Unsupervised and supervised learning: Mutual information between parameters and observations

Didier Herschkowitz* and Jean-Pierre Nadal
Laboratoire de Physique Statistique de l'E.N.S.,† Ecole Normale Supérieure, 24, rue Lhomond-75231 Paris Cedex 05, France
(Received 3 August 1998; revised manuscript received 13 October 1998)

We study the mutual information between parameter and data for a family of supervised and unsupervised learning tasks. The parameter is a possibly, but not necessarily, high-dimensional vector. We derive exact bounds and asymptotic behaviors for the mutual information as a function of the data size and of some properties of the probability of the data given the parameter. We compare these exact results with the predictions of replica calculations. We briefly discuss the universal properties of the mutual information as a function of data size. [S1063-651X(99)00403-1]

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

I. INTRODUCTION

We consider the very general problem of finding the structure underlying a set of data, also called *examples, patterns*, or *training set*. The parametric approach assumes that the structure of the probability density function (PDF) the patterns have been sampled from is known. Only its parameters have to be determined given the examples. We consider both *supervised* and *unsupervised* learning paradigms within the same framework of parameter estimation. The process of determining the parameters is called *unsupervised learning* when the goal is to estimate the probability distribution from the observed data only. In the case of *supervised learning* one is given additional information about the data, that is, each training example is labeled. Several type of labels can be specified, and we will consider two kinds of labels: a *cluster* label, which, in the case of a mixture density, indicates from which PDF the pattern has been produced (to which cluster the pattern belongs); and a *class* label, which is a classification of the observed pattern (e.g., it is the binary classification produced by a teacher perceptron). In all these cases, the PDF of the data and/or the labels can be characterized by a parameter, a vector in a possibly high-dimensional space, and the goal is to estimate the parameter from the observed data.

Recent results on parameter estimation show that the *mutual information* between data and parameter is a relevant tool to derive optimal performances [1–5]. Based on Shannon information quantities (see, e.g., [6]), it quantifies the intuitive idea that our knowledge of the parameter value is limited if we have a finite amount of data. This quantity is independent of any specific algorithm used to estimate the parameter. The best possible estimator of the parameter is the one that is able to extract all this information hidden in the data. If such an estimator exists, its performance should then be related to the mutual information. In fact, one should be able to compute the best possible performance from the mutual information without knowing in advance which algorithms will allow us to achieve this performance. In the context of supervised learning, the mutual information is shown in [7] to have, within the Bayesian framework, the meaning of a cumulative entropic error.

In addition, any model of parameter estimation can be interpreted in the neural coding framework, *via* the duality shown in [8]: the parameter plays the role of the stimulus and each pattern of the training set is then the activity of a coding cell. In this context, the mutual information characterizes the quality of the coding system. Its maximization has been proposed as a possible principle for neural organization in living animals (see, e.g., [9,10]) and is related to coding based on redundancy reduction (see, e.g., [11,12]).

All this motivates the study of the mutual information between data and parameter, which we do in the present paper for a family of unsupervised and supervised learning tasks. We address the question of the behavior of the mutual information as a function of the dimension of parameter space, size of data set, and properties of the PDF generating the data given the parameter. It is already known that universal scaling laws exist for the asymptotic performance of estimators; e.g., the generalization error decreases as p/N for $p \gg N$ in the case of smooth distributions [13]. Our main concern will be to see what types of universal properties exist for the mutual information.

Some of the results we present are very general, but the detailed calculations and analysis will be done for a family of models where the data structure can be characterized by a single symmetry-breaking orientation **B** along which the pattern distribution is nonuniform. Models of this family have been studied extensively with the *replica method* in the framework of statistical mechanics [14–17]. As we will see, the *self-averaged* free energy associated to *Gibbs learning* is directly related to the mutual information; hence, it contains the typical properties of the system.

After introducing the general framework of unsupervised learning (Sec. II), and introducing the mutual information between data and parameters, we show how the computation

*Electronic address: herschko@lps.ens.fr nadal@lps.ens.fr http://www.lps.ens.fr/~risc/rescomp/

†The laboratory is associated with the CNRS (URA 1306), ENS, and the Universities Paris VI and Paris VII.

of the information gain in a supervised learning task can be reduced to that of the mutual information in a related unsupervised problem. As a result we can then work on a family of parameter estimation tasks that can be seen as unsupervised learning problems, some of them having in addition an alternative interpretation as a supervised learning problem.

For this family of models we present first exact results (Sec. III): a linear upper bound valid for any data size and any parameter dimension; an upper bound for the case of supervised learning also for any N and p; the asymptotic behavior of the mutual information for smooth distributions in the limit of the data size p very large compared to N, the parameter dimension, which here is not necessarily large. In the latter case we make use of general results relating the mutual information to the *Fisher information* [1,4,5]. Finally we make use of tools, introduced in [7] in the context of the standard supervised learning framework, to derive upper and lower bounds for both unsupervised learning and supervised learning in the case of patterns correlated with the parameter. A direct application of the techniques of [7] provide the behavior of the mutual informations in the large data size limit. In addition, we show that one can get also explicit upper and lower bounds valid in the large N limit at any given value of $\alpha = p/N$ (the derivation of these bounds is detailed in the Appendix).

Next, in Sec. IV with the relationship mentioned above and to be detailed below between the mutual information and the free energy, we make use of replica calculations already published, giving their interpretation in terms of mutual information. We also present new results, on both previously studied and not previously studied models. These replica calculations are expected to be valid in the case where the number p of observed patterns is of order of the dimension N of the parameter space, in the limit of very large N. We consider first unsupervised learning, with both smooth and discontinuous PDF, and we then deduce the relevant information quantities for the associated supervised learning models. We compare the predictions of the replica calculations made under the replica symmetry ansatz with the exact bounds and asymptotic behaviors presented in Sec. III. In Sec. V we illustrate all these results on specific models. Finally in Sec. VI we use information quantities to derive bounds on performance of specific estimators. In the Conclusion we discuss general features of parameter estimation in view of the results obtained on the particular class of models studied in the present paper.

II. MUTUAL INFORMATION FOR A PARAMETER ESTIMATION TASK

A. Model family

We first introduce the general setup from the point of view of *unsupervised learning*. We assume that a set of patterns $\mathbf{X} = \{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$ is generated by p independent samplings from a nonuniform probability distribution

$$\mathcal{P}(\mathbf{X}|\mathbf{B}) = \prod_{\mu=1}^{p} p(\boldsymbol{\xi}^\mu|\mathbf{B}), \tag{1}$$

where $\mathbf{B} = \{B_1, \ldots, B_N\}$ represents the symmetry-breaking orientation. For the family of models we are considering, the probability of a given pattern ξ can always be written in the form

$$p(\boldsymbol{\xi}|\mathbf{B}) = \frac{1}{\sqrt[N]{2\pi}} \exp\left(-\frac{\boldsymbol{\xi}^2}{2} - V(\lambda) \right), \tag{2}$$

where N is the dimension of the space and

$$\lambda = \mathbf{B}\cdot\boldsymbol{\xi}/\|\mathbf{B}\| \tag{3}$$

is the overlap between the pattern and the direction. According to Eq. (2), the patterns have normal, unit variance distribution, i.e., $\exp(-x^2/2)/\sqrt{2\pi}$ onto the $N-1$ directions orthogonal to $\mathbf{B}$ and the distribution of the overlap in the symmetry-breaking direction is given by

$$P(\lambda) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\lambda^2}{2} - V(\lambda) \right). \tag{4}$$

The potential $V(\lambda)$ characterizes the structure of the data in the symmetry-breaking direction. In particular, if $V(\lambda) \equiv 0$, the patterns are uniformly distributed in all the directions and no special orientation can be detected. The potential $V(\lambda)$ satisfy the normalization condition

$$\int D\lambda \, \exp[-V(\lambda)] = 1, \tag{5}$$

where $D\lambda = d\lambda \, \exp(-\lambda^2/2)/\sqrt{2\pi}$ is the Gaussian measure. Here and in the following, when not explicitly written, integrals go from $-\infty$ to $+\infty$.

As justified within the Bayesian and statistical physics frameworks, one has to consider a *prior* distribution on the parameter space, $\rho(\mathbf{B})$. Convenient choices for detailed calculations in specific models are, e.g., the Gaussian prior $\rho(\mathbf{B}) = \exp(-\mathbf{B}^2/2)/\sqrt[N]{2\pi}$ or the uniform distribution on the unit sphere. From the point of view of inference, there is, however, an optimal prior, the one that maximizes the mutual information [1,4].

B. Unsupervised learning

The mutual information $I(\mathbf{X};\mathbf{B})$ between the examples and the parameter (here the symmetry breaking direction $\mathbf{B}$) is (see, e.g., [6])

$$I(\mathbf{X};\mathbf{B}) = H(\mathbf{X}) - H(\mathbf{X}|\mathbf{B}), \tag{6}$$

where

$$H(\mathbf{X}) = -\int d\mathbf{X} \, \mathcal{P}(\mathbf{X}) \ln \mathcal{P}(\mathbf{X}) \tag{7}$$

is the pattern entropy according to their probability

$$\mathcal{P}(\mathbf{X}) = \int d\mathbf{B} \, \rho(\mathbf{B}) \mathcal{P}(\mathbf{X}|\mathbf{B}) \tag{8}$$

and

$$H(\mathbf{X}|\mathbf{B}) = -\int d\mathbf{B} \, d\mathbf{X} \rho(\mathbf{B}) \mathcal{P}(\mathbf{X}|\mathbf{B}) \ln \mathcal{P}(\mathbf{X}|\mathbf{B}) \tag{9}$$

is the equivocation: the pattern entropy conditional to **B**, averaged over the parameter distribution. Here and in the following, the logarithm are neperian. The unit of the mutual information is then the *nat*. The mutual information represents the mean amount of information the data **X** convey about the variable **B**.

For the model family we are considering defined by Eq. (2), the mutual information can be rewritten

$$I(\mathbf{X};\mathbf{B}) = -p\langle V(\lambda)\rangle - \langle\langle \ln Z(\mathbf{X})\rangle\rangle, \tag{10}$$

where

$$Z(\mathbf{X}) \equiv \int d\mathbf{B}\, \rho(\mathbf{B}) \exp\left(-\sum_{\mu=1}^{p} V(\lambda^\mu) \right). \tag{11}$$

Here and in all this paper the brackets $\langle\cdots\rangle$ stand for the average over the overlap distribution $P(\lambda)$, Eq. (4), and $\langle\langle\cdots\rangle\rangle$ the average over the pattern distribution $\mathcal{P}(\mathbf{X})$, Eq. (8). One should note that $-\langle V(\lambda)\rangle$ is a positive quantity. In the statistical physics literature, the quantity $-\ln Z(\mathbf{X})$ is the ''free energy'' and Z the ''partition function.'' From related studies in the field of statistical physics of disordered systems, one expects the free energy to be a self-averaging quantity, that is $-(1/N)\ln Z(\mathbf{X}) \sim \langle\langle -(1/N)\ln Z(\mathbf{X})\rangle\rangle$ in the large N limit. This means that, in this limit, the properties of the system depend no more on the specific set of patterns **X** but on the patterns distribution $\mathcal{P}(\mathbf{X})$ only. It is interesting that it is precisely this quantity, the averaged free energy, that appears in the mutual information. This shows that, on one hand, it is indeed the mutual information that contains the typical behavior of the system, and on the other hand, that the mean free energy is a relevant quantity even for finite N.

A remark on our notation is in order. Since in the following we will consider relationships between the mutual information associated with different, but related, models, we will attach to the mutual information associated with each model a subscript referring to the particular probability with which the patterns have been generated. In particular, whenever considering a *smooth* potential (that is, such that V is as regular as needed), we will write the mutual information (10) associated with the model (2) as $I_P(\mathbf{X};\mathbf{B})$ where the subscript P refers to the smooth distribution $P(\lambda)$, Eq. (4).

C. Supervised learning

We now turn to the case of *supervised learning* tasks. We will consider two kinds of supervised learning: ''cluster learning'' and ''class learning.'' We show how they are related to smooth and discontinuous unsupervised learning tasks, respectively.

1. Cluster learning

We consider a mixture density made of two smooth PDF's such that the data will appear as two clusters symmetric about the origin: the symmetry-breaking direction is the direction of the axis joining the centers of the two clusters. To each cluster is associated a label $A = \pm 1$. Each pattern is generated in a two step procedure: first one chooses a cluster with equal probability and then the pattern is generated from the corresponding cluster distribution. Denoting by $\mathbf{A} = \{A^\mu\}_{\mu=1}^p$ the set of cluster labels, the model we are considering is thus

$$\mathcal{P}(\mathbf{A},\mathbf{X}|\mathbf{B}) = \prod_{\mu=1}^{p} \tfrac{1}{2}(\delta_{A^\mu,1} + \delta_{A^\mu,-1}) p(\boldsymbol{\xi}^\mu | A^\mu \mathbf{B}). \tag{12}$$

We assume $p(\boldsymbol{\xi}|\mathbf{B})$ to have a smooth overlap distribution $P(\lambda)$.

In the context of supervised learning, the patterns **X** are given with their cluster label $\mathbf{A} = \{A^\mu\}_{\mu=1}^p$. We will denote by $I_{AP}(\mathbf{A},\mathbf{X};\mathbf{B})$ the information the pair of variables $(\mathbf{A},\mathbf{X})$ gives about the symmetry-breaking axis **B**.

Now a pattern $\boldsymbol{\xi}$ coming from the cluster $p(\boldsymbol{\xi}|A\mathbf{B})$ gives the same amount of information about the direction **B** than a pattern $A\boldsymbol{\xi}$ coming from the cluster $p(A\boldsymbol{\xi}|\mathbf{B})$. One can thus proceed as if one was given the set of patterns $\{A^\mu \boldsymbol{\xi}^\mu\}$ generated from a single distribution $A = +1$: one is back to the unsupervised task with the smooth overlap distribution $P(\lambda)$.

This writes

$$I_{AP}(\mathbf{A},\mathbf{X};\mathbf{B}) = I_P(\mathbf{X};\mathbf{B}). \tag{13}$$

The direct proof is straightforward using $p(A\boldsymbol{\xi}|A\mathbf{B}) = p(\boldsymbol{\xi}|\mathbf{B})$.

If the cluster label A^μ is not provided, one has an unsupervised learning task equivalent to having the patterns generated from the symmetric and smooth mixture distribution:

$$\Sigma P(\lambda) \equiv \tfrac{1}{2}[P(\lambda) + P(-\lambda)]. \tag{14}$$

We will denote by $I_{\Sigma P}(\mathbf{X};\mathbf{B})$ the information conveyed by the patterns alone.

Another quantity of interest is the amount of information conveyed by the cluster labels about **B** when the patterns, generated with the probability (14), are known, that is, $I_{\Sigma P}(\mathbf{A};\mathbf{B}|\mathbf{X})$. From information theory one has that the information that the pair of variables $(\mathbf{A},\mathbf{X})$ gives about the symmetry breaking axis is equal to the information that the patterns alone gives about **B**, plus the information that **A** gives about **B** when the patterns are already known:

$$I_{AP}(\mathbf{A},\mathbf{X};\mathbf{B}) = I_{\Sigma P}(\mathbf{X};\mathbf{B}) + I_{\Sigma P}(\mathbf{A};\mathbf{B}|\mathbf{X}). \tag{15}$$

As we will see in Sec. IV, the left-hand side (lhs) of Eq. (13) and the first term of the right-hand side (rhs) can be computed with the replica technique. From these two calculations one then gets the second term in the rhs. Since the information is a positive quantity, from Eqs. (13) and (15) it follows that

$$I_{\Sigma P}(\mathbf{X};\mathbf{B}) \leqslant I_P(\mathbf{X};\mathbf{B}). \tag{16}$$

Equations (13) and (15) relating supervised and unsupervised informations are illustrated in Fig. 5 in the particular case of a Gaussian overlap distribution.

Note: If the (single cluster) distribution $p(\boldsymbol{\xi}|\mathbf{B})$ is symmetric about the origin, the two clusters are indistinguishable, and one has $I_{AP}(\mathbf{A},\mathbf{X};\mathbf{B}) = I_{\Sigma P}(\mathbf{X};\mathbf{B})$.

2. Class learning

We consider now that the patterns $\mathbf{X}=\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$ are generated by p independent samplings from the distribution $p(\boldsymbol{\xi}|\mathbf{B})$ defined as in Eq. (2) with a distribution $P(\lambda)$ associated with a symmetrical and smooth potential $V_P(\lambda)$. In addition, for each pattern a teacher provides a class label, that is a binary classification $S^\mu=\pm 1$. Since we are considering models with a single symmetry-breaking orientation, we assume that the $\mathbf{B}$ vector in the pattern distribution (2) also controls the classification according to

$$S^\mu=\mathrm{sgn}(\mathbf{B}\cdot\boldsymbol{\xi}^\mu). \tag{17}$$

Denoting by $\mathbf{S}=\{S^\mu\}_{\mu=1}^p$ the set of class labels, the model we are considering is thus

$$\mathcal{P}(\mathbf{S},\mathbf{X}|\mathbf{B})=\prod_{\mu=1}^{p}\Theta(S^\mu\boldsymbol{\xi}^\mu\cdot\mathbf{B})p(\boldsymbol{\xi}^\mu|\mathbf{B}). \tag{18}$$

We denote by $I_{PS}(\mathbf{S},\mathbf{X};\mathbf{B})$ the mutual information between the pair of variables $(\mathbf{S},\mathbf{X})$ and the parameter $\mathbf{B}$. It has to be noted that contrary to most supervised learning models previously studied, the patterns themselves carry information about the teacher (the symmetry-breaking direction).

As was pointed out in [16], the classification of the pattern $\boldsymbol{\xi}$ as S automatically implies that the pattern $S\boldsymbol{\xi}$ is classified as $+1$. The overlap distribution of a pattern $S\boldsymbol{\xi}$, denoted by ΘP, readily follows from the original one $P(\lambda)$:

$$\Theta P(\lambda)\equiv 2\Theta(\lambda)P(\lambda), \tag{19}$$

where $\Theta(\lambda)$ is the Heavyside distribution. The corresponding potential is

$$V_{\Theta P}(\lambda)=\infty \quad \text{for } \lambda<0,$$

$$V_{\Theta P}(\lambda)=V_P(\lambda)-\ln 2 \quad \text{for } \lambda>0. \tag{20}$$

The task is thus equivalent to an unsupervised learning task with the discontinuous overlap distribution $\Theta P(\lambda)$. This writes

$$I_{PS}(\mathbf{S},\mathbf{X};\mathbf{B})=I_{\Theta P}(\mathbf{X};\mathbf{B}). \tag{21}$$

Note: this equality is true only when the overlap distribution $P(\lambda)$ is symmetric. Otherwise patterns with classification $S=+1$ and $S=-1$ convey different information about $\mathbf{B}$, and the supervised information $I_{PS}(\mathbf{S},\mathbf{X};\mathbf{B})$ is not in general directly related to an unsupervised problem as in Eq. (21).

If the class label is not given, one is back to the unsupervised learning problem with smooth potential $P(\lambda)$, for which the information is $I_P(\mathbf{X};\mathbf{B})$. The additional amount of information given by the class labels is noted $I_P(\mathbf{S};\mathbf{B}|\mathbf{X})$. Similarly to Eqs. (15) and (16) one has

$$I_{PS}(\mathbf{S},\mathbf{X};\mathbf{B})=I_P(\mathbf{X};\mathbf{B})+I_P(\mathbf{S};\mathbf{B}|\mathbf{X}) \tag{22}$$

and

$$I_P(\mathbf{X};\mathbf{B})\leqslant I_{\Theta P}(\mathbf{X};\mathbf{B}). \tag{23}$$

Equations (21) and (22) relating supervised and unsupervised informations are illustrated in Figs. 8 and 9 for different choices of the overlap distribution.

One may note that we could have considered class learning as a particular case of cluster learning where the single cluster distribution is given by $2\Theta(\lambda)P(\lambda)$. However, what justifies distinguishing the two types of supervised learning is that, as we have seen above, supervised cluster learning is related through Eq. (13) to *smooth* unsupervised learning, and class learning is related through Eq. (21) to *discontinuous* unsupervised learning.

In the following, thanks to these relationships between supervised and unsupervised learning tasks, we will indifferently take either the supervised or the unsupervised point of view according to which is the more convenient or relevant to the current discussion.

III. EXACT BOUNDS AND ASYMPTOTIC BEHAVIORS

We derive now exact bounds and exact asymptotic behaviors for the mutual information. Some of these results are specific to the form (2) of the probability distribution and other are more general. We begin with a linear upper bound.

A. Linear bound

The mutual information, a positive quantity, cannot grow faster than linearly in the amount of data p. Indeed, it is easy to show that

$$I(\mathbf{X};\mathbf{B})\leqslant pI_1(\boldsymbol{\xi};\mathbf{B}), \tag{24}$$

where I_1 is the mutual information between the parameter and a single example (one can check that pI_1-I can be written as a Kullback divergence, a quantity always nonnegative). However, I_1 cannot be easily computed in the general case. We derive the simpler linear upper bound:

$$I(\mathbf{X};\mathbf{B})\leqslant -p\langle V(\lambda)\rangle. \tag{25}$$

This relation is true for all p and all N. We prove the inequality for the case $\langle\lambda\rangle=0$. The extension to the case $\langle\lambda\rangle\neq 0$ is straightforward. As we will see, for the particular family of models that we are considering, in the large N limit this upper bound becomes in fact identical to the bound pI_1.

In the expression (6) of the mutual information, the computation of the second term, the equivocation $H(\mathbf{X}|\mathbf{B})$, is straightforward. One gets

$$H(\mathbf{X}|\mathbf{B})=\frac{pN}{2}\ln(2\pi e)+\frac{p}{2}(\langle\lambda^2\rangle-1)+p\langle V\rangle. \tag{26}$$

The first term on the rhs of Eq. (6), that is, the entropy of the data, $H(\mathbf{X})$, is the quantity difficult to compute. However, one can upperbound this entropy by the entropy of the Gaussian with the same covariance matrix. The covariance matrix of the data is easily obtained as

$$\langle\langle\xi_i^\mu\xi_j^\nu\rangle\rangle=\delta_{\mu\nu}(\delta_{ij}+(\langle\lambda^2\rangle-1)\overline{B_iB_j/\|\mathbf{B}\|^2}), \tag{27}$$

where $\overline{(.)}$ denotes the average over the parameter distribution. One then has

$$H(\mathbf{X}) \leqslant \frac{pN}{2}\ln(2\pi e) + \frac{p}{2}\sum_{i=1}^{N} \ln[1+(\langle\lambda^2\rangle-1)\eta_i], \tag{28}$$

where η_i are the eigenvalues of the matrix $\overline{B_i B_j/\|\mathbf{B}\|^2}$. Putting Eqs. (28) and (26) together with Eq. (6), one gets the linear bound

$$I(\mathbf{X};\mathbf{B}) \leqslant -p\langle V(\lambda)\rangle - \frac{p}{2}(\langle\lambda^2\rangle-1) + \frac{p}{2}\sum_{i=1}^{N} \ln[1+(\langle\lambda^2\rangle-1)\eta_i]. \tag{29}$$

Using the property $\ln(1+x)\leqslant x$ together with

$$1 = \sum_{i=1}^{N} \overline{B_i B_i/\|\mathbf{B}\|^2} = \sum_{i=1}^{N} \eta_i ,$$

one then gets the simpler bound (25).

In fact the bound (29) becomes identical to Eq. (25) in the asymptotic regime $N\to\infty$ whenever all the eigenvalues η_i are of the same order, that is, $1/N$. This is, in particular, true if the *prior* is spherically symmetric, in which case $\eta_i = 1/N$ for all $i=1,\ldots,N$. In these cases, for finite N, the bound (29) reads

$$I(\mathbf{X};\mathbf{B}) \leqslant -p\langle V(\lambda)\rangle - \frac{p}{2}(\langle\lambda^2\rangle-1) + \frac{pN}{2}\ln\left(1+\frac{\langle\lambda^2\rangle-1}{N}\right). \tag{30}$$

In the large N limit, keeping $\alpha = p/N$ fixed, one has then

$$\lim_{N\to\infty}\frac{1}{N} I(\mathbf{X};\mathbf{B}) \leqslant -\alpha\langle V(\lambda)\rangle. \tag{31}$$

From the relationship between mutual information and free energy, Eq. (10), this inequality (25) can also be written as

$$-\langle\langle \ln(Z)\rangle\rangle \leqslant 0 \tag{32}$$

that is, the mean free energy is always negative or null.

B. Asymptotic behavior and Fisher information

The asymptotic limit usually considered in the context of statistical parameter estimation is the one where the dimension of the parameter space, N, is given (and not necessarily large), and the number of examples p is large compared to the dimension N. For smooth structure, it has been proved [1,4,5] that, in that limit $p \gg N$, the mutual information increases as half the logarithm of the determinant of the *Fisher information matrix*. This matrix is a fundamental quantity in parameter estimation: its inverse is a bound on the covariance of any efficient estimator (Cramer-Rao bound, see, e.g., [6]). Hence, in this asymptotic limit of large data size, one has a simple and explicit link between the mutual information and the best possible performance of an estimator. For our model family, this asymptotic behavior of the mutual information reads

$$I(\mathbf{X};\mathbf{B}) \sim I_{\text{Fisher}} \quad \text{for } p \gg N$$

$$I_{\text{Fisher}} \equiv \frac{N}{2}\ln\left(\frac{p}{N}\langle V'^2(\lambda)\rangle\right), \tag{33}$$

where $V'(\lambda) = dV(\lambda)/d\lambda$. We will see in Sec. IV that this asymptotic behavior is correctly predicted by the replica calculation for smooth potentials, in the limit $N\to\infty$ first, then $\alpha = p/N \to \infty$. In the case of nonsmooth distributions, the Fisher information matrix does not exist (it is infinite). One can then expect a different asymptotic behavior for the mutual information, as suggested by the bound derived in the next section.

C. Bound on the class information

We show in this section that the mutual information between the class and the symmetry-breaking orientation given the pattern $I_P(\mathbf{S};\mathbf{B}|\mathbf{X})$ is bounded:

$$I_P(\mathbf{S};\mathbf{B}|\mathbf{X}) \leqslant \ln\Delta(p,N), \tag{34}$$

where

$$\Delta(p,N) \equiv \sum_{k=0}^{\min(p,N)} C_p^k \tag{35}$$

with $C_p^k = p!/[k!(p-k)!]$.

This bound and its proof are the same as for the information capacity of a perceptron studied in [18,8,19]. The argument is as follows. Since the class is a deterministic function of the parameter $\mathbf{B}$, when the pattern is given, the mutual information between the class labels and the parameter is equal to the conditional entropy of the labels given $\mathbf{X}$:

$$I_P(\mathbf{S};\mathbf{B}|\mathbf{X}) = -\sum_{\mathbf{S}} \langle\langle P(\mathbf{S}|\mathbf{X})\ln P(\mathbf{S}|\mathbf{X})\rangle\rangle, \tag{36}$$

where $P(\mathbf{S}|\mathbf{X}) = \int d\mathbf{B}\, P(\mathbf{B}|\mathbf{X})\Pi_{\mu=1}^{p}\Theta(S^\mu \mathbf{B}\cdot\boldsymbol{\xi}^\mu)$ with $P(\mathbf{B}|\mathbf{X}) = \rho(\mathbf{B})P(\mathbf{X}|\mathbf{B})/P(\mathbf{X})$. Let us call $\Delta(\mathbf{X}) \leqslant 2^p$ the number of realizable dichotomies, that is, the number of distinct configurations $\mathbf{S} = \{S^\mu\}_{\mu=1}^{p}$ for which there is at least one parameter $\mathbf{B}$ such that $S^\mu = \text{sgn}(\boldsymbol{\xi}^\mu\cdot\mathbf{B})$ for every $\mu = 1,\ldots,p$. The entropy of the distribution $P(\mathbf{S}|\mathbf{X})$ is maximum when every possible S has the same probability, that is $1/\Delta(\mathbf{X})$. Hence

$$I_P(\mathbf{S};\mathbf{B}|\mathbf{X}) \leqslant \langle\langle \ln\Delta(\mathbf{X})\rangle\rangle. \tag{37}$$

If the patterns are in ''general position,'' one basic result [20] is that $\Delta(\mathbf{X})$ is in fact independent of the particular sample $\mathbf{X}$, and depends only on p and N, being equal to $\Delta(p,N)$ defined in Eq. (35). As a result one then obtains the bound (34). If the patterns are not in general position, the bound remains valid because then $\Delta(\mathbf{X}) \leqslant \Delta(p,N)$.

In the limit $N\to\infty$ and $\alpha = p/N$ fixed, one has the asymptotic behavior

$$\lim_{N\to\infty}\frac{\ln\Delta(p,N)}{N} = \begin{cases} \alpha\ln 2 & \text{if } \alpha \leqslant 2, \\ \alpha H(1/\alpha) & \text{if } \alpha > 2, \\ \sim \ln\alpha & \text{for large } \alpha, \end{cases} \tag{38}$$

where $H(x)=-[x\ln x+(1-x)\ln(1-x)]$. This shows in particular that for $p \gg N$ the mutual information $I_P(\mathbf{S};\mathbf{B}|\mathbf{X})$ increases at most as $\ln p/N$ for p large. We will see in Secs. III D and IV that this behavior is indeed reached for supervised learning tasks. This should be contrasted with the behavior for smooth densities, in $\frac{1}{2}\ln p/N$.

D. Opper-Haussler bounds

In the case of class supervised learning with patterns correlated with the vector **B**, it is not clear at this point which asymptotic behavior for the mutual information between data and parameter should be expected. In the case of supervised learning, with a PDF for the patterns that does not depend on the parameter $p(\boldsymbol{\xi}|\mathbf{B})=p(\boldsymbol{\xi})$, very useful bounds on the mutual information $I(\mathbf{S},\mathbf{X};\mathbf{B})=I(\mathbf{S};\mathbf{B}|\mathbf{X})$ have been derived by Opper and Haussler [19,7].

From these bounds one obtains the asymptotic behavior for the mutual information. For the standard perceptron (that is, for supervised learning with the deterministic rule and patterns uncorrelated with the parameter), the main result is $I(\mathbf{S};\mathbf{B}|\mathbf{X})\sim N\ln p$ in the limit $p\to\infty$.

In the Appendix we apply the techniques of [7] to the case of supervised and unsupervised parameter estimation tasks with patterns correlated to the parameter. Quite interestingly, as we show in the Appendix, these tools introduced in [7] in order to extract the large p behavior of the mutual information, allow also to derive lower and upper bounds for both unsupervised and supervised learning in the regime of large N and large p for *any* given value of $\alpha=p/N$; that is, in the same regime as with the replica calculations. These bounds are shown in Fig. 2 for an unsupervised Gaussian and simple perceptron learning (models 1 and 6, respectively). The details are given in the Appendix, and we present here the main results concerning the limit of large data size.

One deduces from the bounds that in the large p limit, $I\sim(N/2)\ln p$ for smooth unsupervised learning, and $I\sim N\ln p$ for supervised learning. For N large, in the large α limit, one finds for smooth unsupervised learning,

$$\tfrac{1}{2}\ln\left(\alpha\frac{e}{4}\langle V'^2\rangle\right)\leqslant i(\mathbf{X};\mathbf{B})\leqslant\tfrac{1}{2}\ln(\alpha e\langle V'^2\rangle) \tag{39}$$

with $i(\mathbf{X};\mathbf{B})=\lim_{N\to\infty}I(\mathbf{X};\mathbf{B})/N$, in agreement with the exact behavior (33) derived in Sec. III, that is, $I\sim(N/2)\ln(p/N)\langle V'^2\rangle$. One can note the quality of the bounds in this case. For supervised learning, in the same limit

$$\ln\left(\alpha\frac{e}{\pi}e^{-V(0)}\right)\leqslant i(\mathbf{S},\mathbf{X};\mathbf{B})\leqslant\ln\alpha+O(\ln\ln\alpha) \tag{40}$$

with $i(\mathbf{S},\mathbf{X};\mathbf{B})=\lim_{N\to\infty}I(\mathbf{S},\mathbf{X};\mathbf{B})/N$. In the case of the standard perceptron, that is, for $V\equiv 0$, we have the better upper bound given by Eq. (38), which shows that there is no correction of order $\ln\ln\alpha$ to the leading behavior. We will see in the next section that the replica calculations, in agreement with the above inequalities, suggests that there is no such correction for nonzero potentials either.

IV. REPLICA CALCULATIONS

We now compare the previous results with those predicted by replica calculations.

A. Replica calculation of the mutual information

In the limit $N\to\infty$ with α finite, the calculation of the free energy $\langle\langle\ln Z(\mathbf{X})\rangle\rangle$ in Eq. (10) can be performed by standard replica technique. This calculation is the same as those related to Gibbs learning, done in [15–17] but the interpretation of the order parameters is different. Assuming replica symmetry, the result for the total mutual information (10) is as follows:

$$\lim_{N\to\infty}\frac{I(\mathbf{X};\mathbf{B})}{N}=i(\alpha,Q),$$

$$i(\alpha,Q)\equiv-\frac{1}{2}[Q+\ln(1-Q)]-\alpha\langle V(\lambda)\rangle - \alpha\int Dx\,A(x,Q)\ln A(x,Q) \tag{41}$$

with

$$A(x,Q)\equiv\int Dy\,\exp[-V(y\sqrt{1-Q}+x\sqrt{Q})], \tag{42}$$

Dx and Dy being the Gaussian measure. The order parameter $Q=Q(\alpha)$ is solution of the saddle point equation

$$\frac{\partial i}{\partial Q}=0, \tag{43}$$

which reads

$$\frac{Q}{1-Q}=2\alpha\int Dx\,\frac{\partial A(x,Q)}{\partial Q}\ln A(x,Q). \tag{44}$$

The order parameter Q is restricted to the [0,1] interval and can be interpreted as the typical overlap between two directions compatible with the data. The stability of the symmetry ansatz has already been studied for various specific choices of potentials V. The main result [17] is that the replica symmetric solution is stable if

$$\frac{dQ}{d\alpha}>0. \tag{45}$$

Within this hypothesis of replica symmetry, and for a general potential V, one can analyze from Eq. (41) the behavior of the mutual information $i(\alpha)=i(\alpha,Q(\alpha))$ as function of α. Different behaviors will occur depending on some properties of the potential. We will illustrate each case with a specific model in Sec. V.

A first remark concerns the concavity of $i(\alpha)$. One expects the mutual information to be a concave function of the data size p. This is indeed the case for the mutual information computed with the replica technique under the replica symmetry ansatz. Since Q satisfies Eq. (43), one has $di/d\alpha=\partial i/\partial\alpha$, so that from Eq. (41) one can write

$$-\frac{Q}{2}-\tfrac{1}{2}\ln(1-Q)=i(\alpha)-\alpha\frac{di(\alpha)}{d\alpha}. \tag{46}$$

As the lhs is always positive for Q in $[0,1]$, one has $[i(\alpha)]/\alpha \geqslant di/d\alpha$. Under the reasonable hypothesis that the mutual information is a nondecreasing function of α, it follows that $i(\alpha)$ is concave. From Eq. (46) on gets also that $dQ/d\alpha$ has the sign of $-d^2i/d\alpha^2$ (wherever i admits a second derivative), hence $dQ/d\alpha>0$: this is exactly the condition for the stability of the replica symmetric solution.

One can note also the interesting structure of the above equation (46). From the replica calculation one has that the lhs is the (logarithm of the) volume of the domain in parameter space in which two directions taken at random have a typical overlap equal to Q. If we define $j\equiv[di(\alpha)]/d\alpha$, the rhs of Eq. (46) is the Legendre transform $l(j)\equiv i(\alpha)-\alpha j$, which is a function of j alone, that is, of the marginal gain of information for an infinitesimal increase of α.

We consider now the behavior of the mutual information (41) in the small and large α regimes according to the replica calculation.

B. Unsupervised learning

We consider first the case of unsupervised learning. We derived the behavior for small α which is true for all potential. In the large α we consider smooth and discontinuous potentials showing different asymptotic behaviors. In the next section we will deduce the asymptotic behavior for supervised learning from the behaviors obtained for unsupervised learning.

1. Small α

For some potentials one finds Q strictly null from $\alpha=0$ up to a critical value $\alpha_c>0$. This is known as *retarded generalization* in the context of supervised learning [21], and *retarded classification* in the case of unsupervised learning [15]. Explicit calculation gives that such retarded classification occurs whenever $\langle\lambda\rangle=0$, a case illustrated by models 1 and 2 in the next section.

In such case, since from Eq. (42) $A(x,0)=1$ for any x, one gets from Eq. (41) that the mutual information is strictly linear in $[0,\alpha_c]$:

$$i(\alpha)=-\alpha\langle V(\lambda)\rangle. \tag{47}$$

This is a regime where there is no redundancy in the data: each datum conveys some information independent from the information conveyed by the other data. It corresponds, in the context of neural coding, to the regime where full redundancy reduction can be achieved [8,12].

In this regime one saturates the bound (25): one gains from the data the largest possible amount of information about the probability distribution of the patterns. However, $Q=0$ means that no estimation of the parameter **B** is possible for $\alpha<\alpha_c$. To understand better this seemingly paradoxical result, consider the simple case $N=3$ and a overlap distribution $P(\lambda)=\delta(\lambda)$. After receiving a first example $\boldsymbol{\xi}$, we know for sure that the vector **B** lies in the plane orthogonal to this pattern. We have thus gained a large amount of information about the localization of **B**. However, due to the symmetrical nature of the space left for **B**, one cannot give an estimation of this orientation and the direction of the next pattern is still unpredictable.

It is only at α_c when correlations between examples appear that one is able to make prediction on the next sample. Then Q becomes different from zero, and this may happen either continuously or with a jump to a finite value. According to Eq. (46) the linear regime is left smoothly in the continuous case, and with a discontinuity in the slope in the discontinuous case. In any case, the mutual information itself is continuous at the transition since the information is bounded by Eq. (25), and it cannot decrease (one cannot have less information with more examples). It follows also that the mean free energy must leave his zero level continuously.

For $\langle\lambda\rangle\neq0$, the bias in the distribution of λ allows to build a nontrivial estimate of **B** even with a very small number of examples. Then the mutual information cannot saturate the linear bound. Indeed, in the $\alpha\to0$ limit, one finds the following behavior for the mutual information when $\langle\lambda\rangle\neq0$:

$$i(\alpha)=-\alpha\langle V(\lambda)\rangle-\tfrac{1}{4}\alpha^2\langle\lambda\rangle^4+O(\alpha^3). \tag{48}$$

2. Large α limit

We consider now the $\alpha\to\infty$ limit. First, one can see the relationship between the asymptotic behaviors of $Q\to1$ and $i(\alpha)$ from Eq. (46). If for large α Eq. (43) for Q gives

$$1-Q=(\alpha C)^{-\nu} \tag{49}$$

for some exponent $\nu>0$ and constant C, then Eq. (46) gives

$$i(\alpha)\sim\frac{\nu}{2}\ln(\alpha C)+\frac{\nu-1}{2}. \tag{50}$$

In already studied models one finds $\nu=1$ for smooth PDF's, and $\nu=2$ for standard supervised learning tasks (see, e.g., [2] and the papers cited in Sec. IV A). This implies a behavior in $1/2\ln\alpha$ and $\ln\alpha$ for smooth and non smooth potentials, respectively. More precisely, the asymptotic behaviors are as follows.

For smooth potentials, a straightforward expansion of Eqs. (41) and (44) for $Q\to1$ leads to

$$i(\alpha)=\tfrac{1}{2}\ln[\alpha\langle V'^2(\lambda)\rangle]+O(\alpha^{-1}). \tag{51}$$

This is in agreement with Eq. (33) and the bounds (39). Two examples of smooth unsupervised learning are detailed in Sec. V, models 1 and 2.

We study now the interesting case of a discontinuity of the form

$$\begin{aligned} &P(\lambda)=0, \quad \lambda<\lambda_0,\\ &P(\lambda)\ \text{ smooth}, \quad \lambda>\lambda_0,\\ &\lim_{\lambda\to\lambda_0} P(\lambda)=\Delta P>0. \end{aligned} \tag{52}$$

By closer inspection of Eqs. (41) and (44), one finds that in the limit $Q\to1$, in the region that contributes the most in the

integrations on x and λ, one can replace $\exp -V(z)$ with $z \equiv y\sqrt{1-Q}+x\sqrt{Q}$ by $\Theta(z-\lambda_0)\exp -V(\max[\lambda_0,x])$. This yields the leading order in the asymptotic expansion:

$$i(\alpha)\sim \ln(\alpha \Delta P K) \tag{53}$$

with

$$K=\sqrt{e}\int Dx\, x \ln \int_{-x}^{\infty} Dy. \tag{54}$$

The numerical value of this constant is $K\sim 1.489$.

This behavior (53) is in good agreement with the lower bound in Eq. (40). For such discontinuous probability, the rate of information gain given by the patterns is twice the rate for smooth potentials. This rate is controlled by the value of the discontinuity ΔP. An example of unsupervised learning with discontinuous potential is detailed in Sec. V, model 3.

C. Supervised learning

We have shown in Sec. II C, Eqs. (13), (21), and (22), how the information quantities related to supervised and unsupervised tasks is related. Having computing in the preceding section the asymptotic behaviors in the unsupervised case, we then deduce the asymptotic behavior for the mutual informations related to supervised learning. We give below the main results for cluster and class learning. The notations are the same as in Sec. II C.

1. Cluster learning

Let $P(\lambda)$ be a smooth distribution. For small α, one gets

$$\begin{aligned} \lim_{N\to\infty}\frac{I_{AP}(\mathbf{A},\mathbf{X};\mathbf{B})}{N} &= -\alpha\langle V_P\rangle + O(\alpha^2), \\ \lim_{N\to\infty}\frac{I_{\Sigma P}(\mathbf{A};\mathbf{B}|\mathbf{X})}{N} &= -\alpha(\langle V_P\rangle - \langle V_{\Sigma P}\rangle) + O(\alpha^2) \end{aligned} \tag{55}$$

and for large α

$$\begin{aligned} \lim_{N\to\infty}\frac{I_{AP}(\mathbf{A},\mathbf{X};\mathbf{B})}{N} &= \tfrac{1}{2}\ln(\alpha\langle V_P'^2\rangle) + O(\alpha^{-1}), \\ \lim_{N\to\infty}\frac{I_{\Sigma P}(\mathbf{A};\mathbf{B}|\mathbf{X})}{N} &= \tfrac{1}{2}\ln\left(\frac{\langle V_P'^2\rangle}{\langle V_{\Sigma P}'^2\rangle}\right) + O(\alpha^{-1}). \end{aligned} \tag{56}$$

The amount of information given by the cluster label converges toward a constant. Then almost all the information comes from the patterns alone. This is illustrated in Sec. V, model 4. In the special case of two nonoverlapping cluster distributions, that is, $P(\lambda)$ and $P(-\lambda)$ are not different from zero together, we have $\langle V_{\Sigma P}'^2\rangle = \langle V_P'^2\rangle$ and the label information converges to zero. With a large number of patterns, the vector $\mathbf{B}$ becomes localized with high accuracy. Now, since the patterns with $A=-1$ and $A=+1$ are well separated in this model, the label of the patterns become predictable and give no additional information. This behavior is illustrated in Sec. V, model 5.

2. Class learning

In this section $P(\lambda)$ is a symmetrical smooth distribution. For small α, with $V_{\Theta P}$ defined in Eq. (20), one has

$$\begin{aligned} \lim_{N\to\infty}\frac{I_{PS}(\mathbf{S},\mathbf{X};\mathbf{B})}{N} &\sim -\alpha\langle V_{\Theta P}(\lambda)\rangle + O(\alpha^2), \\ \lim_{N\to\infty}\frac{I_P(\mathbf{S};\mathbf{B}|\mathbf{X})}{N} &\sim \alpha\ln 2 + O(\alpha^2). \end{aligned} \tag{57}$$

The $O(\alpha^2)$ is a negative contribution. This is in agreement with the bound (34), (38).

Consider first the particular case where the patterns have no statistical dependency in the vector $\mathbf{B}$, that is $V(\lambda)\equiv 0$. Then the pattern themselves carry no information about $\mathbf{B}$, and the task is the standard supervised learning task by a simple perceptron. One gets the asymptotic behavior for the mutual information $I_{\text{Perceptron}}(\mathbf{S},\mathbf{X};\mathbf{B}) = I_{\text{Perceptron}}(\mathbf{S};\mathbf{B}|\mathbf{X})$:

$$\lim_{N\to\infty} I_{\text{Perceptron}}(\mathbf{S};\mathbf{B}|\mathbf{X})/N \sim \ln\left(\alpha\sqrt{\frac{2}{\pi}}K\right) \tag{58}$$

where K is given by Eq. (54), in agreement with the computation of the free energy done in [22]. In this case the bound (34),(38) is asymptotically saturated. This is illustrated in Sec. V, model 6.

Consider now the case where the patterns are correlated with the direction $\mathbf{B}$, that is, $V(\lambda)\not\equiv 0$. The distribution (19) has the form (52) with $\lambda_0=0$ and $\Delta P = 2P(0)$. The asymptotic behaviors are given by

$$\begin{aligned} \lim_{N\to\infty}\frac{I_{PS}(\mathbf{S},\mathbf{X};\mathbf{B})}{N} &\sim \ln\left(\alpha\sqrt{\frac{2}{\pi}}K\, e^{-V(0)}\right), \\ \lim_{N\to\infty}\frac{I_P(\mathbf{S};\mathbf{B}|\mathbf{X})}{N} &\sim \tfrac{1}{2}\ln\left(\alpha\frac{2K^2e^{-2V(0)}}{\pi\langle V_P'^2\rangle}\right), \end{aligned} \tag{59}$$

where K is given by Eq. (54). The information rate given by the pair $(\mathbf{S},\mathbf{X})$ behaves as $\ln\alpha$, as for the simple perceptron, but here half of the information comes from the patterns alone and half from the class information. These results are illustrated in Sec. V, models 6 and 7.

V. SPECIFIC MODELS

We illustrate on specific models the different behaviors of the mutual information discussed in the preceding section. We compare the predictions of the replica calculations with the exact results from Sec. III. Some of the models presented here have been previously treated in the replica symmetry approach. For those models, the behavior of the order parameter can be found in the cited references.

A. Unsupervised learning

Model 1: smooth Gaussian learning. The simplest model is obtained for a Gaussian overlap distribution. The replica calculation of the free energy has been performed in [23]. We use the following parameterization:

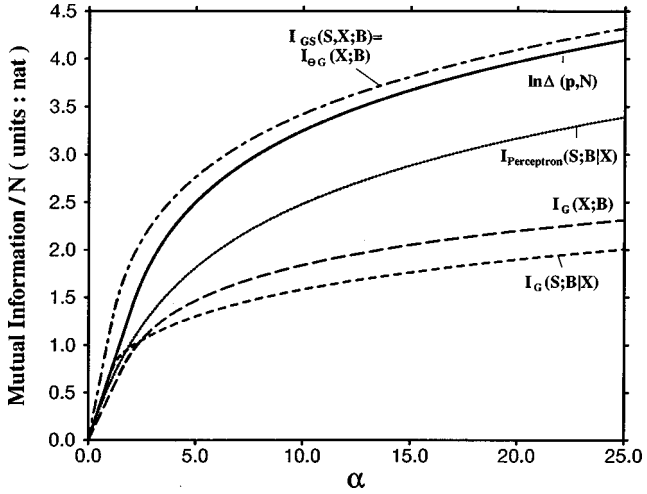

FIG. 1. The smooth unsupervised Gaussian learning $I_G(\mathbf{X};\mathbf{B})$ from model 1. For large α it behaves as $\sim\frac{1}{2}\ln\alpha$. The supervised class information $I_{GS}(\mathbf{S},\mathbf{X};\mathbf{B})$ and the additional class information $I_G(\mathbf{S};\mathbf{B}|\mathbf{X})$ from model 6 with $\sigma=1/\sqrt{6}$ and $\rho=0$. Their asymptotic behavior are respectively, $\sim\ln\alpha$ and $\sim\frac{1}{2}\ln\alpha$. Shown also is the class information $I_{\text{Perceptron}}(\mathbf{S};\mathbf{B}|\mathbf{X})$ for the simple perceptron from model 6 and the bound on the class label information $\ln\Delta(p,N)$ from Eq. (38). The simple perceptron asymptotically saturates the bound. Both of them have a $\sim\ln\alpha$ asymptotic behavior.

$$P(\lambda)=G(\lambda;\rho,\sigma)\equiv\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\lambda+\rho)^2}{2\sigma^2}\right). \quad (60)$$

The mutual information $I_G(\mathbf{X},\mathbf{B})$ is shown in Fig. 1 with parameters value $\rho=0$ and $\sigma=1/\sqrt{6}$.

As $\langle\lambda\rangle=0$, retarded classification occurs. For large α, the information behaves as $\sim\frac{1}{2}\ln(\alpha\langle V_G'^2\rangle)$ in agreement with Eq. (33). The behavior is similar to the one in model 2 below, for which we give a more detailed analysis. In Fig. 2(a) the information as computed with the replica technique is compared with the lower I_{lb} and upper I_{ub} bound from Sec. III D computed respectively by Eqs. (A14) and (A33). The bounds are in very good agreement with the replica calculation.

Model 2: smooth mixture distribution. The data are generated from a Gaussian mixture distribution with an overlap distribution given by

$$P(\lambda)=GG(\lambda;\rho,\sigma)\equiv\frac{1}{2}\sum_{A=\pm1}G(A\lambda;\rho,\sigma), \quad (61)$$

where G is the Gaussian distribution (60) introduced in model 1. We will see how this particular overlap distribution is also related to supervised cluster and supervised class learning (see models 4 and 7).

The behavior of the order parameter Q and the mean free energy are given in Fig. 3. The mutual information $I_{GG}(\mathbf{X};\mathbf{B})$ and the bound (25) associated with the distribution (61) are shown in Fig. 4. In both figures the parameters are $\rho=1.2$ and $\sigma=0.5$.

Since $\langle\lambda\rangle=0$, *retarded classification* occurs: up to a critical value α_c, the order parameter Q is null, the free energy is null and the mutual information saturates the linear bound, being given by Eq. (47). At α_c the mutual information leaves this linear regime. In the large α limit, the asymptotic behavior is $\sim\frac{1}{2}\ln\alpha$. This is the same behavior as in model 1.

In the replica symmetry ansatz, the true minimum of the free energy is given by $\langle\langle\ln Z\rangle\rangle=0$ until $\alpha=\alpha_1$, and then the solution $\alpha_1\rightarrow P_3$ shown on Fig. 3(b). The corresponding behavior of the order parameter is shown on Fig. 3(a): Q is null until α_1 and follows the lower branch until $P3$ where it jumps to the upper branch. In this scenario we thus have $\alpha_c=\alpha_1$.

However, it had been suggested in [17] that the order parameter Q can reach the upper branch well before $\alpha(P_3)$. As we have seen, the mean free energy cannot be positive and must be continuous (see Sec. IV B 1). It results that the only possibility of a jump to the upper branch before $\alpha(P_3)$ (that is, by following a metastable solution), would be that the free energy follows the path $0\rightarrow\alpha_2\rightarrow P_3$ (see Fig. 3). In such case the order parameter is null until α_2, where it jumps to the upper branch. This would give $\alpha_c=\alpha_2$.

Model 3: discontinuous Gaussian learning. This case has been treated in [16]. The data are generated from

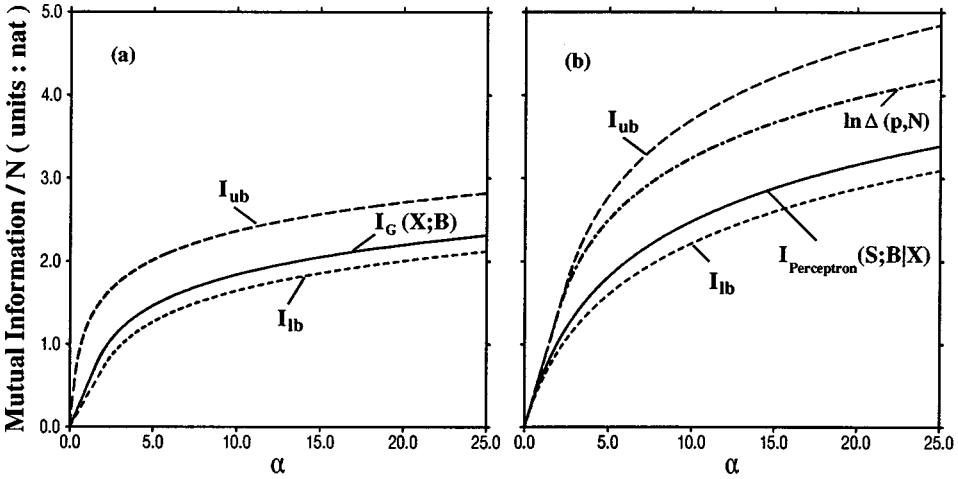

FIG. 2. The lower I_{lb} and upper I_{ub} bound on the mutual information from Sec. III D. These are computed with Eqs. (A14) and (A33), respectively, and compared with the mutual information computed with the replica technique (a) for the smooth unsupervised Gaussian learning, model 1 and (b) for the supervised learning of the simple perceptron, model 6.

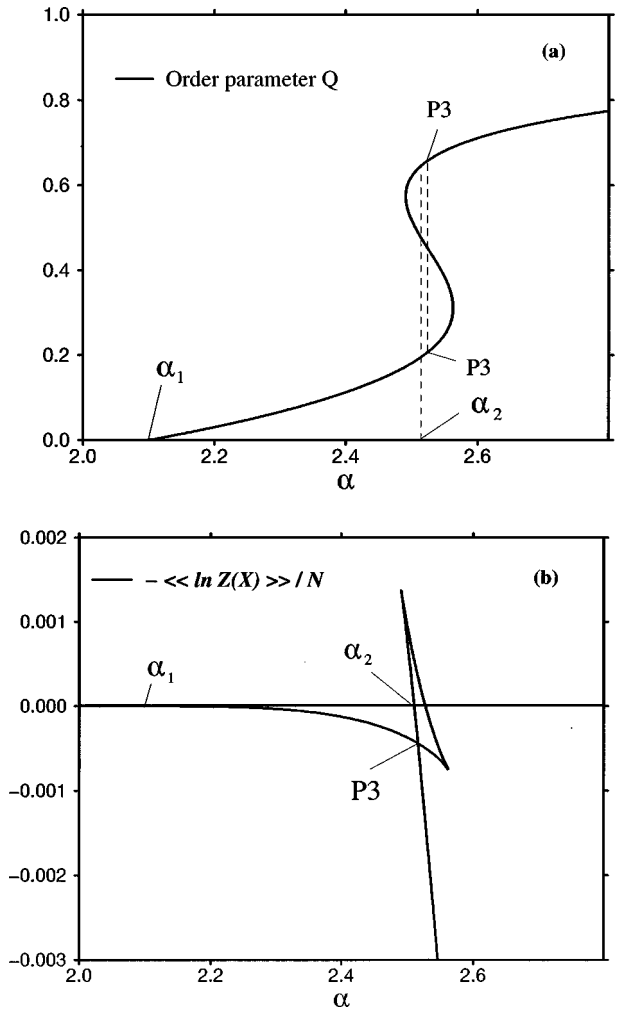

FIG. 3. (a) the order parameter Q and (b) the free energy as functions of α, for the smooth unsupervised mixture learning, model 2 with $\rho=1.2$ and $\sigma=0.5$, as computed in (Buhot and Gordon, 1998) under the replica symmetry ansatz. In the range of α values shown on these graphs the mean field equation for Q, Eq. (43), accepts several solutions (in particular $Q=0$ is always a solution). The stability analysis (not shown) and our results allow to eliminate some of them. In particular the values giving a positive free energy must be rejected. The solution corresponding to the absolute minimum of the free energy follows $0\rightarrow\alpha_1\rightarrow P3$, which gives $\alpha_c=\alpha_1$. Another metastable pathway is $0\rightarrow\alpha_2\rightarrow P3$ (see text). $\alpha_1=2.10$, $\alpha_2=2.515$, and $\alpha(P3)=2.527$.

the discontinuous overlap distribution obtained from the truncated Gaussian distribution, $\Theta G(\lambda;\sigma,\rho)=2\Theta(\lambda)G(\lambda;\sigma,\rho)$. The mutual information $I_{\Theta G}(\mathbf{X},\mathbf{B})$ is given in Fig. 1 with $\sigma=1/\sqrt{6}$ and $\rho=0$. The asymptotic behavior for large α is $\sim\ln\alpha$, see Eq. (53). For large data size, it is the patterns near the discontinuity which give the largest information about the localization of **B**.

B. Supervised learning

Model 4: Gaussian cluster learning. As a particular instance of cluster learning, Eq. (12), we consider the Gaussian mixture (61) introduced in model 2 in which the two clusters $A=\pm1$ have Gaussian distributions:

$$P_A(\lambda)=G(A\lambda;\rho,\sigma), \tag{62}$$

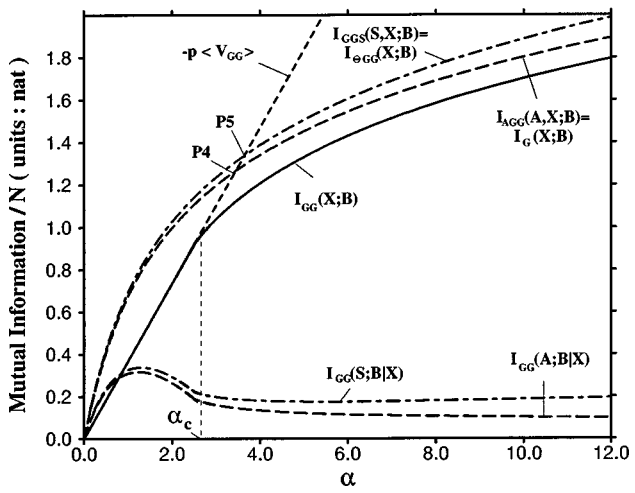

FIG. 4. As function of α, in the large N limit: (i) the mutual information for smooth unsupervised mixture learning $I_{GG}(\mathbf{X};\mathbf{B})$ for model 2, with $\sigma=0.5$ and $\rho=1.2$, together with the associated linear bound. This information is strictly linear up to α_c. The special structure near α_C visible on the order parameter and the free energy (figure 3) is not visible here due to the graph scale. (ii) The supervised Gaussian cluster information $I_{AGG}(\mathbf{A},\mathbf{X};\mathbf{B})$ and the cluster information $I_{GG}(\mathbf{A};\mathbf{B}|\mathbf{X})$ from model 4. (iii) The supervised discontinuous class information $I_{GGS}(\mathbf{S},\mathbf{X};\mathbf{B})$ and the class information $I_{GG}(\mathbf{S};\mathbf{B}|\mathbf{X})$ from model 7. All these models are linked together (see Figs. 5 and 9). $\alpha(P_4)=3.45$ and $\alpha(P_5)=3.65$ are upper bounds on α_c (see text, models 4 and 7).

where G is the Gaussian distribution (60). Each pattern is generated from one of the two clusters with equal probability and the cluster label $A=\pm1$ is given. Noting $I_{AGG}(\mathbf{A},\mathbf{X};\mathbf{B})$ the information the patterns and their labels give about **B** and $I_{GG}(\mathbf{A};\mathbf{B}|\mathbf{X})$ the cluster information, relations (13) and (15) relating supervised and unsupervised learning are illustrated in Fig. 5.

$I_G(\mathbf{X};\mathbf{B})$ and $I_{GG}(\mathbf{X};\mathbf{B})$ have been calculated, respectively, in models 1 and 2. All of this information is plotted in figure 4 for $\sigma=0.5$ and $\rho=1.2$. For small α, the cluster information grows. As the estimation of direction **B** becomes more and more accurate with the number of data, the cluster

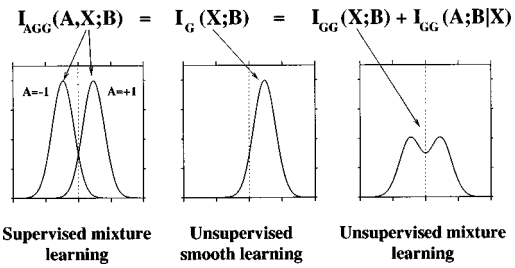

FIG. 5. Illustration of Eqs. (13) and (15) for the particular case of Gaussian cluster learning. The information $I_{AGG}(\mathbf{A},\mathbf{X};\mathbf{B})$ the patterns and their labels give about **B** is equal to the information $I_G(\mathbf{X};\mathbf{B})$ given in an unsupervised smooth learning with examples drawn from the overlap probability $G(\lambda;\sigma,\rho)$. This information is also equal to sum of the information $I_{GG}(\mathbf{X};\mathbf{B})$ the patterns without any cluster information give about **B** (the unsupervised mixture information associated to model 2), plus the cluster information $I_{GG}(\mathbf{A};\mathbf{B}|\mathbf{X})$ the labels convey about **B** when the patterns are known.

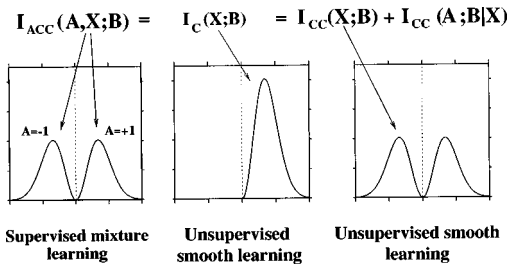

FIG. 6. Illustration of Eqs. (13) and (15) for the PDF $P(\lambda) = CC(\lambda;a)$ (see text, Sec. II C 2 and model 5). The two clusters are well separated and all the distributions are smooth.

to which the patterns belong becomes more predictable. This explains the decrease of the information that converges towards a constant (56). Due to the smooth nature of the PDF $G(\lambda;\sigma,\rho)$, the large α supervised information behavior is $I_{AGG}(\mathbf{A},\mathbf{X};\mathbf{B}) \sim \frac{1}{2}\ln\alpha$.

We show that the linear bound on the mutual information can be used to obtain bounds on the value α_c. Let $\alpha(P_4)$ be the intersection of the information $I_{AGG}(\mathbf{A},\mathbf{X};\mathbf{B})$ with $-\alpha\langle V_{GG}\rangle$, that is, the linear bound for the unsupervised information $I_{GG}(\mathbf{X};\mathbf{B})$. Since the supervised information is always bigger than the unsupervised one, see Eq. (16), one has

$$\alpha_c \leqslant \alpha(P_4). \tag{63}$$

This is illustrated in Fig. 4.

Model 5: nonoverlapping cluster learning. We consider the cluster distribution $P_A(\lambda) = CC(A\lambda;\alpha)$ with

$$CC(\lambda;a) \equiv 2\Theta(\lambda)\frac{(1+a)^{3/2}}{\sqrt{2\pi}}\lambda^2\exp\left(-\frac{(1+a)\lambda^2}{2}\right). \tag{64}$$

The model is similar to model 4 but now the two clusters do not overlap. Relations between supervised information $I_{ACC}(\mathbf{C},\mathbf{X};\mathbf{B})$, unsupervised smooth information $I_C(\mathbf{X};\mathbf{B})$, unsupervised smooth learning $I_{CC}(\mathbf{X};\mathbf{B})$, and the cluster information $I_{CC}(\mathbf{A};\mathbf{B}|\mathbf{X})$ are illustrated in Fig. 6.

The information behaviors and the linear bound associated with distribution (64) are shown in Fig. 7 for $a=0.9$. The unsupervised information $I_{CC}(\mathbf{X};\mathbf{B})$ shows a similar behavior as that encountered in model 2. $I_{CC}(\mathbf{X};\mathbf{B})$ and $I_{ACC}(\mathbf{C},\mathbf{X};\mathbf{B})$ converges to the same limit in $\sim\frac{1}{2}\ln\alpha$. The cluster information vanishes due to the fact that the clusters do not overlap (the cluster label becomes predictable with high accuracy).

Model 6: supervised perceptron. The data are generated by the overlap distribution $G(\lambda;\sigma,\rho)$ considered in model 1 and a teacher provide the class label $S=\mathrm{sgn}(\mathbf{B}\cdot\boldsymbol{\xi})$ for each pattern $\boldsymbol{\xi}$ of the data set. Relations between supervised information $I_{GS}(\mathbf{S},\mathbf{X};\mathbf{B})$, unsupervised discontinuous information $I_{\Theta G}(\mathbf{X};\mathbf{B})$, calculated in model 3, unsupervised smooth learning $I_G(\mathbf{X};\mathbf{B})$, calculated in model 1, and the class information $I_G(\mathbf{S};\mathbf{B}|\mathbf{X})$ are illustrated in Fig. 8.

These information quantities are shown in Fig. 1 for $\sigma = 1/\sqrt{6}$ and $\rho=0$. Also shown is the bound (34) on the class information in the large N limit. For small α, the class information $I_G(\mathbf{S};\mathbf{B}|\mathbf{X})$ almost saturate this bound in agreement

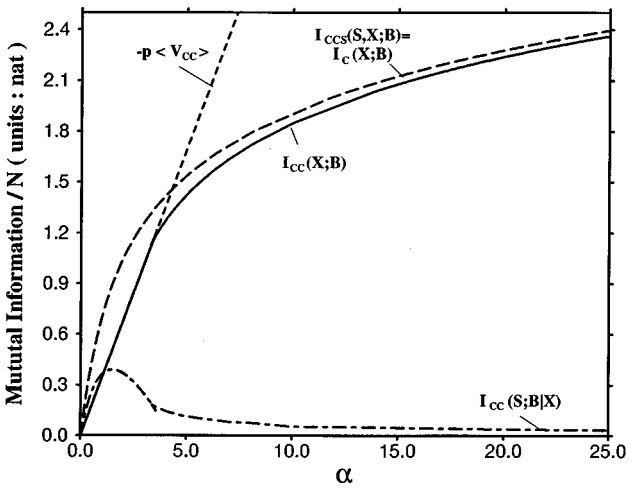

FIG. 7. As function of α the information quantities appearing in Fig. 6 for model 5 with $a=0.9$: $I_{ACC}(\mathbf{A},\mathbf{X};\mathbf{B})$ (supervised learning), $I_C(\mathbf{X};\mathbf{B})$ (smooth unsupervised), together with the associated linear bound $-p\langle V_{CC}\rangle$, and $I_{CC}(\mathbf{A};\mathbf{B}|\mathbf{X})$ (cluster information). The supervised and unsupervised informations have the same asymptotic behavior, $\sim\frac{1}{2}\ln\alpha$. The class information vanishes because the cluster label becomes easily predictable for large α.

with Eq. (57). For large α it behaves as $\sim\frac{1}{2}\ln\alpha$. In this limit the label information of examples near the boundary separating $S=-1$ and the $S=+1$ example give valuable information about $\mathbf{B}$. The smooth unsupervised part $I_G(\mathbf{X};\mathbf{B})$ behaves also as $\sim\frac{1}{2}\ln\alpha$. This implies that the total supervised information $I_{GS}(\mathbf{S},\mathbf{X};\mathbf{B})$ behaves as $\sim\ln\alpha$.

The special case $\sigma=1$, $\rho=0$ corresponds to $V(\lambda)\equiv 0$, that is to the standard supervised learning task with a teacher perceptron. The patterns are symmetrically distributed in all the directions and are not correlated with the symmetry-breaking orientation. Then $I_G(\mathbf{X};\mathbf{B})=0$ and $I_{GS}(\mathbf{S},\mathbf{X};\mathbf{B}) = I_{\mathrm{Perceptron}}(\mathbf{S};\mathbf{B}|\mathbf{X})$. This information is plotted in Fig. 1. According to Eq. (58) it asymptotically saturates the bound. In Fig. 2(b) the information as computed with the replica technique is compared with the lower I_{lb} and upper I_{ub} bound from Sec. III D computed, respectively, by Eqs. (A14) and (A33). One sees that the replica calculation is in good agreement with the bounds. If one believes that it gives indeed the exact result, then one can see the very good quality of the

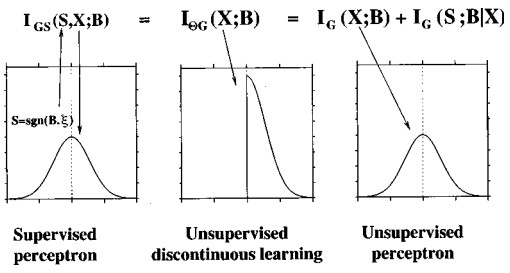

FIG. 8. Illustration of Eqs. (21) and (22) relating supervised and unsupervised learning with a Gaussian pdf to unsupervised learning with a discontinuous distribution. For each mutual information the subscript refers to the distribution from which the examples are drawn (see text, Sec. II C 2 and model 6). The particular case $\sigma=1$, for which $I_G(\mathbf{X};\mathbf{B})=0$, corresponds to the standard supervised learning task by a perceptron.

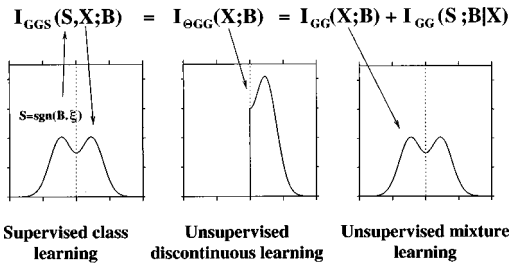

FIG. 9. Illustration of Eqs. (21) and (22) relating supervised class learning to unsupervised learning with discontinuous distribution. For each mutual information the subscript refers to the distribution from which the examples are drawn (see text, Sec. II C 2 and model 7).

lower bound at any value of α, whereas the upper bound is less precise due to the presence of the term of order $\ln\ln\alpha$ (see Sec. III D).

Model 7: class learning. The patterns are generated with the mixture distribution $GG(\lambda;\sigma,\rho)$ considered in model 2, but now a teacher provides the class label for each pattern, that is, $S=\mathrm{sgn}(\mathbf{B}\cdot\boldsymbol{\xi})$. We note $I_{GGS}(\mathbf{S},\mathbf{X};\mathbf{B})$ the information the patterns and their labels give about $\mathbf{B}$, and $I_{GG}(\mathbf{S},\mathbf{B}|\mathbf{X})$ the label information. The relations (21) and (22) are illustrated in Fig. 9 where $I_{\Theta GG}(\mathbf{X};\mathbf{B})$ in an unsupervised discontinuous learning from examples drawn from the discontinuous overlap probability $\Theta GG(\lambda;\sigma,\rho)=2\Theta(\lambda)GG(\lambda;\sigma,\rho)$ and $I_{GG}(\mathbf{X};\mathbf{B})$ has been calculated in model 2.

These informations are drawn in Fig. 4 for $\sigma=0.5$ and $\rho=1.2$. For not too large α, the behavior of the class information $I_{GGS}(\mathbf{S},\mathbf{X};\mathbf{B})$ and $I_{GG}(\mathbf{S};\mathbf{B}|\mathbf{X})$ is similar to the behavior of their corresponding cluster information $I_{AGG}(\mathbf{A},\mathbf{X};\mathbf{B})$ and $I_{GG}(\mathbf{A};\mathbf{B}|\mathbf{X})$. For large α this is no more true due to the discontinuous nature of the class learning. The large α behavior is similar to the one encountered in model 6.

As the supervised information is always bigger than the unsupervised one, similarly to (63) one gets that $\alpha(P_5)$ is an upper bound on α_c (see Fig. 4). It has to be noted for supervised learning that in some region, especially for small α, one can gain more than one bit of information per example: one bit from the binary classification plus the information conveyed by the patterns themselves.

VI. BOUNDS FOR SPECIFIC ESTIMATORS

Given the data $\mathbf{X}$, one wants to find an estimate $\mathbf{J}$ of the parameter $\mathbf{B}$ (see Fig. 10). Although this paper is not primarily concerned with the question of estimating the performance of estimators, we show in this section that making use of the mutual information one can derive simple bounds on the performance of some specific estimators.

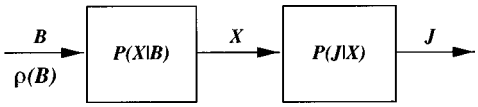

FIG. 10. The flow of information. First an orientation $\mathbf{B}$ is drawn from a *prior* distribution $\rho(\mathbf{B})$. Then, patterns are generated according to $P(\mathbf{X}|\mathbf{B})$. In the last process, an estimation $\mathbf{J}$ of the original orientation is extracted from the examples. The information decreases at each step.

The amount of information $I(\mathbf{X};\mathbf{B})$ limits the performance of any estimator. Indeed, since processing cannot increase information [6], one has

$$I(\mathbf{J};\mathbf{B})\leqslant I(\mathbf{X};\mathbf{B}). \tag{65}$$

This basic relationship allows to derive interesting bounds based on the choice of particular estimators. We consider first *Gibbs learning*, which consists in sampling a direction $\mathbf{J}$ from the *a posteriori* probability $P(\mathbf{J}|\mathbf{X})=\mathcal{P}(\mathbf{X}|\mathbf{J})\rho(\mathbf{J})/\mathcal{P}(\mathbf{X})$. In this particular case, the differential entropy of the estimator $\mathbf{J}$ and of the parameter $\mathbf{B}$ are equal, $H(\mathbf{J})=H(\mathbf{B})$. If $1-Q_g^2$ is the variance of the *Gibbs* estimator, from Eq. (6) and using again the fact that the entropy of a Gaussian distribution is greater than the entropy of any distribution with the same variance, one gets the relations for a Gaussian prior on $\mathbf{B}$

$$-\frac{N}{2}\ln(1-Q_g^2)\leqslant I_{\mathrm{Gibbs}}(J;B)\leqslant I(\mathbf{X};B). \tag{66}$$

These relations together with the linear bound (25) allows to bound the order parameter Q_g for small α, where this bound is of interest.

The *Bayes estimator* consists in taking for $\mathbf{J}$ the center of mass of the *a posteriori* probability. In the limit $\alpha\to\infty$, this distribution becomes Gaussian centered at its most probable value.

We can thus assume $P_{\mathrm{Bayes}}(\mathbf{J}|\mathbf{B})$ to be Gaussian with mean $Q_b\mathbf{B}$ and variance $1-Q_b^2$. Then the first inequality in Eq. (66) (with Q_g replaced by Q_b and *Gibbs* by *Bayes*) becomes an equality. Using the Cramer-Rao bound on the variance of the estimator one can then bound the mutual information for the Bayes estimator,

$$I_{\mathrm{Bayes}}(\mathbf{J};\mathbf{B})\leqslant\frac{N}{2}\ln[1+\alpha\langle V'^2(\lambda)\rangle]. \tag{67}$$

The rhs is the Fisher information (33). For $\alpha\to\infty$ all these quantities have the same asymptotic behavior. They are shown in Fig. 11 from replica calculation, when the data are generated with the Gaussian overlap distribution $G(\lambda;\rho,\sigma)$ from model 1.

The fact that Q_b, as computed with the replica technique, asymptotically saturates the Cramer-Rao bound was first noted in [24]. We have shown here that this manifests itself in the behavior of the mutual information and in the related quantity I_{Bayes} defined above.

VII. CONCLUSION

We have studied the mutual information between data and parameter in a family of unsupervised and supervised clustering tasks. We derived exact bounds, exact asymptotic behavior, and have compared these results with replica calculations.

We have restricted the analysis to continuous parameters. The case of discrete parameters is discussed in [3]. In such cases the mutual information is upper bounded by the entropy of the prior distribution on the parameter space, and

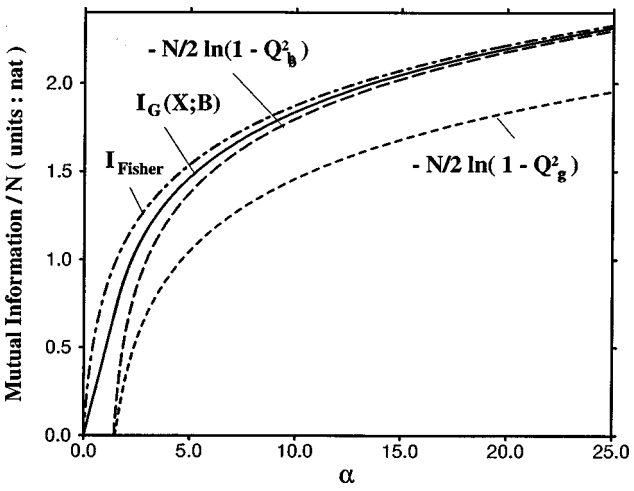

FIG. 11. The mutual information $I_G(\mathbf{X},\mathbf{B})$ for Gaussian unsupervised learning (model 1 with $\sigma=1/\sqrt{6}$, $\rho=0$). It limits the performance of any estimator $\mathbf{J}$, since $I(\mathbf{J};\mathbf{B})\leqslant I(\mathbf{X};\mathbf{B})$. The curve $-\frac{1}{2}\ln(1-Q_g^2)$ is a lower bound on the mutual information between the *Gibbs* estimator and $\mathbf{B}$ (which would be equal to this bound if the conditional probability distribution of the estimator were Gaussian with mean $Q_g\theta$ and variance $1-Q_g^2$. Shown also is the analogous curve $-\frac{1}{2}\ln(1-Q_b^2)$ for the *Bayes* estimator with $Q_b=\sqrt{Q_g}$. Q_g is computed from Eq. (44). In the limit $\alpha\to\infty$ these two latter curves and the replica information $I_G(\mathbf{X};\mathbf{B})$, all converge toward the exact asymptotic behavior, which can be expressed as $I_{\rm Fisher}=\frac{1}{2}\ln[1+\alpha\langle V'^2(\lambda)\rangle]$. This latter expression is, for any p, an upper bound for the two Gaussian curves.

converges exponentially to this value.

Most of the results concerning the behavior of the mutual information, observed for this particular family, are ''universal,'' in that they will be qualitatively the same for any problem that can be formulated as either a parameter estimation task or a neural coding task. This is in particular the case for the linear bound—that is, the information cannot grow faster than linearly in the data size—and the asymptotic behaviors. For smooth PDF, the large data size ($p\gg N$) behavior for I, given by the *Fisher information*, is $I\sim(N/2)\ln(p/N)$. In the case of potentials with a discontinuity, or equivalently in the case of supervised learning of a binary classification, the asymptotic behavior is $I\sim N\ln(p/N)$. These behaviors can be seen to be valid for any learning machine (see [1,4,5] for the smooth case, [7] for supervised learning), N being understood as the number of independent parameters. In particular, this results in optimal performances, which depends on p/N and not on $p/d_{\rm VC}$, where $d_{\rm VC}$ is the Vapnik-Chervonenkis dimension [26] (see, in particular, [13] for the smooth case).

We have obtained bounds and exact asymptotic behaviors by extending the results in [7] to the case of unsupervised and supervised learning with patterns correlated to the parameter. An interesting feature is that, for the supervised learning tasks where the patterns are correlated with the symmetry-breaking direction, half of the mutual information comes from the patterns alone, and half from the class information (given the patterns).

Besides the asymptotic regime p large, N arbitrary, we have also considered the case of large N at any given value of $\alpha=p/N$. In this regime we have both replica calculations and exact bounds, in particular, an upper bound for the class information and explicit upper and lower bounds for the mutual information obtained with the techniques of [7]. The results suggest that the replica symmetry ansatz give the correct solution. The lower bound is then quite good whereas the upper bound overestimate the mutual information by a factor that keeps increasing with the data size. We have also seen that our linear bound is particularly interesting in the case of retarded classification for $p/N\sim\alpha_c$. This critical value α_c gives the value of α at which the mutual information ceases to increase linearly with the amount of data, and where generalization begins. Contrary to the asymptotic regime, it can be seen to be related to the Vapnik-Chervonenkis dimension. This fact is confirmed by the analysis in [25] of the respective roles of N and $d_{\rm VC}$ in a supervised learning task for a model with $N\neq d_{\rm VC}$.

The analysis of the mutual information between data and parameter we have presented, suggests that it should be interesting to study other models with the same set of techniques, e.g., non smooth potentials with a singularity which is not a simple discontinuity, or models with a more complicated structure such as multilayer networks, or support-vector machines [26], which have been recently studied with statistical mechanics techniques [27].

ACKNOWLEDGMENTS

We gratefully acknowledge Nicolas Brunel, Arnaud Buhot, and Mirta Gordon for useful discussions. This work has been partly supported by the French Ministere de la Defense under Contract No. DGA96 2557A/DSP.

APPENDIX: OPPER-HAUSSLER BOUNDS

In this section we derive lower and upper bounds for the mutual information following Opper and Haussler [7]. We will write $\mathcal{P}(\mathbf{D}|\mathbf{B})$ for the data distribution given the parameter, where $\mathbf{D}$ is $\mathbf{X}$ or $(\mathbf{S},\mathbf{X})$ depending on the particular model considered.

1. Lower bound

Following [7], we make use of

$$I(\mathbf{D};\mathbf{B})=J_1(\mathbf{D};\mathbf{B})\geqslant J_t(\mathbf{D};\mathbf{B}) \tag{A1}$$

for any $0\leqslant t\leqslant 1$, with

$$J_t(\mathbf{D};\mathbf{B})\equiv-\int d\mathbf{B}\,\rho(\mathbf{B})\int d\mathbf{D}\,\mathcal{P}(\mathbf{D}|\mathbf{B}) \times\ln\int d\mathbf{W}\,\rho(\mathbf{W})\left(\frac{\mathcal{P}(\mathbf{D}|\mathbf{W})}{\mathcal{P}(\mathbf{D}|\mathbf{B})}\right)^t \tag{A2}$$

(where $\int d\mathbf{D}$ means the integration over the continuous data and the summation over the discrete data, if any). This holds because

$$J_t-I=\int d\mathbf{D}\,\mathcal{P}(\mathbf{D})\int d\mathbf{B}\,\mathcal{P}(\mathbf{B}|\mathbf{D})\ln\frac{\mathcal{P}(\mathbf{B}|\mathbf{D})}{Q_t(\mathbf{B}|\mathbf{D})}\geqslant 0,$$

is the average over the data of the Kullback divergence of $\mathcal{P}(\mathbf{B}|\mathbf{D})$ relative to $Q_t(\mathbf{B}|\mathbf{D})$ defined by

$$Q_t(\mathbf{B}|\mathbf{D})\equiv\rho(\mathbf{B})[\mathcal{P}(\mathbf{D}|\mathbf{B})]^t\Big/\int d\mathbf{W}\,\rho(\mathbf{W})[\mathcal{P}(\mathbf{D}|\mathbf{W})]^t.$$

Using the convexity of the logarithm, one lower bounds J_t by putting in Eq. (A2) the average over the data inside the logarithm. One then makes use of the independency of the examples given the parameter, leading to, for any t different from 0 and 1,

$$I \geqslant I_{\rm lb} \equiv -\int d\mathbf{B}\,\rho(\mathbf{B}) \ln \int d\mathbf{W}\,\rho(\mathbf{W})[\,\omega_t(\mathbf{B},\mathbf{W})]^p \tag{A3}$$

with, in the case of smooth unsupervised learning, that is $\mathcal{P}(\mathbf{D}|\mathbf{B}) = \mathcal{P}(\mathbf{X}|\mathbf{B})$ defined by Eq. (1),

$$\omega_t(\mathbf{B},\mathbf{W}) = \gamma_t(\mathbf{B},\mathbf{W}) \tag{A4}$$

and, in the case of class learning, that is for $\mathcal{P}(\mathbf{D}|\mathbf{B}) = \mathcal{P}(\mathbf{S},\mathbf{X}|\mathbf{B})$ defined by Eq. (18):

$$\omega_t(\mathbf{B},\mathbf{W}) = \gamma_t(\mathbf{B},\mathbf{W}) - \epsilon_t(\mathbf{B},\mathbf{W}), \tag{A5}$$

where

$$\gamma_t(\mathbf{B},\mathbf{W}) \equiv \int d\boldsymbol{\xi}\, p(\boldsymbol{\xi}|\mathbf{B})^{(1-t)} p(\boldsymbol{\xi}|\mathbf{W})^t \tag{A6}$$

and

$$\epsilon_t(\mathbf{B},\mathbf{W}) \equiv \int d\boldsymbol{\xi}\, p(\boldsymbol{\xi}|\mathbf{B})^{(1-t)} p(\boldsymbol{\xi}|\mathbf{W})^t \times \sum_{S=\pm 1} \Theta(S\boldsymbol{\xi}\cdot\mathbf{B})\Theta(-S\boldsymbol{\xi}\cdot\mathbf{W}). \tag{A7}$$

In the particular case where the patterns ξ are independent of the parameter, $p(\boldsymbol{\xi}|\mathbf{W}) = p(\boldsymbol{\xi})$, one recovers the supervised learning case considered in [7]: $\gamma_t(\mathbf{B},\mathbf{W}) = 1$ and $\epsilon_t(\mathbf{B},\mathbf{W})$ is the probability that $\mathbf{B}$ and $\mathbf{W}$ disagree on the classification of an example. Note that γ_t is the normalization factor that makes the mixture $p(\boldsymbol{\xi}|\mathbf{B})^{(1-t)} p(\boldsymbol{\xi}|\mathbf{W})^t / \gamma_t$ a well defined PDF for $\boldsymbol{\xi}$.

We perform the rest of the analysis working with Eqs. (A3) and (A5) for both supervised and smooth unsupervised learning, keeping in mind that for the latter case the term ϵ_t must be dropped. It will appear that, precisely, the asymptotic behavior will be governed by properties of γ_t in the case of smooth learning, and of ϵ_t in the case of discontinuous learning, leading respectively to the $(N/2)\ln p$ and $N \ln p$ behaviors.

Since the quantity $\gamma_t(\mathbf{B},\mathbf{W}) - \epsilon_t(\mathbf{B},\mathbf{W})$ lies in [0,1], for p large the integral in Eq. (A5) is dominated by the $\mathbf{W}$ such that $\gamma_t - \epsilon_t$ is close to 1. Similarly to [7], one will get that, *if* the volume $V_\delta(\mathbf{B})$ of $\mathbf{W}$ such that $\gamma_t(\mathbf{B},\mathbf{W}) - \epsilon_t(\mathbf{B},\mathbf{W}) \geqslant 1 - \delta$ behaves as $\delta^{d(\mathbf{B})}$ as $\delta \to 0$, then for large p the lower bound behaves as $d \ln p$, with $d = \int d\mathbf{B}\,\rho(\mathbf{B}) d(\mathbf{B})$. For the particular model family we are considering, this coefficient and in fact the detailed behavior of the bound for both the limit p large and the limit N large with p/N fixed, are easily derived, as we show now.

We thus take into account the special structure (2) for the PDF $p(\boldsymbol{\xi}|\mathbf{B})$. In this case the quantities γ_t and ϵ_t depend only on the scalar product of the two parameters, $q \equiv \mathbf{B}\cdot\mathbf{W}/\|\mathbf{B}\|\|\mathbf{W}\|$. We have

$$I_{\rm lb} = -\int d\mathbf{B}\,\rho(\mathbf{B}) \ln \int_{-1}^{1} dq\,\Omega(q,\mathbf{B})[\,\gamma_t(q) - \epsilon_t(q)]^p, \tag{A8}$$

where

$$\gamma_t(q) = \int Dx \int Dy\, \exp -(1-t)V(x) - tV(xq + y\sqrt{1-q^2}), \tag{A9}$$

$$\epsilon_t(q) = 2\int Dx\,\Theta(x) \int Dy\,\Theta(-xq - y\sqrt{1-q^2}) \times \exp -(1-t)V(x) - tV(xq + y\sqrt{1-q^2}). \tag{A10}$$

Dx and Dy being the Gaussian measures, and with Ω the volumic fraction of parameters having an overlap q with $\mathbf{B}$:

$$\Omega(q,\mathbf{B}) = \int d\mathbf{W}\rho(\mathbf{W})\,\delta(q - \mathbf{B}\cdot\mathbf{W}/\|\mathbf{B}\|\|\mathbf{W}\|). \tag{A11}$$

In the above expressions we have assumed for simplicity the potential to be symmetric, although there is no difficulty in considering nonsymmetric potentials in the case of unsupervised learning. Also, for simplicity let us restrict ourselves to the case of the uniform prior on the unit sphere, in which case the above volume does not depend on $\mathbf{B}$. Denoting by S_N the surface of the unit sphere in N dimensions, we have

$$\Omega(q) = \frac{S_{N-1}}{S_N}(1-q^2)^{(N-2)/2} \tag{A12}$$

and

$$I_{\rm lb} = -\ln \int dq\,\Omega(q)[\,\gamma_t(q) - \epsilon_t(q)]^p. \tag{A13}$$

Consider first the large N limit with $p = \alpha N$. The lower bound $I_{\rm lb}$ can then be computed by the saddle point method. One has

$$i_{\rm lb} \equiv \lim_{N\to\infty} \frac{I_{\rm lb}}{N} = -\tfrac{1}{2}\ln(1-q^2) - \alpha \ln[\,\gamma_t(q) - \epsilon_t(q)], \tag{A14}$$

where q satisfies the saddle point equation

$$\frac{\partial}{\partial q} i_{\rm lb}(q) = 0. \tag{A15}$$

One can see that $\gamma_t = \gamma_{1-t}$, so that the best lower bound is obtained for $t = \frac{1}{2}$. For a given model, the saddle point equation can be solved numerically [setting $\epsilon_t = 0$ in Eq. (A14) if one consider a smooth unsupervised learning model]. More explicit calculations can be performed for the simplest cases. For the Gaussian unsupervised learning defined in Eq. (60), with $\rho = 0$, we have

$$\gamma_t(q) = \left[1 + t(1-t)\frac{(1-\sigma^2)^2}{\sigma^2}(1-q^2)\right]^{-1/2} \tag{A16}$$

and for the standard perceptron supervised learning, we have $\gamma_t = 1$ and

$$\epsilon_t(q) = \epsilon_0(q) = \frac{1}{\pi}\arccos(q). \tag{A17}$$

The resulting bounds are shown on Figs. 2(a) and 2(b).

Expression (A14) is very reminiscent of the expression of the mutual information obtained with the replica techniques, see Eq. (41). If we identify q with $\sqrt{Q}$, the ''order parameter'' q that appears here must be identified as the Bayes parameter, whereas Q, in the replica approach, is the Gibbs parameter (see Sec. VI).

The large α limit is obtained by taking the leading behavior for $q \to 1$. One gets

$$\gamma_t(q) \sim 1 - t(1-t)(1-q)\langle V'^2 \rangle, \tag{A18}$$

$$\epsilon_t(q) \sim \sqrt{1-q}\,\frac{\sqrt{2}}{\pi} e^{-V(0)}. \tag{A19}$$

For large α one then gets that the lower bound on the mutual information behaves for smooth learning as

$$i_{\rm lb} \sim \tfrac{1}{2} \ln\left(\alpha \frac{e}{4} \langle V'^2 \rangle \right), \tag{A20}$$

where we have taken $t = 1/2$ which gives the largest lower bound, and for supervised learning as

$$i_{\rm lb} \sim \ln\left(\alpha \frac{e}{\pi} e^{-V(0)} \right), \tag{A21}$$

whatever t is.

As can be seen starting from Eq. (A13), the leading term in the large p limit for N finite, is the one given by expressions (A20) and (A21), that is $I \sim N/2 \ln p$ and $I \sim N \ln p$, respectively.

2. Upper bound

Again we follow closely [7]. The first step is to consider the inequality

$$I(\mathbf{D};\mathbf{B}) \leqslant \int d\mathbf{B}\, \rho(\mathbf{B}) d\mathbf{D}\, \mathcal{P}(\mathbf{D}|\mathbf{B}) \ln \frac{\mathcal{P}(\mathbf{D}|\mathbf{B})}{\mathcal{Q}(\mathbf{D})}, \tag{A22}$$

which is true for an arbitrary PDF $\mathcal{Q}(\mathbf{D})$ (this follows from the fact that the difference between the rhs and the lhs can be written as a Kullback divergence, hence an always nonnegative quantity). Taking $\mathcal{Q}$ as $\mathcal{Q}(\mathbf{D}) = \int d\mathbf{W} \rho(\mathbf{W}) \mathcal{Q}(\mathbf{D}|\mathbf{W})$, and rewriting $\ln[\mathcal{P}(\mathbf{D}|\mathbf{B})]/[\mathcal{Q}(\mathbf{D})]$ as $-\ln \int d\mathbf{W} \rho(\mathbf{W}) \exp -\ln[\mathcal{P}(\mathbf{D}|\mathbf{B})]/[\mathcal{Q}(\mathbf{D}|\mathbf{W})]$, the second step consists in upper bounding the rhs of the above inequality by passing the mean over the data inside the exponential. One gets

$$I(\mathbf{D};\mathbf{B}) \leqslant -\int d\mathbf{B}\, \rho(\mathbf{B}) \ln \int d\mathbf{W} \rho(\mathbf{W}) \times \exp - \int d\mathbf{D}\, \mathcal{P}(\mathbf{D}|\mathbf{B}) \ln \frac{\mathcal{P}(\mathbf{D}|\mathbf{B})}{\mathcal{Q}(\mathbf{D}|\mathbf{W})}. \tag{A23}$$

Optimizing with respect to $\mathcal{Q}$, this inequality is a variational method for bounding the mutual information. For a smooth distribution, hence for the case of smooth unsupervised learning, $\mathcal{P}(\mathbf{D}|\mathbf{B}) = \mathcal{P}(\mathbf{X}|\mathbf{B})$ being defined by Eq. (1), the optimal choice is simply

$$\mathcal{Q}(\mathbf{D}|\mathbf{W}) = \mathcal{P}(\mathbf{D}|\mathbf{W}). \tag{A24}$$

In the case of a discontinuity or of supervised learning [that is, $\mathcal{P}(\mathbf{D}|\mathbf{B}) = \mathcal{P}(\mathbf{S},\mathbf{X}|\mathbf{B})$ defined by Eq. (18)], the ratio $\mathcal{P}(\mathbf{S},\mathbf{X}|\mathbf{B})/\mathcal{Q}(\mathbf{S},\mathbf{X}|\mathbf{W})$ is not bounded. Following [7] we thus take

$$\mathcal{Q}(\mathbf{D}|\mathbf{W}) = \prod_{\mu=1}^{p} q_\delta(S^\mu|\boldsymbol{\xi}^\mu,\mathbf{B}) p(\boldsymbol{\xi}^\mu|\mathbf{B}) \tag{A25}$$

with q_δ a noisy version of the deterministic rule:

$$q_\delta(S|\boldsymbol{\xi},\mathbf{W}) = (1-\delta)\Theta(S\boldsymbol{\xi}\cdot\mathbf{W}) + \frac{\delta}{2}, \tag{A26}$$

where δ is a parameter smaller than 1 over which optimization will be done to get the best possible upper bound.

The upper bound then becomes, for supervised learning,

$$I \leqslant I_{\rm ub} \equiv -\int d\mathbf{B}\, \rho(\mathbf{B}) \ln \int d\mathbf{W} \rho(\mathbf{W}) \times \exp[-p\mathcal{D}_s(\mathbf{B},\mathbf{W}) - p\mathcal{D}_d(\mathbf{B},\mathbf{W},\delta)], \tag{A27}$$

where $\mathcal{D}_s$ and $\mathcal{D}_d$ are Kullback divergences related to the smooth and discontinuous parts of the PDF, respectively:

$$\mathcal{D}_s(\mathbf{B},\mathbf{W}) = \int d\boldsymbol{\xi}\, p(\boldsymbol{\xi}|\mathbf{B}) \ln \frac{p(\boldsymbol{\xi}|\mathbf{B})}{p(\boldsymbol{\xi}|\mathbf{W})} \tag{A28}$$

and

$$\mathcal{D}_d(\mathbf{B},\mathbf{W},\delta) = \int d\boldsymbol{\xi}\, p(\boldsymbol{\xi}|\mathbf{B}) \sum_{S=\pm 1} \Theta(S\boldsymbol{\xi}\cdot\mathbf{B}) \ln \frac{\Theta(S\boldsymbol{\xi}\cdot\mathbf{B})}{q_\delta(S|\boldsymbol{\xi},\mathbf{W})}. \tag{A29}$$

As it is clear from the above equations, the case of smooth unsupervised learning is obtained by simply dropping the term $\mathcal{D}_d$. Conversely, in the case considered in [7] where there is no correlation between the patterns and the parameter, the quantity $\mathcal{D}_s$ is not present (it is zero). We perform the rest of the analysis for both discontinuous (supervised) and smooth (unsupervised) learning, keeping in mind that for the latter case the term $\mathcal{D}_d$ must be dropped.

We note that $\mathcal{D}_s$ and $\mathcal{D}_d$ are simply related to the quantities that appear in the lower bounds, γ_t and ϵ_t defined in Eqs. (A6) and (A7), as follows:

$$\mathcal{D}_s = -\left[\frac{\partial}{\partial t} \gamma_t \right]_{t=0} \tag{A30}$$

and

$$\mathcal{D}_d = -\ln\left(1 - \frac{\delta}{2}\right) - \epsilon_0 \ln \frac{\delta}{2-\delta}. \tag{A31}$$

The next step in [7] is to upper bound again $I_{\rm ub}$ in such a way that both the lower and the upper bounds depend in a simple way on the same quantity (namely, $\epsilon_{1/2}$). Here we will instead keep the (slightly) better bound $I_{\rm ub}$, since it can be easily computed for the particular model family we are considering.

We thus specify now the analysis to the case where the PDF $p(\boldsymbol{\xi}|\mathbf{B})$ has the special structure (2). Following the same procedure as for the lower bound in the preceding section, we get

$$I_{\rm ub}=-\ln\int dq\,\Omega(q)\exp[-p\mathcal{D}_s(q)-p\mathcal{D}_d(q,\delta)] \tag{A32}$$

with Ω given by Eq. (A12), and $\mathcal{D}_s(q)$ and $\mathcal{D}_d(q,\delta)$ related to $\gamma_t(q)$ and $\epsilon_t(q)$ at $t=0$ according to Eqs. (A30) and (A31).

Consider first the limit $N\to\infty$ with $\alpha=p/N$ fixed. Using the saddle point method one has

$$i_{\rm ub}\equiv\lim_{N\to\infty}\frac{I_{\rm ub}}{N}=-\tfrac{1}{2}\ln(1-q^2)+\alpha\mathcal{D}_s(q)+\alpha\mathcal{D}_d(q,\delta), \tag{A33}$$

where q is given by the saddle point equation

$$\frac{\partial}{\partial q}i_{\rm ub}(q)=0. \tag{A34}$$

One should bear in mind that the term $\mathcal{D}_d$ is not present in the case of smooth learning. For supervised learning, at any given α the optimal choice for δ is solution of $(\partial/\partial\delta)\mathcal{D}_d(q,\delta)=0$, that is,

$$\frac{\delta(\alpha)}{2}=\epsilon_0(q). \tag{A35}$$

This is an implicit equation for δ, since q, the solution of Eq. (A34), depends on δ. One should note that $\epsilon_0(q)$ is the error rate that results from using a parameter $\mathbf{W}$ having an overlap q with the parameter $\mathbf{B}$ defining the rule. At the optimum $\mathcal{D}_d$ takes the nice expression of the binary entropy associated to the error rate $\epsilon_0(q)$:

$$\mathcal{D}_d=-\epsilon_0(q)\ln\epsilon_0(q)-[1-\epsilon_0(q)]\ln[1-\epsilon_0(q)]. \tag{A36}$$

As for the lower bound, the saddle point equation can be solved at least numerically for any specific model. For the Gaussian case we have

$$\mathcal{D}_s(q)=\frac{1}{2}\frac{(1-\sigma^2)^2}{\sigma^2}(1-q^2), \tag{A37}$$

and for the standard perceptron $\mathcal{D}_s(q)=0$ and $\mathcal{D}_d(q)$ is given by (A31) and (A17). The upper bounds are shown on Fig. 2 for these two models.

Consider now the large α behavior. We have to take the limit $q\to1$. From Eqs. (A30) and (A31) using Eqs. (A18) and (A19) one gets

$$\mathcal{D}_s(q)\sim(1-q)\langle V'^2\rangle \tag{A38}$$

and

$$\mathcal{D}_d(q)\sim-\ln\left(1-\frac{\delta}{2}\right)-\sqrt{1-q}\frac{\sqrt{2}}{\pi}e^{-V(0)}\ln\frac{\delta}{2-\delta}. \tag{A39}$$

One then gets the behavior of the upper bound as α goes to infinity, for smooth learning

$$i_{\rm ub}\sim\tfrac{1}{2}\ln(\alpha e\langle V'^2\rangle) \tag{A40}$$

and for supervised learning

$$i_{\rm ub}\sim-\alpha\ln\left(1-\frac{\delta}{2}\right)+\ln\left(\frac{\alpha e}{\pi}e^{-V(0)}\ln\frac{2-\delta}{\delta}\right). \tag{A41}$$

In this limit the optimal value of δ is given by

$$\alpha=\frac{2}{\delta\ln(2/\delta)} \tag{A42}$$

which gives the upper bound

$$i_{\rm ub}=\ln\alpha+O(\ln\ln\alpha). \tag{A43}$$

One can see that both the upper and lower bounds have a qualitative behavior very similar to that of the mutual information not only at large α but also at finite α. In particular, when retarded classification occurs, they have a linear regime over a finite range of α values (see the related discussion Sec. IV B 1).

As for the lower bound, one can check that the leading term in the large p limit, for N finite, is correctly predicted by expressions (A40) and (A43), that is, $I\sim(N/2)\ln p$ and $I\sim N\ln p+O(N\ln\ln p)$, respectively.

[1] B. S. Clarke and A. R. Barron, IEEE Trans. Inf. Theory **36**, 453 (1990).

[2] M. Opper and W. Kinzel, in *Physics of Neural Networks*, edited by E. Domany J. L. van Hemmen, and K. Schulten (Springer, Berlin, 1995), p. 151.

[3] D. Haussler and M. Opper, in *Proceedings of the VIIIth Annual Workshop on Computational Learning Theory* (COLT95) (ACM, New York, 1995), pp. 402–411.

[4] J. Rissanen, IEEE Trans. Inf. Theory **42**, 40 (1996).

[5] N. Brunel and J.-P. Nadal, Neural Comput. **10**, 1731 (1998).

[6] R. E. Blahut, *Principles and Practice of Information Theory* (Addison-Wesley, Cambridge, Massachusetts, 1998).

[7] M. Opper and D. Haussler, Phys. Rev. Lett. **75**, 3772 (1995).

[8] J.-P. Nadal and N. Parga, Neural Comput. **6**, 489 (1994).

[9] R. Linsker, Computer **21**, 105 (1988).

[10] J. H. van Hateren, J. Comp. Physiol. A **171**, 157 (1992).

[11] J. J. Atick, Network **3**, 213 (1992).
[12] J.-P. Nadal and N. Parga, Network **5**, 565 (1994).
[13] S. I. Amari and N. Murata, Neural Comput. **5**, 140 (1992).
[14] M. Biehl and A. Mietzner, Europhys. Lett. **24**, 421 (1993); J. Phys. A **27**, 1885 (1994).
[15] T. Watkin and J.-P. Nadal, J. Phys. A **27**, 1899 (1994).
[16] P. Reimann and C. Van den Broeck, Phys. Rev. E **53**, 3989 (1996).
[17] A. Buhot and M. Gordon, Phys. Rev. E **57**, 3326 (1998).
[18] N. Brunel, J.-P. Nadal, and G. Toulouse, J. Phys. A **25**, 5017 (1992).
[19] M. Opper, Phys. Rev. E **51**, 3613 (1994).
[20] T. M. Cover, IEEE Trans. Electron. Comput. **14**, 326 (1965).
[21] D. Hansel, G. Mato, and C. Meunier, Europhys. Lett. **20**, 471 (1992).
[22] G. Györgyi and N. Tishby, in *Neural Networks and Spin Glasses*, Proceedings of the Statphys 17 Workshop on Neural Networks and Spin Glasses, edited by W. K. Theumann and R. Korberle (World Scientific, Singapore, 1990), pp. 3–36.
[23] P. Reimann, C. Van den Broeck, and G. J. Bex, J. Phys. A **29**, 3521 (1996).
[24] C. Van den Broeck (unpublished).
[25] M. Opper, Phys. Rev. Lett. **72**, 2113 (1994).
[26] V. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 1995).
[27] A. Buhot and M. Gordon, Phys. Rev. Lett. (to be published).