first step the cost function is minimized with respect to the slightly enlarged set of examples without changing the potential $V(\lambda)$. Since, as we demonstrated in this paper, the assumption of replica symmetry seems to be valid, the cavity concepts (see [9] and further references therein) strongly suggest that the minimizing $\boldsymbol{J}$ changes little and can be updated, e.g. by gradient descent. In a second step, the potential is adapted to the slightly increased α-value according to (33). Again, one expects that in this way the minimizing $\boldsymbol{J}$ changes only a little and can be updated by standard methods. We verified this procedure by extensive numerical simulations as discussed in more detail in appendix D. From the results shown in figure 4 we conclude that our algorithm indeed approaches the Bayes limit for asymptotically large N. We finally mention that the most probable *a posteriori* learning rule for large α, where it reaches the Bayes limit, can also be implemented very efficiently by a suitably tailored on-line algorithm [13].

Throughout our investigation we assumed to know $V^*(\lambda)$ from (2) and it turned out

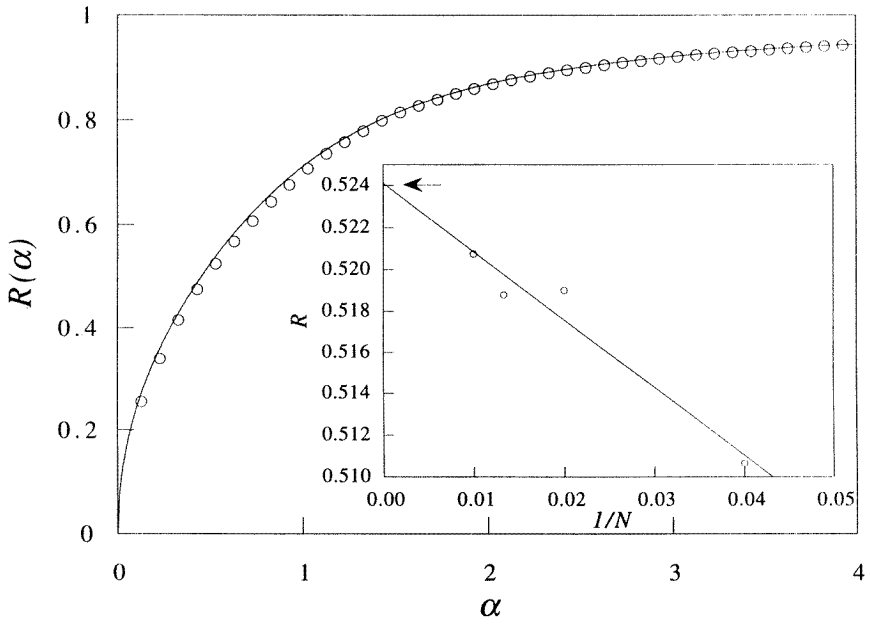

Figure 4. The numerical overlap $R(\alpha)$ (circles) obtained by the iterative gradient descent-type algorithm on the optimal *ad hoc* potential proposed in section 5 (see also appendix D) for $N = 100$, $a = b = 3$. The full curve is the theoretical Bayes limit $R_{\mathrm{B}}(\alpha)$ from (10), (11). The inset exemplifies the extrapolation $N \to \infty$ (from data for $N = 25$, 50, 75, 100) for $\alpha = 0.5$. The theoretical Bayes limit is indicated by the arrow ($R_{\mathrm{B}}(\alpha = 0.5) = 0.524$).

that this knowledge is crucial to find a good hypothesis $\boldsymbol{J}$ for the unknown 'true' $\boldsymbol{B}$. We will show elsewhere how an unknown $V^*(\lambda)$ can be determined *exactly* from the same training set $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$ in the thermodynamic limit $N \to \infty$ with $\alpha = p/N > 0$ fixed. This underscores the relevance of the present results in a more general context.

Acknowledgments

Discussions with Eddy Lootens and Lüder Reimers are gratefully acknowledged. We thank the Program on Inter-University Attraction Poles of the Belgian Government, the NFWO Belgium, the Holderbank Foundation (Switzerland), the Hungarian Ministry for Culture and Education and the Swiss National Science Foundation for financial support.

Appendix A. Derivation of (11)

One can conclude from (6) that

$$\frac{\mathrm{d}\, G(q = R, R)}{\mathrm{d}\, R} = \frac{\alpha}{2}\left[B^2 + \frac{A^2\, R}{1 - A\, R}\right] - \frac{1}{2}\,\frac{R}{1 - R}\,. \tag{A1}$$

Since this derivative (A1) is smooth for $0 < R < 1$, non-negative for $R \to 0$ and approaches $-\infty$ for $R \to 1$, we can focus on the determination of its zeros in the interval $[0, 1]$ in order to find the maximizing $R = R_{\mathrm{G}}(\alpha)$. The unique such zero is given by (11). This solution is also consistent with the tacit restriction to $1 + cx > 0$ in (6) (see also the second footnote in section 2) due to $cx = a(1 - q) > -1$.

Appendix B. Derivation of (22)–(26)

By closer inspection one can see that an extremum of (6) as specified in the second footnote of section 2 exists, is generically unique and satisfies $\partial G(q, R)/\partial q = 0$, $\partial G(q, R)/\partial R = 0$ for any fixed $0 < \beta < \infty$.

We now consider separately the cases where the extremizing $q = q(\alpha)$ in (6) stays below 1 or converges to 1 for $\beta \to \infty$. In the former case $q(\alpha) < 1$ we can restrict ourselves to $c > 0$ since $c < 0$ is incompatible with the side condition $1 + cx > 0$. Then, equation (6) can be rewritten (up to an irrelevant additive constant) in the form

$$G(q, R) = \frac{1-\alpha}{2} \ln(1-q) + \frac{1 - R^2 - \alpha F(R)}{2(1-q)} \tag{B1}$$

where $F(R)$ is defined in (27). From the extremization condition $\partial G(q, R)/\partial R = 0$ one recovers (25), while $\partial G(q, R)/\partial q = 0$ yields (24). The side condition $1+cx > 0$ is trivially fulfilled, whereas both the minimization condition with respect to q, $\partial^2 G(q, R)/\partial q^2 > 0$, and the stability condition (8) are fulfilled if and only if $\alpha < 1$. Setting $q(\alpha) = 1$ in (24) and eliminating $R(\alpha)$ by means of (25) one obtains a quadratic equation for α with the unique solution (22) in the admitted domain $0 \leqslant \alpha < 1$.

Next we address the case that $q(\alpha) \to 1$ for $\beta \to \infty$. Since the minima of the cost function (3), (4) are of quadratic order, $x = \beta(1-q)$ is the sensitive quantity in this limit [9] and the extremization of $G(q, R)$ from (6) becomes equivalent to the extremization of

$$\tilde{G}(x, R) = \frac{1-R^2}{2x} - \frac{\alpha}{2} \frac{cF(R)}{1+cx}. \tag{B2}$$

We recall that we have to minimize $\tilde{G}(x, R)$ with respect to $x \in [0, \infty)$ and then to maximize with respect to $R \in [-1, 1]$. Further, let us restrict ourselves for the moment to $c > 0$. For any fixed $R \in [-1, 1]$ the corresponding minimizing $x = x(R)$ is then readily obtained:

$$x(R) = \sqrt{1-R^2}/[ch(R)] \tag{B3}$$

$$h(R) := \sqrt{\alpha F(R)} - \sqrt{1-R^2} \tag{B4}$$

and we are left with the maximization of

$$\tilde{G}(x(R), R) = -c\, h(R)^2/2 \tag{B5}$$

provided $h(R) > 0$ for all $R \in [-1, 1]$. In the opposite case, $h(R) \leqslant 0$ for at least one $R \in [-1, 1]$, our initial assumption that the extremizing $x = x(\alpha)$ is finite breaks down, indicating that actually $q(\alpha)$ stays below 1 for $\beta \to \infty$. Since $A < 1$ it follows that $F(R)$ from (27) is positive for $-1 \leqslant R \leqslant 1$. Consequently, there must exist a critical α value $\tilde{\alpha}_c$ above which the condition $h(R) > 0$ for all $R \in [-1, 1]$ is satisfied and below which it is violated. As one expects, it will turn out that this $\tilde{\alpha}_c$ agrees with α_c from (22). In the following we will determine the value $R = R(\alpha)$ that maximizes $\tilde{G}(x(R), R)$ from (B5) and thus minimizes $h(R)$ from (B4) provided $\alpha \geqslant \tilde{\alpha}_c$. We thus know that with decreasing α, $h(R(\alpha))$ vanishes for the first time at $\alpha = \tilde{\alpha}_c$. Since additionally $h(R(\alpha)) < 0$ for all $\alpha < \tilde{\alpha}_c$, the value of $\tilde{\alpha}_c$ is uniquely fixed through

$$h(R(\tilde{\alpha}_c)) = 0. \tag{B6}$$

Exploiting the fact that $h(R) \geqslant 0$ for $\alpha \geqslant \alpha_c$ we can infer that the side condition $1+cx > 0$ is always fulfilled by $x(R)$ from (B3).

We now turn to the minimization of $h(R)$ from (B4). One readily sees that the minimum $R = R(\alpha)$ cannot be at $R = 1$ nor in the domain $-1 \leqslant R \leqslant 0$ due to the term $\sqrt{1-R^2}$ in

(B4) and $B > 0$, $D > 0$ in (27), respectively (for simplicity, the cases $B = 0$ and $D = 0$ are treated here as limits of small positive B and D, see also [6]). We can thus concentrate on solutions of $h'(R) = 0$ in the region $0 < R < 1$. Differentiating (B4), this implies that $R(\alpha)$ must be a solution of (26) (in particular, $0 \leqslant R(\alpha) \leqslant R_0$ must be satisfied). By means of a straightforward but somewhat tedious discussion of this fourth-order equation (26) one can show that a solution exists and is unique in the prescribed interval $[0, R_0]$ for any α. This unique solution of (26) can thus be identified with $R(\alpha)$ provided α is above the critical value $\tilde{\alpha}_c$. At this critical value, $R = R(\tilde{\alpha}_c)$ satisfies both (B6) and (26) and after eliminating R one recovers that $\tilde{\alpha}_c$ indeed coincides with α_c from (22).

For $c < 0$ one can show by a similar line of reasoning that $q(\alpha) = 1$ and $R(\alpha)$ following from (26) is a solution of the extremization problem plus side condition $1 + cx > 0$, but now for arbitrary $\alpha > 0$. By closer inspection one finally finds that these solutions with $q(\alpha) \to 1$ for $\beta \to \infty$ for both $c > 0$ and $c < 0$ satisfy the stability condition (8).

In conclusion, we identified for all possible parameter- and α-values the relevant global solution of the extremization problem (6) plus side condition $1 + cx > 0$ and we verified the stability condition (8) in all cases.

Appendix C. Derivation of (33)

We want to show that for given values of a, b, and α we recover $R(\alpha) = R_{\mathrm{B}}(\alpha)$ if we choose c and d according to (33). To this end it is sufficient to verify that with c and d from (33) and $R = R_{\mathrm{B}}(\alpha)$ equation (26) is solved (due to the uniqueness of such a solution found in appendix B) and, in the case that $a > 0$, additionally $h(R)$, defined in (B4), is non-negative for $R = R_{\mathrm{B}}(\alpha)$, guaranteeing that $\alpha \geqslant \alpha_c$. Introducing (33) and $R = R_{\mathrm{B}}(\alpha)$ into the definition (27) one obtains

$$A\,F(R) = (1 - A\,R^2)\,\tilde{h}(R) \tag{C1}$$

$$\tilde{h}(R) := A - B^2 + B\,D/R\,. \tag{C2}$$

Exploiting (C1) and the fact that (A1) is zero at $R_{\mathrm{G}}(\alpha) = R_{\mathrm{B}}(\alpha)^2$ one finds that $R = R_{\mathrm{B}}(\alpha)$ satisfies $\alpha\,\tilde{h}(R)^2(1 - R^2) = F(R)$ and that

$$\alpha\,\tilde{h}(R) = \frac{1 - A\,R^2}{A(1 - R^2)} \begin{cases} > 1 & \text{for } a \geqslant 0 \\ < 0 & \text{for } a < 0\,. \end{cases} \tag{C3}$$

From equations (C3), (C2) and (28) one can infer that $R < R_0$. Since $R \geqslant 0$ is trivially true for our choice $R = R_{\mathrm{B}}(\alpha)$ the verification of (26) is completed. We are left to prove $h(R) \geqslant 0$ in the case that $a > 0$. This is readily achieved by introducing (26) into the definition (B4) and then making use of the inequality (C3).

Appendix D. Simulations

To verify whether the Bayes limit can indeed be reached by means of the algorithm proposed in section 5 we performed extensive numerical simulations. Patterns were generated according to the distribution (1), (2) by drawing the first component from a Gaussian distribution with mean value $m = b/(1 + a)$ and standard deviation $\sigma^2 = 1/(1 + a)$, while the remaining $N - 1$ components were taken to be normally distributed random variables.

Since for asymptotically small α the Hebb perceptron reaches the Bayes limit (cf the end of section 4.3) we used the Hebb rule as a starting point for the simulations:

$$\boldsymbol{J} = \frac{1}{p}\sum_{\mu=1}^{\alpha_s p}\boldsymbol{\xi}^{\mu} \tag{D1}$$

with $\alpha_s \approx 0.05$. The additional $p - \alpha_s N$ patterns were added one by one to the data set. At each step, the corresponding $\boldsymbol{J}$-vector was updated by minimization of the cost function $E(\boldsymbol{J})$:

$$E(\boldsymbol{J}) = \sum_{\mu=1}^{p} V(\boldsymbol{\xi}^{\mu}\cdot\boldsymbol{J}/\sqrt{N}) + \gamma(\boldsymbol{J}^2 - N)^2 \tag{D2}$$

where the potential $V(\lambda)$ is fixed through (4), (10), (11), (33) and the last term is added to ensure proper normalization of $\boldsymbol{J}$ (γ being typically equal to 10.0). Minimization with respect to this cost function was performed by means of the Fletcher–Reeves–Polak–Ribiere algorithm which is essentially a conjugate gradient-descent method. For each value of α, we then calculated the overlap of the generated hypothesis $\boldsymbol{J}$ with the 'true' preferential direction $\boldsymbol{B}$. The resulting curve for $N = 100$ is shown in figure 4. The overlap has been averaged over 400 independent random sets of patterns. The small deviations from the theoretical Bayes limit (less than 2% for $\alpha > 0.2$) can be attributed to finite-size effects as is illustrated in the inset of figure 4.

References

[1] Barkai N, Seung H S and Sompolinsky H 1993 *Phys. Rev. Lett.* **70** 3167
[2] Biehl M and Mietzner A 1994 *Europhys. Lett.* **24** 421; 1994 *J. Phys. A: Math. Gen.* **27** 1885
[3] Watkin T H L and Nadal J-P 1994 *J. Phys. A: Math. Gen.* **27** 1899
[4] Lootens E and Van den Broeck C 1995 *Europhys. Lett.* **30** 381
[5] Mietzner A, Opper M and Kinzel W 1995 *J. Phys. A: Math. Gen.* **28** 2785
[6] Reimann P and Van den Broeck C 1996 Learning by examples from a non-uniform distribution *Phys. Rev.* E at press
[7] Widrow B and Hoff M E 1960 *WESCO Convention Report IV* (San Fransisco, CA: Western Periodicals)
[8] Griniasty M and Gutfreund H 1991 *J. Phys. A: Math. Gen.* **24** 715
[9] Bouten M, Schietse J and Van den Broeck C 1995 *Phys. Rev.* E **52** 1958
[10] de Almeida J R L and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983
[11] Hansel D, Mato G and Meunier C 1992 *Europhys. Lett.* **20** 471
[12] Derényi I, Geszti T and Györgyi G 1994 *Phys. Rev.* E **50** 3192
[13] Van den Broeck C and Reimann P Unsupervised learning from examples: on-line versus off-line *Phys. Rev. Lett.* at press
[14] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Reading, MA: Addison-Wesley)