Phase transitions in optimal unsupervised learning

Arnaud Buhot and Mirta B. Gordon*
*Département de Recherche Fondamentale sur la Matière Condensée, CEA/Grenoble, 17 rue des Martyrs,
38054 Grenoble Cedex 9, France*

(Received 25 September 1997; revised manuscript received 12 November 1997)

We determine the optimal performance of learning the orientation of the symmetry axis of a set of $P=\alpha N$ points that are uniformly distributed in all the directions but one on the N-dimensional space. The components along the symmetry breaking direction, of unitary vector $\mathbf{B}$, are sampled from a mixture of two Gaussians of variable separation and width. The typical optimal performance is measured through the overlap $R_{\mathrm{opt}}=\mathbf{B}\cdot\mathbf{J}^*$, where $\mathbf{J}^*$ is the optimal guess of the symmetry breaking direction. Within this general scenario, the learning curves $R_{\mathrm{opt}}(\alpha)$ may present first order transitions if the clusters are narrow enough. Close to these transitions, high performance states can be obtained through the minimization of the corresponding optimal potential, although these solutions are metastable, and therefore not learnable, within the usual Bayesian scenario. [S1063-651X(98)07303-6]

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

I. INTRODUCTION

In this paper we address a very general problem in the statistical analysis of large amounts of data points, also called *examples*, *patterns*, or *training set*, namely the problem of discovering the structure underlying the data set. Whether this determination is possible or not depends on the assumptions one is willing to accept [1]. Several algorithms allowing to detect structure in a set of points exist. Among them, principal component analysis finds the directions of higher variance, projection pursuit methods [2] seek directions in input space onto which the projections of the data maximize some measure of departure from normality, whereas self-organizing clustering procedures [3] allow to determine prototype vectors representative of clouds of data. The parametric approach assumes that the structure of the probability density function from which the patterns have been sampled is known. Only its parameters have to be determined given the examples. A frequent guess is that the probability density is either Gaussian or a mixture of Gaussians. The process of determining the corresponding parameters is called *unsupervised learning*, because we are not given any additional information about the data, in contrast with *supervised learning*, in which each training example is labeled.

It has recently been shown that finding the principal component of a set of examples, clustering data with a mixture of Gaussians, and learning pattern classification from examples with neural networks may be cast as particular cases of unsupervised learning [4]. In all these problems, the examples are drawn from a probability density function (PDF) with axial symmetry, and the symmetry-breaking direction has to be determined given the training set. As this direction may be found through the minimization of a cost function, the properties of unsupervised learning may be analyzed with statistical mechanics. This approach allows to establish the properties of the typical solution, determined in the thermodynamic limit, i.e., the space dimension $N\rightarrow+\infty$, the number of examples $P\rightarrow+\infty$, with the fraction of examples $\alpha=P/N$ constant.

Besides these general results, the statistical mechanics framework allows to deduce the expression of an *optimal cost function* [5–7], whose minimum is the best solution that may be expected to be learned given the data. The optimal cost function depends on the functional structure of the PDF from which the examples are sampled, and on the fraction α of available examples. Its main interest is that it allows to deduce the upper bound for the typical performance that may be expected from any learning algorithm. On the other hand, Bayes' formula of statistical inference allows to determine the probability of the symmetry-breaking direction given the training set. Sampling the direction with Bayes probability is called Gibbs learning [8]. The average of the solutions obtained through Gibbs learning, weighted with the corresponding probability, is called the *Bayesian* solution. It is widely believed that the Bayesian solution is optimal. Moreover, this has been so in all the scenarios considered so far.

In the present paper, we consider a very general two-cluster scenario, which contains results already reported as particular cases. In fact, two different situations, in which the pattern distribution is a Gaussian of zero mean and unit variance in all the directions but one, have been considered so far: a Gaussian scenario [9] and a two-cluster scenario [10,11,8]. In the former, the components of the examples parallel to the symmetry-breaking direction are sampled from a single Gaussian. In the latter these components are drawn from a mixture of two Gaussians, each one having unit variance. The learning process has to detect differences between the PDF along the symmetry-breaking direction and the distributions in the orthogonal directions. Several *ad hoc* cost functions allowing to determine the symmetry-breaking direction have been analyzed for both scenarios. Typically, if the PDF has a nonzero mean value in the symmetry-breaking direction, learning is ‘‘easy’’: the quality of the solution increases monotonically with the fraction α of examples, starting at $\alpha=0$. In contrast, if the PDF has zero mean, the deviations of the PDF along the symmetry-breaking direction from the PDF in the orthogonal directions depend on the

*Also at Centre National de la Recherche Scientifique.

second and higher moments. In this case, a phenomenon called *retarded learning* [8] appears: learning the symmetry-breaking direction becomes impossible when the fraction of examples falls below a critical value α_c.

Since we have considered the case of clusters of variable width, we could determine the entire phase diagram of the two-cluster scenario. Several new learning phases appear, depending on the mean and the variance of the clusters. In particular, if the second moment of the individual clusters is smaller than the second moment of the PDF in the orthogonal directions, first order transitions from low to high performance learning may occur as a function of α. Close to these, high performance metastable states exist above the stable states of Gibbs learning, in the thermodynamic limit. One of the most striking results of this paper is that these high performance metastable states can indeed be learned through the minimization of an optimal α-dependent potential, although they cannot be obtained through Bayesian learning.

Our results have been obtained within the replica approach with the replica-symmetry hypothesis. We show below that this assumption is equivalent to the more intuitive requirement that the optimal learning curves $R_{\rm opt}(\alpha)$ are increasing functions of the fraction of examples α. To our knowledge, this fact has not been noticed before.

The paper is organized as follows. A short presentation of the problem and the replica calculation are given in Sec. II. In Sec. III we deduce the optimal cost functions within the replica-symmetry hypothesis, as well as the condition of replica-symmetry stability. In Sec. IV we deduce and discuss the optimal learning curves for the general two-cluster scenario. The typical properties of the optimal cost functions in the complete range of α, presented in Sec. V, show that Bayesian learning may not be optimal. Finally, the complete phase diagram is described in Sec. VI, as a function of the two clusters' parameters.

II. GENERAL FRAMEWORK AND REPLICA CALCULATION

We consider the general case of N-dimensional vectors $\boldsymbol{\xi}$, the patterns or examples of the training set, drawn from an axially symmetric probability density $P^*(\boldsymbol{\xi}|\mathbf{B})$ of the form

$$P^*(\boldsymbol{\xi}|\mathbf{B}) \equiv \frac{1}{(2\pi)^{N/2}} \exp\left\{-\frac{\boldsymbol{\xi}\cdot\boldsymbol{\xi}}{2} - V^*(\lambda)\right\}, \qquad (1)$$

where $\mathbf{B}$ is a unitary vector in the symmetry-breaking direction, i.e., $\mathbf{B}\cdot\mathbf{B}=1$ (notice that this is *not* the usual convention), and $\lambda \equiv \boldsymbol{\xi}\cdot\mathbf{B} = \Sigma_i \xi_i B_i$. According to Eq. (1), the patterns have normal distributions, i.e., $P(x) = \exp(-x^2/2)/\sqrt{2\pi}$ onto the $N-1$ directions orthogonal to $\mathbf{B}$. The distribution (1) in the symmetry-breaking direction is

$$P^*(\lambda) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\lambda^2}{2} - V^*(\lambda)\right\}. \qquad (2)$$

Thus, $V^*(\lambda)$ introduces a modulation parallel to $\mathbf{B}$; if $V^*=0$ the patterns' distribution is normal in all the directions. Normalization of P^* requires

$$\int_{-\infty}^{+\infty} D\lambda \, \exp[-V^*(\lambda)] = 1, \qquad (3)$$

where $D\lambda = \exp(-\lambda^2/2)d\lambda/\sqrt{2\pi}$. The different moments $\langle\lambda^n\rangle$ of (2) are

$$\langle\lambda^n\rangle \equiv \int (\boldsymbol{\xi}\cdot\mathbf{B})^n P^*(\boldsymbol{\xi}|\mathbf{B}) d\boldsymbol{\xi} = \int_{-\infty}^{+\infty} \lambda^n P^*(\lambda)\, d\lambda. \qquad (4)$$

Several examples of functions V^* have been treated in the literature so far [4,7–11]. In the particular case of supervised learning of a linearly separable classification task by a single unit neural network, the symmetry-breaking direction $\mathbf{B}$ is the *teacher's* vector, orthogonal to the hyperplane separating the classes. The class of pattern $\boldsymbol{\xi}$ is $\tau \equiv \mathrm{sgn}(\mathbf{B}\cdot\boldsymbol{\xi})$. The corresponding PDF is $P^*(\tau\lambda) = 2\,\Theta(\tau\lambda)\exp(-\lambda^2/2)/\sqrt{2\pi}$, i.e., $V^*(\lambda) = -\ln 2$ for $\tau\lambda > 0$ and $+\infty$ for $\tau\lambda < 0$.

In the following, we concentrate on the problem of unsupervised learning. We are given a *training* set $\mathcal{L}_\alpha = \{\boldsymbol{\xi}^\mu\}_{\mu=1,\ldots,P}$ of $P=\alpha N$ vectors sampled independently with probability density $P^*(\boldsymbol{\xi}|\mathbf{B})$. We have to *learn* the unknown symmetry-breaking direction $\mathbf{B}$ from the examples knowing the functional dependence of P^* on $\mathbf{B}$. Using Bayes' rule of inference, the probability of a direction $\mathbf{J}$ (with $\mathbf{J}\cdot\mathbf{J}=1$) given the data is

$$P(\mathbf{J}|\mathcal{L}_\alpha) = \frac{1}{\mathcal{Z}} \prod_\mu \exp\{-\boldsymbol{\xi}^\mu\cdot\boldsymbol{\xi}^\mu/2 - V^*(\boldsymbol{\xi}^\mu\cdot\mathbf{J})\} P_0(\mathbf{J}), \qquad (5)$$

where $P_0(\mathbf{J}) = \delta(\mathbf{J}\cdot\mathbf{J}-1)$ is the assumed prior probability and $\mathcal{Z} = \int d\mathbf{J}\Pi_\mu \exp\{-\boldsymbol{\xi}^\mu\cdot\boldsymbol{\xi}^\mu/2 - V^*(\boldsymbol{\xi}^\mu\cdot\mathbf{J})\}P_0(\mathbf{J})$ is the probability of the training set. By analogy with supervised learning, sampling the direction with probability (5) is called *Gibbs learning* [8].

We consider learning procedures where the direction $\mathbf{J}$ is found through the minimization of a cost function or energy $E(\mathbf{J};\mathcal{L}_\alpha)$. As the patterns are independently drawn, this energy is an additive function of the examples. The contribution of each pattern $\boldsymbol{\xi}^\mu$ to E is given by a *potential* V that depends on the direction $\mathbf{J}$ and on $\boldsymbol{\xi}^\mu$ through the projection (called *local field*) $\gamma^\mu = \mathbf{J}\cdot\boldsymbol{\xi}^\mu$:

$$E(\mathbf{J};\mathcal{L}_\alpha) = \sum_{\mu=1}^{P} V(\gamma^\mu). \qquad (6)$$

As the training set only carries partial information on the symmetry-breaking direction $\mathbf{B}$, the direction $\mathbf{J}$ determined by the minimization of Eq. (6) will generally differ from $\mathbf{B}$. The quality of a solution $\mathbf{J}$ may be characterized by the overlap $R=\mathbf{B}\cdot\mathbf{J}$. If $R=0$, $\mathbf{J}$ does not give any information about the symmetry-breaking direction. Conversely, if $R=1$ the symmetry-breaking direction is perfectly determined.

The statistical mechanics approach allows to calculate the expected overlap $R(\alpha)$ for any general distribution V^* and any general potential V, in the thermodynamic limit $N,P \to +\infty$ with $\alpha \equiv P/N$ finite. In this limit, we expect that the energy is self-averaging: its distribution is a δ peak centered at its expectation value independently of the particular realization of the training patterns. Given the modulation V^*, different values of R may be reached, depending on the potential used for learning. In the following, we sketch the main lines that allow to derive the typical value of R corresponding to a general potential V.

The free energy F corresponding to the energy (6) with a given potential $V(\gamma)$ is

$$F(\beta,N,\mathcal{L}_\alpha)=-\frac{1}{\beta}\ln Z(\beta,N,\mathcal{L}_\alpha), \tag{7}$$

where β is the inverse temperature and Z the partition function:

$$Z(\beta,N,\mathcal{L}_\alpha)=\int d\mathbf{J}\, \exp\{-\beta E(\mathbf{J};\mathcal{L}_\alpha)\}\delta(\mathbf{J}^2-1). \tag{8}$$

As mentioned before, in the thermodynamic limit the free energy is self-averaging, i.e.,

$$\lim_{N\to+\infty}\frac{1}{N}F(\beta,N,\mathcal{L}_\alpha)=\lim_{N\to+\infty}\frac{1}{N}\overline{F(\beta,N,\mathcal{L}_\alpha)}, \tag{9}$$

where $\overline{(\cdots)}$ stands for the average over all the possible training sets. The average in the right-hand side of Eq. (9) is calculated using the replica method:

$$\overline{\ln Z}=\lim_{n\to 0}\frac{1}{n}\ln\overline{Z^n}, \tag{10}$$

which reduces the problem of averaging $\ln Z$ to the one of averaging the partition function of *n replicas* of the original system, and taking the limit $n\to 0$. The properties of the minimum of the cost function are those of the zero temperature limit ($\beta\to+\infty$) of the free energy. In the case of differentiable potentials V, the integrals are dominated by the saddle point, and the zero temperature free energy reads [4]

$$\begin{aligned}f(R,c)&=\lim_{\beta\to+\infty}\lim_{N\to+\infty}\frac{1}{N}\overline{F(\beta,N,\mathcal{L}_\alpha)}\\&=-\frac{1}{2c}\left\{1-R^2-2\alpha\int Dt\, W(t;c)\right.\\&\qquad\left.\times\int Dz\, \exp[-V^*(\lambda)]\right\},\end{aligned} \tag{11}$$

where

$$\lambda\equiv z\sqrt{1-R^2}+Rt. \tag{12}$$

In Eq. (11), R is the overlap between the symmetry-breaking direction $\mathbf{B}$ and a minimum $\mathbf{J}$ of the cost function (6); $c=\lim_{\beta\to+\infty}\beta(1-q)$, where q is the overlap between minima of the cost function (6) for two different replicas, and

$$W(t;c)=\min{}_{\gamma}[cV(\gamma)+(\gamma-t)^2/2], \tag{13}$$

is the saddle point equation. The extremum conditions of the free energy (11) with respect to R and c, $\partial f/\partial R=\partial f/\partial c=0$, give the following equations for R and c:

$$1-R^2=\alpha\int_{-\infty}^{+\infty}Dt\, [\gamma(t;c)-t]^2\int_{-\infty}^{+\infty}Dz\, \exp[-V^*(\lambda)], \tag{14a}$$

$$\begin{aligned}R\sqrt{1-R^2}=\alpha\int_{-\infty}^{+\infty}Dt\, [\gamma(t;c)-t]\\\times\int_{-\infty}^{+\infty}Dz\, z\, \exp[-V^*(\lambda)],\end{aligned} \tag{14b}$$

where λ is defined in Eq. (12) and $\gamma(t;c)$ is the solution that minimizes Eq. (13). Introduction of Eq. (14) into Eq. (11) gives the free energy at zero temperature:

$$f(R,c)=\alpha\int Dt\, V(\gamma(t;c))\int Dz\, \exp[-V^*(\lambda)]. \tag{15}$$

If the potential $V(\gamma)$ is not convex, Eq. (14) may have more than one solution. In that case, the one minimizing Eq. (15) with respect to R should be kept.

These results were obtained under the assumption of replica symmetry. A necessary condition for the replica-symmetry hypothesis to be satisfied is

$$\alpha\int_{-\infty}^{+\infty}Dt\, [\gamma'(t;c)-1]^2\int_{-\infty}^{+\infty}Dz\, \exp[-V^*(\lambda)]<1, \tag{16}$$

with $\gamma'(t;c)\equiv\partial\gamma/\partial t$.

III. OPTIMAL POTENTIAL AND REPLICA-SYMMETRY STABILITY CONDITION

Given any modulation V^*, the typical overlap R obtained through the minimization of a differentiable potential V may be determined as a function of α by solving Eqs. (14). The result is consistent if condition (16) is verified. In this section, we are interested in the *best* performances that may be expected. Recently, a general expression for the optimal potential allowing to find the solution with maximum overlap $R_{\rm opt}$ has been deduced [4]. This *optimal potential* $V_{\rm opt}$ depends implicitly on α through $R_{\rm opt}(\alpha)$, and on the probability distribution P^* through the modulation V^*. It was obtained under the assumption of replica symmetry, which has been shown to be correct for the particular cases investigated so far. In fact, the stability condition of replica symmetry for optimal learning is verified whenever the slope of the learning curves is positive, as will be shown below. For the sake of completeness, we first describe an alternative derivation of the optimal potential. Following the same lines we used for supervised learning [6], $V_{\rm opt}$ is determined through a functional maximization of R, given by Eq. (14), with respect to V at constant α. As discussed in [6], the parameter c sets the energy units and may be arbitrarily chosen. We used $c=1$ throughout, without any lack of generality. After a straightforward calculation we obtain that the optimal overlap $R_{\rm opt}$ is given by the inversion of

$$\alpha(R_{\mathrm{opt}})=R_{\mathrm{opt}}^2\left\{\int_{-\infty}^{+\infty}Dt\,\frac{\left[\int Dz\;z\;\exp(-V^*(\lambda))\right]^2}{\int Dz\;\exp(-V^*(\lambda))}\right\}^{-1}, \tag{17}$$

where λ, given by Eq. (12), reads $\lambda\equiv z\sqrt{1-R_{\mathrm{opt}}^2}+R_{\mathrm{opt}}\,t$. Notice that Eq. (17) may be not invertible, i.e., $R_{\mathrm{opt}}(\alpha)$ may be multivalued. In this case, the correct solution has to be selected.

V_{opt} is determined through the integration of

$$V'_{\mathrm{opt}}(\gamma_{\mathrm{opt}}(t))=\frac{1-R_{\mathrm{opt}}^2}{R_{\mathrm{opt}}^2}\,\frac{d}{dt}\left[\ln\int_{-\infty}^{+\infty}Dz\;\exp(-V^*(\lambda))\right], \tag{18}$$

where the argument of V'_{opt} is given by the saddle-point equation (13) with $c=1$, i.e.,

$$\gamma_{\mathrm{opt}}(t)=t-V'_{\mathrm{opt}}(\gamma_{\mathrm{opt}}(t)). \tag{19}$$

Since R is parametrized by α, the cost function leading to optimal performance is different for different training set sizes.

Equations (17) and (18) were previously derived by Van den Broeck and Reiman [7], who showed that the typical overlap R_{b} of Bayesian learning satisfies the same equation (17) as R_{opt}. However, this only guarantees that Bayesian learning is optimal if Eq. (17) is invertible. In that case its unique solution is $R_{\mathrm{b}}=R_{\mathrm{opt}}$. Otherwise, as is discussed in the example of Sec. IV, solutions with $R_{\mathrm{opt}}>R_{\mathrm{b}}$ may exist.

The results derived so far are valid under the replica-symmetry hypothesis, and must thus satisfy Eq. (16). Taking Eqs. (17) and (19) into account, a cumbersome but straightforward calculation gives

$$1-\alpha\int_{-\infty}^{+\infty}Dt\;[\gamma'(t;c)-1]^2\int_{-\infty}^{+\infty}Dz\;\exp[-V^*(\lambda)]$$
$$=\frac{R_{\mathrm{opt}}^2(1-R_{\mathrm{opt}}^2)}{\alpha}\,\frac{d\alpha(R_{\mathrm{opt}})}{dR_{\mathrm{opt}}^2}. \tag{20}$$

Therefore, in the case of optimal learning, the necessary condition of replica-symmetry stability (16) *is equivalent* to the natural requirement that the learning curve $R_{\mathrm{opt}}(\alpha)$ is an increasing function of the fraction of examples α for $R_{\mathrm{opt}}\neq 0,1$. This relation, which does not seem to have been noticed before, is independent of the distribution (1) from which the data set is sampled.

In the cases where the analytic function α (R_{opt}) given by Eq. (17) is not invertible, only the branches with positive slope have to be considered, as they trivially satisfy the replica-symmetry condition. Examples of such a behavior are shown in the next section.

Hence, given any modulating function V^* sufficiently derivable, as far as $R_{\mathrm{opt}}\neq 0,1$ there exists an optimal potential $V_{\mathrm{opt}}(\gamma)$, consistent with the assumptions of the replica calculation, which depends implicitly on α through $R_{\mathrm{opt}}(\alpha)$, and on V^*. The minimum $\mathbf{J}^*$ of the corresponding energy (6) maximizes the overlap R between $\mathbf{J}^*$ and the symmetry-breaking direction $\mathbf{B}$.

The development of α (R_{opt}) for small R_{opt} shows that $R_{\mathrm{opt}}>0$ for all $\alpha>0$ if and only if $\langle\lambda\rangle\neq 0$. In that case, for $\alpha\ll 1$, $R_{\mathrm{opt}}\approx\langle\lambda\rangle\sqrt{\alpha}$, as with Hebb's learning rule [4]. If $\langle\lambda\rangle=0$, two different behaviors may arise: either a continuous transition from $R_{\mathrm{opt}}=0$ to $R_{\mathrm{opt}}\sim\sqrt{\alpha-\alpha_c}$ occurs at $\alpha_c\equiv(1-\langle\lambda^2\rangle)^{-2}$, or the overlap jumps from $R_{\mathrm{opt}}=0$ to $R_{\mathrm{opt}}>0$ through a first order transition at $\alpha_1\leqslant\alpha_c$. In particular, if $\langle\lambda^2\rangle=1$, only a discontinuous transition may occur since $\alpha_c=+\infty$. Discontinuities between two finite values of R_{opt} also may arise for $\alpha>\alpha_c$. All these phase transitions appear in the two-cluster scenario that we analyze in the next section.

IV. A CASE STUDY: TWO-CLUSTER DISTRIBUTIONS

Consider the general two-Gaussian-clusters scenario, in which the modulation along the symmetry-breaking direction (2) is

$$P^*(\lambda;\rho,\sigma)=\frac{1}{2\sigma\sqrt{2\pi}}\sum_{\epsilon=\pm 1}\exp\left[-\frac{(\lambda+\epsilon\rho)^2}{2\sigma^2}\right]. \tag{21}$$

This distribution is a generalization of the one studied by Watkin and Nadal [8], who considered optimal learning for clusters with $\sigma=1$. If $\rho=0$, Eq. (21) corresponds to the single Gaussian scenario studied by Reimann *et al.* [4]. In this paper we investigate the complete phase diagram in the plane ρ,σ.

The first two moments of Eq. (21) are

$$\langle\lambda\rangle=\;0, \tag{22}$$

$$\langle\lambda^2\rangle=\;\rho^2+\sigma^2. \tag{23}$$

Thus, if $\sigma=1$ only distributions with $\langle\lambda^2\rangle>1$ are considered. The optimal solution in that case is close to the one obtained with a quadratic potential [8]. Quadratic potentials detect the direction extremizing the variance of the training set, which we call *variance learning*. We show below that the optimal overlap may be much larger than the one obtained through variance learning if the clusters have $\sigma<1$.

Introducing the expression of V^* obtained from Eqs. (21) and (2) into Eq. (17) gives α as a function of R_{opt}. It turns out that, for some values of α, this function has three different roots for $R_{\mathrm{opt}}(\alpha)$, as is apparent in Figs. 1 and 2. The one lying on the branch with negative slope violates the assumption of replica symmetry. The two others correspond to minimas of the corresponding free energies. Figures 1, 2, and 3 show the optimal learning curves for several values of ρ and σ in the range not investigated before. The two branches $R_{\mathrm{opt}}(\alpha)$ with positive slope that satisfy condition (16), and the dotted line of negative slope (inconsistent with the assumption of replica symmetry), are presented for illustration. The value of α at which the jump from one branch to the other occurs is discussed in the next section. The performance obtained through learning with simple quadratic po-

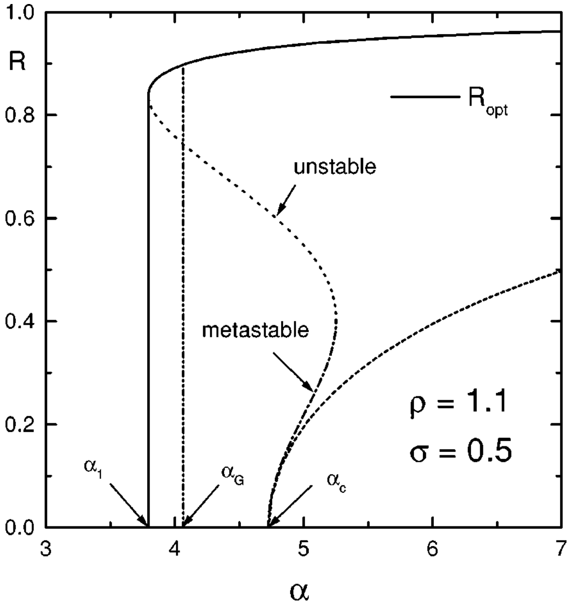

FIG. 1. Learning curves for the two-cluster scenario, for cluster parameters corresponding to the lowest small square of Fig. 6. Full line, optimal learning; dash-dotted lines, lower branch of metastable solutions to optimal learning. Also shown is the replica-symmetry unstable curve (dotted line). The lowest dashed line corresponds to learning with a quadratic potential (variance learning). Here, $\alpha_1 = 3.79$, $R_{\rm opt}(\alpha_1) = 0.84$; the Bayesian first order transition occurs at $\alpha_{\rm G} = 4.07$, $R_{\rm opt}(\alpha_{\rm G}) = 0.9$; the critical α for variance learning is $\alpha_c = 4.73$.

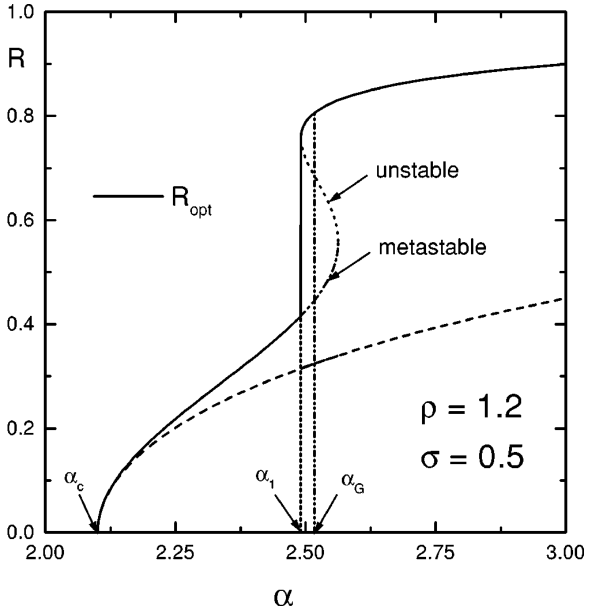

FIG. 2. Learning curves for the two-cluster scenario, for cluster parameters corresponding to the central small square of Fig. 6. Full line, optimal learning; dash-dotted lines, lower branch of metastable solutions to optimal learning. Also shown is the replica-symmetry unstable curve (dotted line). The lowest dashed line corresponds to learning with a quadratic potential (variance learning). Here, $\alpha_1 = 2.49$, $R_{\rm opt}(\alpha_1) = 0.76$; the Bayesian first order transition occurs at $\alpha_{\rm G} = 2.52$, $R_{\rm opt}(\alpha_{\rm G}) = 0.81$; the critical α for variance learning is $\alpha_c = 2.10$.

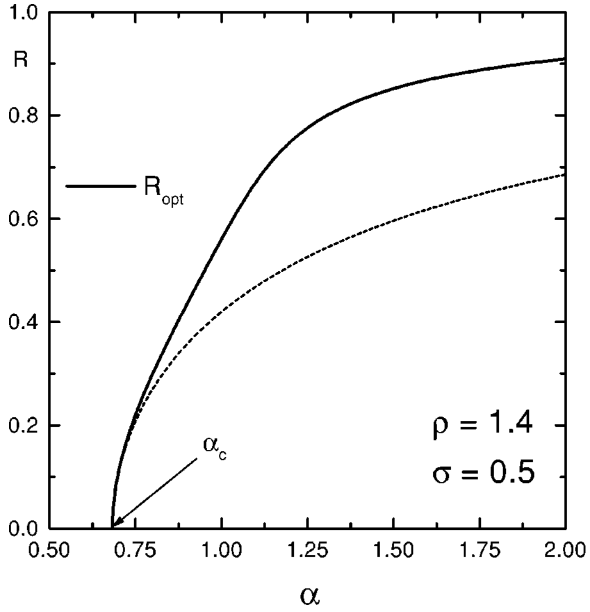

FIG. 3. Optimal learning curves (full line) for the two-cluster scenario, for cluster parameters corresponding to the upper small square of Fig. 6. The lowest dashed line corresponds to learning with a quadratic potential (variance learning). Here, $\alpha_c = 0.68$.

tentials is also presented, to show the dramatic improvement of optimal learning with respect to variance learning for double clusters with $\sigma < 1$.

V. BAYESIAN VERSUS OPTIMAL SOLUTIONS

As pointed out in Sec. III, Eq. (17) may be deduced in two different ways: through the determination of the Bayesian learning performance, or through functional optimization. This procedure yields of a cost function for each training set size α whose minimum gives the solution with maximal overlap.

The Bayesian solution to the learning problem is given by the average of solutions sampled with Gibbs' probability. A simple argument [8] shows that the typical Bayesian performance satisfies $R_{\rm b} = \sqrt{R_{\rm G}}$, where $R_{\rm G}$ is the typical overlap between a solution drawn with probability (5) and the symmetry-breaking direction $\mathbf{B}$. $R_{\rm G}$ minimizes the free energy with potential $V(\gamma) = V^*(\gamma)$ at inverse temperature $\beta = 1$ [8,7].

As Eq. (17) is satisfied both by $R_{\rm b}$ and $R_{\rm opt}$, it is tempting to conclude that Bayesian learning is optimal. If Eq. (17) has a unique solution, this is obviously the case. However, Eq. (17) may not be invertible. This arises in the two-cluster scenario presented in the preceding section, where two branches of solutions consistent with the assumption of replica symmetry exist for some values of α. In the case of Bayesian learning, these branches result from the fact that Gibbs' free energy has two local minima as a function of R. $R_{\rm G}$, the thermodynamically stable state, corresponds to the absolute minimum. When α changes, $R_{\rm G}$ jumps from one branch to the other through a first order phase transition at $\alpha = \alpha_{\rm G}$, where both minima have the same free energy [12]. Therefore the Bayesian solution, which is the average of the solutions sampled with Gibbs' probability, presents a jump at the same value $\alpha_{\rm G}$ as Gibbs' performance. Thus, the meta-

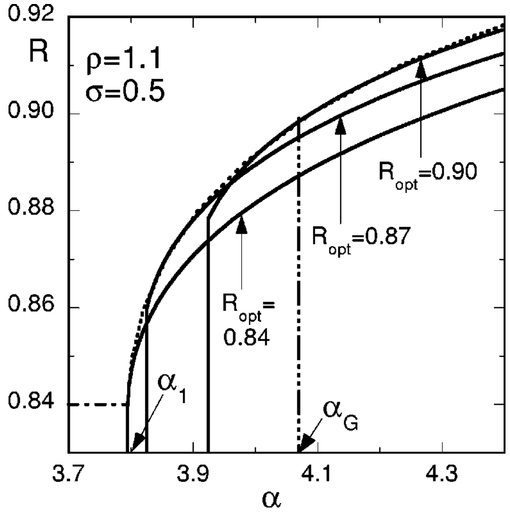

FIG. 4. Learning curves for $\rho=1.1$, $\sigma=0.5$ obtained with the optimal potentials corresponding to $R_{\rm opt}=0.84$, $R_{\rm opt}=0.87$, and $R_{\rm opt}=0.90$ (full lines). Only the solutions consistent with the replica-symmetry hypothesis are shown. Dotted lines: optimal solution.

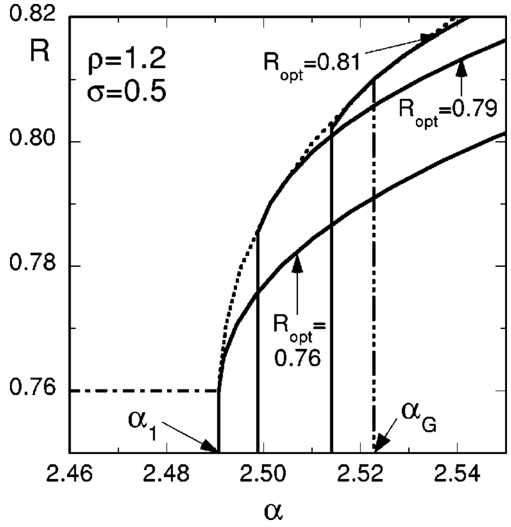

FIG. 5. Learning curves for $\rho=1.2$, $\sigma=0.5$ obtained with the optimal potentials corresponding to $R_{\rm opt}=0.76$, $R_{\rm opt}=0.79$ and $R_{\rm opt}=0.81$ (full lines). Only the solutions consistent with the replica- symmetry hypothesis are shown. Dotted lines: optimal solution.

stable states of higher performance than $R_{\rm b}$, which exist for $\alpha<\alpha_{\rm G}$, cannot be obtained through Bayesian learning.

On the other hand, in Sec. III we determined optimal potentials whose minimization allows to obtain performance $R_{\rm opt}$. These potentials exist for all the pairs $(\alpha, R_{\rm opt}(\alpha))$ lying on the monotonically increasing branches of $R_{\rm opt}(\alpha)$, which satisfy the hypothesis of replica symmetry. Potentials allowing to reach the performances of the upper (Gibbs-metastable) branch thus exist. It should be noticed that we cannot determine the position of the jump of $R_{\rm opt}$ through the comparison of the free energies corresponding to solutions on different branches at the same α, as was done to determine $\alpha_{\rm G}$, because a *different* potential has to be minimized for each pair $(\alpha, R_{\rm opt}(\alpha))$ and, as discussed in Sec. III, these potentials are measured in the arbitrary units determined by our choice $c=1$.

In order to clarify this problem, we studied the performance of the minima of the optimal potentials. In fact, the properties of each of the potentials $V_{\rm opt}(\lambda)$ may be determined for any value of α (besides the value for which it has been optimized) in the same way as those of other *ad hoc* potentials, by solving numerically Eq. (14). Figures 4 and 5 present several learning curves $R(\alpha)$ obtained with potentials $V_{\rm opt}$ optimized for overlaps lying on the upper metastable branch of Gibbs' learning. They correspond to the same clusters' parameters as Figs. 1 and 2. Each learning curve is tangent to the optimal learning curve at the point $(\alpha(R_{\rm opt}), R_{\rm opt})$ at which the potential was determined. This result holds in particular for all the points lying on the high-performance metastable branch of Bayesian learning, i.e., for $\alpha_1<\alpha<\alpha_{\rm G}$. It is important to point out that the free energy (11) presents a *unique* replica-symmetric minimum as a function of R for all these potentials. Thus, these results show that the corresponding optimal potentials $V_{\rm opt}$ allow to select, among the metastable states of Gibbs learning, the one of largest overlap. In particular, the Gibbs' metastable states in the upper branch for $\alpha<\alpha_{\rm G}$ are learnable through the minimization of the corresponding optimal potential. Thus, in the range $\alpha_1<\alpha<\alpha_{\rm G}$, Bayesian learning is not optimal. This surprising behavior may arise whenever the curve $R_{\rm G}(\alpha)$ of Gibbs learning presents first order phase transitions.

It is worth noting that, besides the solutions that verify the replica-symmetric condition (16), solutions unstable under replica-symmetry breaking with smaller R and slightly higher free energy also exist. The nature of these states is very different from that of the metastable states of Gibbs learning. Whether the typical performance in the case of the double cluster distributions is the one described by the replica-symmetric solution or not remains an open problem.

VI. THE PHASE DIAGRAM

In this section we describe, on the ρ-σ plane, all the possible learning phases that may arise in unsupervised learning within the two-Gaussian-cluster scenario. As shown in Fig. 6, depending on the values of ρ and σ, qualitatively different behaviors of the learning curves $R_{\rm opt}(\alpha)$ may appear. They are correlated with the form of the corresponding optimal potentials.

The regions marked with an ''S'' are regions of variance-type learning: the optimal potential is a single well with $V_{\rm opt}\rightarrow+\infty$ for $\lambda\rightarrow\pm\infty$ if $\sigma^2<1$, and $V_{\rm opt}\rightarrow-\infty$ for $\lambda\rightarrow\pm\infty$ if $\sigma^2>1$. In these regions, the learning curves increase monotonically with α, starting at $\alpha_c=|\langle\lambda^2\rangle-1|^{-2}$, as for quadratic potentials [4].

For parameter values outside the ''S'' regions, $V_{\rm opt}\rightarrow+\infty$ for $\lambda\rightarrow\pm\infty$, even in the large variance region $\langle\lambda^2\rangle>1$ where naively one would expect the potential to have the same asymptotic behavior as for $\sigma^2>1$. Depending on the value of $R_{\rm opt}$, the optimal potential may be a double-well function of the local field γ. In the latter case, the optimal learning strategy looks for structure in the data distribution rather than for directions extremizing the variance. This is more striking on the line $\langle\lambda^2\rangle=1$ corresponding to distribu-

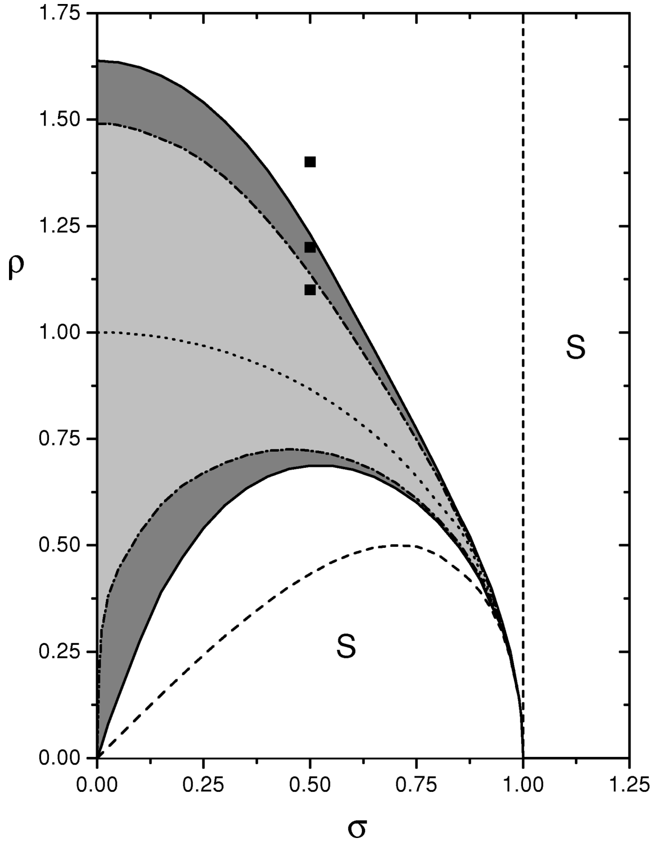

FIG. 6. Phase diagram of the two-cluster scenario. The three small squares correspond to the learning curves of Figs. 1, 2, and 3.

tions with the same second moment in all the directions. On this line, variance learning is impossible and $\alpha_c = \infty$. However, in the entire light-gray region including this line, performant learning is achieved if the adequate potential is minimized. The optimal overlap presents jumps from $R_{\rm opt} = 0$ to finite R at a fraction of examples $\alpha < \alpha_c$. In the high-performance branch, the optimal potential is double-well, with the two minima close to $\pm \rho$, as shown in Fig. 7. Thus, the potential is sensitive to the two-cluster structure, and its minimization results in high performance learning. For ρ and σ in the dark-gray regions, a first order transition to large R also takes place, but for $\alpha > \alpha_c$. Below the transition, optimal learning is mainly controlled by the variance of the training set.

In the white regions on both sides of the dark-gray ones, no first order phase transitions to high performance learning occur as a function of α. In the white region just below the dark-gray one, the potential changes smoothly from a single to a double well with increasing $R_{\rm opt}$. The two minimas appear at $\gamma = 0$, and move away with increasing $R_{\rm opt}$, as shown in Fig. 8. However, as far as these minimas are not sufficiently apart, $R_{\rm opt}$ remains close to the values obtained with simple quadratic potentials. Conversely, in the upper white region, which corresponds to $\langle \lambda^2 \rangle \gg 1$, the minima of the optimal potential are far apart, in a region of large local fields, where the patterns' distribution is vanishingly small. Thus, in the range of pertinent values of γ the potential is concave ($V''_{\rm opt} < 0$), and here also, as in the lower white region, the values of $R_{\rm opt}$ are close to those obtained with quadratic potentials [4].

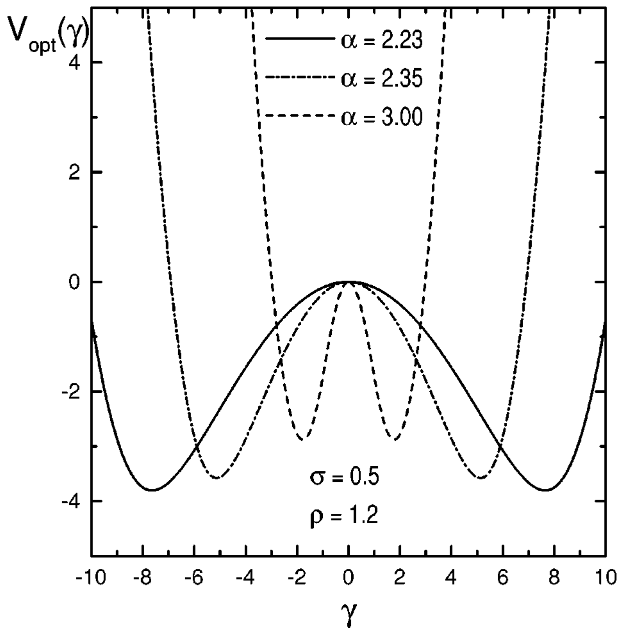

FIG. 7. Potentials for optimal learning in the grey regions of the phase diagram, showing the evolution of the separation between minima with α.

VII. CONCLUSION

Learning the symmetry-breaking direction of a distribution of patterns with axial symmetry in high dimensions is a difficult problem. In this paper we determined the optimal performances that may be reached if the patterns distribution has a double-cluster structure in the symmetry-breaking direction. Depending on the clusters' size and separation, the learning curves may present several phases with increasing α, including novel first order transitions from low-performance variance learning to high-performance structure

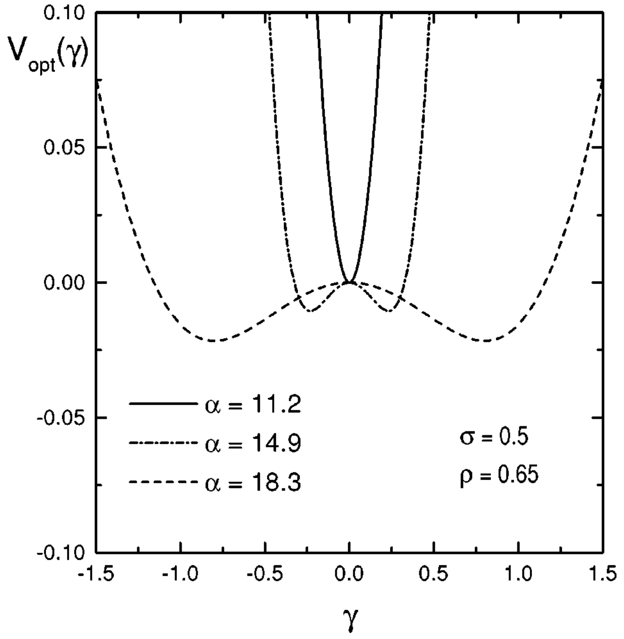

FIG. 8. Potentials for optimal learning in the white regions of the phase diagram, showing the appearance of the two minima that get farther apart with increasing α.

detection. We showed that when the optimal learning curves present such discontinuities, Bayesian learning may be not optimal. These results rely on the assumption that the solution with replica symmetry is the absolute minimum of the free energies studied. Although we showed that our solutions satisfy the replica symmetry stability condition, we cannot rule out the existence of states of lower energy, but having broken replica symmetry.

ACKNOWLEDGMENTS

We thank Chris Van den Broeck and Mauro Copelli for helpful discussions at the early stages of this work.

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (John Wiley and Sons, New York, 1973).

[2] B. Ripley, *Pattern Recognition and Neural Networks* (Cambridge University Press, Cambridge, 1996).

[3] T. Kohonen, *Self-Organizing Maps* (Springer, Berlin, 1995).

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

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

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

[7] C. Van den Broeck and P. Reimann, Phys. Rev. Lett. **76**, 2188 (1996).

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

[9] P. Reimann, C. Van den Broeck, and G. J. Bex, J. Phys. A **29**, 3521 (1996).

[10] M. Biehl and A. Mietzner, Europhys. Lett. **24**, 421 (1993).

[11] M. Biehl and A. Mietzner, J. Phys. A **27**, 1885 (1994).

[12] M. Copelli (private communication).