Statistical-ensemble theory of redundancy reduction and the duality between unsupervised and supervised neural learning

Gustavo Deco* and Bernd Schürmann
Siemens AG, Corporate Research and Development, ZFE T SN 4 Otto-Hahn-Ring 6, 81739 Munich, Germany
(Received 13 April 1995)

The aim of this paper is twofold. First, we derive a statistical-mechanics-based model of unsupervised learning defined by redundancy reduction between the output components of neural nets and entropy conservation from inputs to outputs. We obtain an approximate expression for the probability distribution of the output components for a new data point, which is essentially determined by the probability distribution given by the best network of neural ensembles and by the square root of the ratio between the determinants of the Fisher information without and with the new point. Second, we pose the problem of supervised learning as an unsupervised one. The ensemble theory derived for unsupervised learning results in one for supervised learning by using the ensemble theory based on the maximum-likelihood principle. An upper bound for the prediction probability of a new point not included in the training data is derived. This upper bound is essentially given by the ratio between the Fisher information, determined for the training sets without and with the inclusion of the new point. This upper bound may be used as a mechanism to decide actively on the novelty of new data (mechanism of query learning). An illustrative example is given for the case where the training error possesses a Gaussian distribution.

PACS number(s): 87.10.+e, 02.50.Ph, 02.50.Wp, 05.90.+m

I. INTRODUCTION

The problem of learning and generalization from examples by using neural networks has been treated in the framework of both statistics [1–3] and statistical physics [4–8]. In the statistical physics approach an ensemble of neural networks is used to address the problem of generalization of learning from a finite number of noisy training example. The ensemble treatment of neural networks [4–8] assumes the final model to be probabilistic, built by an integration of single models weighted with the corresponding probability distribution. Gibbs's distribution is obtained from the maximum entropy principle [8], or alternatively by imposing the equivalence of the minimum error and the maximum-likelihood criteria for training the network [4–7]. Learning is defined as a maximization of the Kullback-Leibler entropy of the network distribution in parameter space. It reduces the ensemble volume, with the initial volume being fixed by the *a priori* distribution. Unfortunately, the integration in parameter space needed for deriving the partition function is impossible to perform in the case of standard neural network models such as multilayer perceptrons. Some approximations employing the replica method are possible (cf. Ref. [4]). Tishby and co-workers [5–7] use the annealed approximated which, though simpler, yields the correct qualitative behavior in many cases.

Here we focus our attention on the problem of both unsupervised and supervised learning by neural networks from given examples.

The task in supervised learning consists in approximating a general continuous input-output relation via a nonlinear parametric model, i.e., a neutral network. The supervised learning process considers a set of training examples for finding the relation between input and output. The ultimate goal is to *generalize,* i.e., to find a model that describes the relation between input and output for all possible examples. The quality of modeling clearly depends on the architecture used and on the complexity of the task. The classical definition of learning considers a cost function which is a measure of the error on the training examples, the aim being to define a procedure that finds the set of parameters (weights of the network) minimizing this cost function. Statistical physics relates this task to the problem of finding the state of the system (given by the parameters) that minimizes the energy (cost function). Alternatively, if a multiple hypothesis explains the given task, i.e., if several possible different networks model the given relation, the problem of learning is defined as the decrease of the hypothesis (model) entropy. In other words, the statistical-mechanics approach basically models a given relation by using an ensemble of models combined and weighted in an optimal way. The combination of the networks in the ensemble may be defined by using the maximum entropy principle or the Bayes theorem. Using these methods, it is possible to arrive at the Gibbs distribution which describes the *a posteriori* probability of each model given the training data. The only free parameter that can still be adjusted is the temperature of the ensemble, and it is related to the stochasticity of the input-output relation which in general is given by the noise. The learning procedure then aims at finding the right temperature of the ensemble. This temperature regulates the combination of models describing the final ensemble model. As we will see, the basic prob-

*FAX: +49 89 636 49767. Electronic address: Gustavo.Deco@zfe.siemens.de

lem is to integrate the ensemble of networks. The statistical-mechanics formulation of supervised learning by neural networks was proposed by Denker *et al.* [9]. The general form was given by Tishby, Levin, and Solla [5] and Levin, Tishby, and Solla [4] and is the one we review in the second section of this paper.

On the other hand, unsupervised learning was formulated [10,11] by a neural implementation of the biological principle of redundancy reduction (cf. Barlow [12,13]). The brain performs a statistical decorrelation of the input environment in order to extract statistically independent relevant information. The goal of redundancy reduction is to factorize the output probability distribution without losing information. In the linear case Barlow's principle yields a learning rule that performs a principal component analysis (PCA). In fact, PCA can be derived as a linear transformation which conserves transmission of information and minimizes the mutual information between the outputs by decorrelating them. Deco and Schürmann [10] devised an architecture and a learning paradigm for the unsupervised extraction of statistical correlations, performing Barlow's unsupervised learning in the most general fashion and implementing a nonlinear independent component analysis.

In this paper we formulate a statistical-mechanics-based theory for unsupervised learning as modeled in Ref. [10]. Furthermore, we may at the same time pose the problem of supervised learning as an unsupervised learning one, such that we obtain an ensemble theory for supervised learning based on the maximum-likelihood principle. We remark that there exists other work in the literature on the duality between unsupervised and supervised learning (e.g., Ref. [14]), however, with approaches and aims which differ from ours in an essential way. An upper bound for the prediction probability of a new point not included in the training data is found. This upper bound is determined essentially by the ratio between the Fisher information for the training set and that for a set including the training data and the new point. It is possible to use this upper bound as a mechanism to decide actively on the novelty of new data and therefore to use it as a mechanism of query learning. Query learning aims at improving the generalization ability of a network that continuously learns by actively selecting optimal nonredundant data, i.e., data that contain new information for the model. Several investigations [15–20] in the field of neural networks address the topical problem of active data selection, known also under the more typical names of "query learning" and "on-line learning." The idea behind on-line learning is to actively decide whether new data should be learned or not, depending on the previously learned examples. This active selection is of fundamental importance for the generalization capabilities of the model, since by selecting data that are nonredundant, i.e., that carry new information for the model, overtraining of some region of input space where data are redundantly arriving in a large amount is avoided. For clarification, let us suppose that due to the characteristics of the problem most of the time the data are clustered in a narrow region of input space, but the model should be valid in a wider region of input space where from time to time data are also present. If we train the network with all arriving data the model will in this case concentrate on the redundant data, trying to use all available resources for modeling this region and getting overtrained at the cost of forgetting those regions which have little data. Another possible scenario is the case where off-line data are available, and in order to build a model with good interpolation capability over the whole input space, data should be chosen ("experiment design" [21–24]) so that they are nonredundant. Therefore the essential problem is to define a measure of informativeness of new data for a given model architecture and a given set of viewed example patterns.

In Sec. II we review the ensemble theory for unsupervised learning. Section III is devoted to the formulation of an ensemble theory for unsupervised factorial learning. In Sec. IV supervised and unsupervised learning are dually defined. This is used for a formulation of a statistical theory for supervised learning based on the maximum-likelihood principle, which we denote *statistical-mechanical theory of supervised factorial learning*. In Sec. V the measure of novelty defined in Sec. IV is specified to the Gaussian case. Numerical experiments are presented in Sec. VI, and final conclusions are given in Sec. VII.

II. PROBABILITY INFERENCE WITH AN ENSEMBLE OF NEURAL NETWORKS

Let us consider a feedforward neural network parametrized by a weight vector $\vec{x}$. As notations, we use $\vec{x}$ for the N-dimensional input vector, $\vec{y}$ for the M-dimensional teacher output vector, and $\vec{f}(\vec{x},\vec{w})$ for the M-dimensional actual output vector of the network. The statistical physics approach models the input-output relation by considering an *ensemble* of neural networks instead of only one. The goal of supervised learning is, given a set of P example patterns

$$D^{(P)}=\{(\vec{x}^{(q)},\vec{y}^{(q)}),1\le q\le P\}\ , \tag{1}$$

to model the probability $p(\vec{y}/\vec{x},D^{(P)})$ of predicting a new input-output pair $(\vec{x},\vec{y})$. Let us define the conditional probability $p(\vec{y}/\vec{x},\vec{w})$ as the likelihood of the pair $(\vec{y},\vec{x})$ for the network $\vec{w}$. In the ensemble approach the model consists of a combination of all possible networks. Mathematically we can express the prediction probability of the ensemble of neural networks by the equation

$$p(\vec{y}/\vec{x},D^{(P)})=\int p(\vec{w}/D^{(P)})p(\vec{y}/\vec{x},\vec{w})d\vec{w}\ , \tag{2}$$

where $p(\vec{w}/D^{(P)})$ is called the *a posteriori* probability of the ensemble in parameter space. Clearly, knowing this *a posteriori* probability, no learning process is necessary for defining the final model. The *a posteriori* probability $p(\vec{w}/D^{(P)})$ in parameter space may be defined by means of the maximum entropy principle. The essential idea is that a prescribed constraint should be satisfied on the training data set (for example, the additive quadratic error should be minimal). Otherwise, the model should be combined without assuming extra information, i.e., the ensemble entropy

$$\int p(\vec{w}/D^{(P)})\ln[p(\vec{w}/D^{(P)})]d\vec{w} \tag{3}$$

should be maximal (maximum entropy distribution [25]). The solution to this problem is the Gibbs distribution which in this case is defined by

$$P(\vec{w}/D^{(P)})=\frac{\exp-\beta\sum_{q=1}^{P}\|\vec{y}^{(q)}-\vec{f}(\vec{x}^{(q)},\vec{w})\|^2}{Z(P)}\ , \tag{4}$$

where the normalization factor Z is the partition function of the ensemble of networks and is defined by

$$Z(P)=\int\exp\left[-\beta\sum_{q=1}^{P}\|\vec{y}^{(q)}-\vec{f}(\vec{x}^{(q)},\vec{w})\|^2\right]d\vec{w}\ . \tag{5}$$

The probability of the new input-output pair $(\vec{x},\vec{y})$ for the network $\vec{w}$ is modeled by the Gaussian distribution

$$p(\vec{y}/\vec{x},\vec{w})=\frac{e^{-\beta\|\vec{y}-\vec{f}(\vec{x},\vec{w})\|^2}}{(\sqrt{\pi/\beta})^M}\ . \tag{6}$$

Hence the final model is given by

$$p(\vec{y}/\vec{x},D^{(P)})=\int\frac{\exp\left[-\beta\sum_{q=1}^{P}\|\vec{y}^{(q)}-\vec{f}(\vec{x}^{(q)},\vec{w})\|^2\right]}{Z(P)}\times\frac{e^{-\beta\|\vec{y}-\vec{f}(\vec{x},\vec{w})\|^2}}{(\sqrt{\pi/\beta})^M}d\vec{w}\ . \tag{7}$$

The problem is seen to be reduced to the calculation of the partition function Z, which, due to the nonlinearity in the parameters of the network, in most cases is nonintegrable without approximations. It is important to remark that the only remaining free parameter is β, which in thermodynamics is associated with the inverse of the ensemble temperature. This parameter is related to the stochasticity of the data, meaning that the problem is to have a model which possesses the same stochasticity as the data. A special case is the error-free learning problem, i.e., the problem is noise-free and realizable. In this case, the result of Denker *et al.* [9] can be recovered in the limit $\beta\to\infty$, i.e., when the ensemble of networks yields a deterministic model.

III. STATISTICAL THEORY OF UNSUPERVISED FACTORIAL LEARNING

In this section, we formulate an ensemble theory for unsupervised learning by making use of the basic principles discussed above. We employ a single-layer architecture that attempts to extract correlations. The architecture is always reversible, conserves the volume, and therefore conserves the transmitted information. In general the environment is non-Gaussian distributed and nonlinearly correlated. The learning rule decorrelates statistically the elements of the output by minimizing the mutual information between the output components.

The aim of this section is to derive a statistical-mechanics-based model of the unsupervised learning mechanism devised by Deco and Schürmann [10] and by Deco and Brauer [11]. We concentrate on the one-layer volume-conserving triangular architecture of Fig. 1, and denote the n-dimensional input and output vectors for the unsupervised architecture of Fig. 1 by $\vec{\xi}$ and $\vec{\Upsilon}$, respectively. The output vector $\vec{\Upsilon}$ is defined by

$$\Upsilon_i=\xi_i+f_i(\xi_0,\ldots,\xi_j,\vec{\omega}_i)\ ,\quad\text{with } j<i\ , \tag{8}$$

$\vec{\omega}_i$ being the parameter vector which determines the parametrical function (for example, a neural network) f_i. Note that independent of the functions f_i, the network is always volume conserving, i.e., the determinant of the Jacobi matrix of Eq. (8) is equal to unity. In particular, f_i can be calculated by another neutral network, by a sigmoid neutron, by polynomials (higher order neurons), etc. The triangular structure of this network not only assures conservation of entropy in the transmission from the inputs to the outputs but also a transformation that attempts to decorrelate a component from only the past components, which is the kind of correlation that we need in time series modeling.

Let us denote the entropy of a random variable X by $h(X)$, i.e.,

$$h(X)=-\int dx\,p(x)\ln[p(x)]\ , \tag{9}$$

$p(x)$ being the probability distribution of the random variable X. Unsupervised learning minimizes the redundancy (i.e., the mutual information) at the output components given by

$$R=\sum_{j=1}^{n}h(\Upsilon_j)-h(\vec{\Upsilon})\ . \tag{10}$$

By using the fact that entropy is conserved [due to the fact that the transformation $\vec{\Upsilon}=\vec{U}(\vec{\xi})$ has a Jacobi determinant equal to unity], i.e.,

$$h(\vec{\Upsilon})=h(\vec{\xi})=\text{const}\ , \tag{11}$$

minimization of the redundancy is reduced to minimizing the term $\sum_{j=1}^{n}h(\Upsilon_j)$.

We are now able to formulate an ensemble theory for unsupervised learning. The training set consists of P example patterns,

$$T^{(P)}=\{\vec{\xi}^{(q)},1\le q\le P\}\ . \tag{12}$$

The ensemble of networks given by different parameters $\vec{w}=\{\vec{w}_1,\ldots,\vec{w}_n\}$ is weighted by again using the maximum entropy principle. In this case the macroscopic constraint is the minimization of redundancy, so that the Gibbs function now is defined by

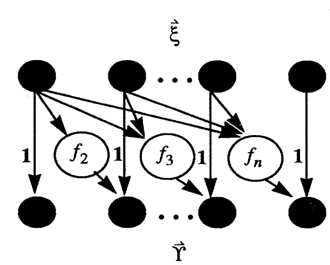

FIG. 1. Volume-conserving triangular architecture for unsupervised learning.

$$p(\vec{w}/T^{(P)})=\frac{\exp\left[\beta\sum_{i=1}^{P}\sum_{j=1}^{n}\ln[p(\Upsilon_j^{(i)}/\vec{w})]\right]}{Z(P)}, \tag{13}$$

with the partition function given by

$$Z(P)=\int\exp\left[\beta\sum_{i=1}^{P}\sum_{j=1}^{n}\ln[p(\Upsilon_j^{(i)}/\vec{w})]\right]d\vec{w}\ . \tag{14}$$

We perform a Taylor expansion of the exponent around $\vec{w}_P$ defined as that point where the empirical entropy multiplied by the number of patterns P,

$$H^{(P)}=\sum_{i=1}^{P}\sum_{j=1}^{n}\ln[p(\Upsilon_j^{(i)}/\vec{w})]\ , \tag{15}$$

is minimal. At this point the gradient is equal to zero, i.e.,

$$\vec{\nabla}H^{(P)}|_{\vec{w}_p}=\vec{0}\ . \tag{16}$$

Hence the Taylor expansion up to second order is given by

$$H^{(P)}\cong H^{(P)}|_{\vec{w}_P}+\tfrac{1}{2}(\vec{w}-\vec{w}_P)^T\vec{\nabla}\vec{\nabla}H^{(P)}|_{\vec{w}_P}(\vec{w}-\vec{w}_P)\ , \tag{17}$$

where $\vec{\nabla}\vec{\nabla}A$ denotes the Hessian matrix of A. Inserting Eq. (17) into Eq. (14), the integrand adopts a Gaussian form and therefore the integral can be easily calculated, yielding

$$Z(P)\cong e^{\beta H^{(P)}}|_{\vec{w}_P}(2\pi)^{D/2}[\det(\beta F^{(P)})]^{-1/2}\ , \tag{18}$$

where $D=\dim(\vec{w})$ and the non-negative definite matrix $F^{(P)}$ is defined by

$$F^{(P)}=-\vec{\nabla}\vec{\nabla}H^{(P)}|_{\vec{w}_P}=-\sum_{i=1}^{P}\sum_{j=1}^{n}\vec{\nabla}\vec{\nabla}\ln[p(\Upsilon_j^{(i)}/\vec{w}_P)]\ . \tag{19}$$

For $\vec{P}\infty$ we obtain

$$\begin{aligned}F^{(P)}&\to-\int p(\vec{e}/\vec{w})\vec{\nabla}\vec{\nabla}\ln[p(\vec{e}/\vec{w})]d\vec{e}\\&=\int p(\vec{e}/\vec{w})\vec{\nabla}\ln p(\vec{e}/\vec{w})\vec{\nabla}\ln[p(\vec{e}/\vec{w})]d\vec{e}\ .\end{aligned} \tag{20}$$

The integrals in (20) are, according to Ref. [26], the Fisher information which is a measure of the amount of "information" about $\vec{w}$ that is present in the data. We need to calculate also $Z(P+1)$ because we want to study the effect of adding a new pattern. To do so we make a Taylor expansion of $H^{(P+1)}$ around the point $\vec{w}_P$ defined above. We obtain

$$\begin{aligned}H^{(P+1)}=&H^{(P+1)}|_{\vec{w}_P}+(\vec{w}-\vec{w}_P)^T\vec{\nabla}H|_{\vec{w}_P}\\&+\tfrac{1}{2}(\vec{w}-\vec{w}_P)^T\vec{\nabla}\vec{\nabla}H^{(P+1)}|_{\vec{w}_P}(\vec{w}-\vec{w}_P)+\cdots\ ,\end{aligned} \tag{21}$$

where

$$H=\sum_{j=1}^{n}\ln[p(\Upsilon_j/\vec{w}_P)] \tag{22}$$

is defined at the new point. Bringing $H^{(P+1)}$ into a binomial form (apart from terms independent of $\vec{w}$) leads to

$$\begin{aligned}H^{(P+1)}=&H^{(P+1)}|_{\vec{w}_P}+\tfrac{1}{2}(\vec{w}-\vec{w}_P+\vec{b})^T\vec{\nabla}\vec{\nabla}H^{(P+1)}|_{\vec{w}_P}\\&\times(\vec{w}-\vec{w}_P+\vec{b})\\&-\tfrac{1}{2}(\vec{\nabla}H|_{\vec{w}_P})^T(\vec{\nabla}\vec{\nabla}H^{(P+1)}|_{\vec{w}_P})^{-T}(\vec{\nabla}H|_{\vec{w}_P})\ ,\end{aligned} \tag{23}$$

with

$$\vec{b}=-(\vec{\nabla}\vec{\nabla}H^{(P+1)}|_{\vec{w}_P})^{-1}\vec{\nabla}H|_{\vec{w}_P}\ . \tag{24}$$

The integrand of $Z(P+1)$ is again a Gaussian, and after performing the integration we obtain

$$\begin{aligned}Z(P+1)=&e^{\beta H^{(P+1)}}|_{\vec{w}_p}(2\pi)^{D/2}[\det(\beta F^{(P+1)})]^{-1/2}\\&\times(e^{(1/2)\beta[\vec{g}^T(F^{(P+1)})^{-T}\vec{g}]})\ ,\end{aligned} \tag{25}$$

where

$$\vec{g}=\vec{\nabla}H|_{\vec{w}_P}\ . \tag{26}$$

Now we are in a position to write the probability distribution $p(\Upsilon,T^{(P)})$ of the new point as

$$\begin{aligned}p(\Upsilon,T^{(P)})=\int&\frac{\exp\left[\beta\sum_{i=1}^{P}\sum_{j=1}^{n}\ln[p(\Upsilon_j^{(i)}/\vec{w})]\right]}{Z(P)}\\&\times p(\vec{\Upsilon}/\vec{w},\beta)d\vec{w}\ .\end{aligned} \tag{27}$$

We assume for the probability of each network the *escort* distribution

$$p(\vec{\Upsilon}/\vec{w},\beta)=\frac{e^{\beta\ln[p(\vec{\Upsilon}/\vec{w})]}}{z}\ . \tag{28}$$

According to Ref. [27], escort distributions have the ability to scan the structure of the original probability distribution. The reason for using the particular form (28) is to write the distribution of $\vec{\Upsilon}$ in the compact form

$$\begin{aligned}p(\vec{\Upsilon},T^{(P)})=&\frac{Z(P+1)}{zZ(P)}\\\cong&\left[\frac{1}{z}e^{\beta H+(\beta/2)\vec{g}^T(F^{(P+1)})^{-T}\vec{g}}\right]\\&\times\det{}^{1/2}[F^{(P)}(F^{(P+1)})^{-1}]\ .\end{aligned} \tag{29}$$

Thus we have derived an approximate expression for the probability distribution of the output components which is essentially based on the probability distribution given by the best network (the one with parameter vector $\vec{w}_P$) and by the square root of the ratio between the determinants of the Fisher information without and with inclusion of the new point.

In the next section we will use these results for obtaining an ensemble theory of supervised learning based on the maximum-likelihood principle.

IV. DUALITY BETWEEN UNSUPERVISED AND SUPERVISED LEARNING

We pose the problem of supervised learning as an unsupervised one. As we will demonstrate, for unsupervised learning the ensemble theory derived above then results in a theory for supervised learning by making use of ensembles based on the maximum-likelihood principle. The input vector $\vec{\xi}$ for the unsupervised architecture of Fig. 2 is defined to be composed of two components $\vec{x}$ and $\vec{y}$ of dimensions N and M, respectively, i.e., $\vec{\xi}=\{\vec{x},\vec{y}\}$. These two vectors are related through a probability distribution $\rho(\vec{x},\vec{y})$, so that they can be regarded as input and output of a relation to be learned by supervised learning. The input vector $\vec{\xi}$ is given empirically by the set of the training data, as defined by Eq. (1). Let us denote the output of the triangular architecture by $\vec{\Upsilon}$ which is also composed of two vectors such that $\vec{\Upsilon}=\{\vec{x},\vec{e}\}$. The network output component $\vec{e}$ is defined by

$$\vec{e}=\vec{y}-\vec{f}(\vec{x},\vec{w}) , \tag{30}$$

where in supervised learning $\vec{e}$ is the error and $\vec{w}$ is the parameter vector which describes the general function $\vec{f}$. The maximum-likelihood principle for supervised learning requires $\vec{w}$ to be chosen so that the empirical likelihood

$$L=\frac{1}{P}\sum_{i=1}^{P}\ln[p(\vec{y}^{(i)}/\vec{x}^{(i)},\vec{w})] \tag{31}$$

is maximal. In Eq. (31), the conditional probability $p(\vec{y}^{(i)}/\vec{x}^{(i)},\vec{w})$ should be regarded as a measure of the compatibility of the pairs $(\vec{x}^{(i)},\vec{y}^{(i)})$. On the other hand, as discussed above, the goal of unsupervised learning is redundancy minimization. The architecture of Fig. 2 can only minimize the redundancy inherent in the relation between the vectors $\vec{x}$ and $\vec{y}$, i.e., it aims to exact the correlations between these vectors, which is the goal of supervised learning. In fact, unsupervised learning minimizes the redundancy at the output components given by

$$R=\sum_{j=1}^{N}h(x_j)+\sum_{j=1}^{M}h(e_j)-h(\vec{\Upsilon}) . \tag{32}$$

By using the fact that entropy is conserved, i.e.,

$$h(\vec{\Upsilon})=h(\vec{\xi})=\text{const} , \tag{33}$$

and assuming that the distribution of $\vec{x}$ is stationary, i.e.,

$$\sum_{j=1}^{N}h(x_j)=\text{const} , \tag{34}$$

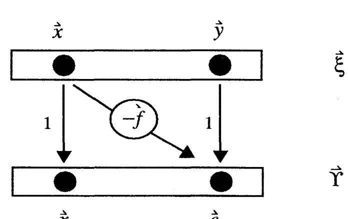

FIG. 2. Unsupervised architecture for supervised learning.

minimization of redundancy is reduced to minimization of the term $\sum_{j=1}^{M}h(e_j)$. Due to the fact that

$$h(\vec{e})\le\sum_{j=1}^{M}h(e_j) \tag{35}$$

and taking into account that

$$h(\vec{e})=-L \tag{36}$$

because of

$$\begin{aligned}p(\vec{y}/\vec{x},\vec{w})&=p(\vec{y}-\vec{f}(\vec{x})\equiv 0/\vec{x},\vec{w})\\&=p(\vec{e}\equiv 0/\vec{x},\vec{w})=p(\vec{e}/\vec{x},\vec{w}) ,\end{aligned} \tag{37}$$

minimization of redundancy of the output components $\vec{\Upsilon}$ is equivalent to maximization of the likelihood L. Put differently, maximum-likelihood supervised learning is equivalent to minimizing the entropy of the error, which is the goal of unsupervised reduction of redundancy. This is the dual formulation of supervised and unsupervised learning.

We now make use of the results previously obtained for the ensemble theory of unsupervised learning. Writing the prediction probability of a new point as

$$\begin{aligned}p(\vec{y}/\vec{x},D^{(P)})&=p(\vec{e}/\vec{x},D^{(P)})\\&=\int\frac{\exp\left[\beta\sum_{i=1}^{P}\ln[p(\vec{e}^{(i)}/\vec{x}^{(i)},\vec{w})]\right]}{Z(P)}\\&\quad\times p(\vec{e}/\vec{x},\vec{w},\beta)d\vec{w} ,\end{aligned} \tag{38}$$

and again assuming for each network the escort distribution

$$p(\vec{e}/\vec{x},\vec{w},\beta)=\frac{e^{\beta\ln[p(\vec{e}/\vec{x},\vec{w})]}}{z} , \tag{39}$$

we obtain for the distribution of $\vec{y}$

$$\begin{aligned}p(\vec{y}/\vec{x},D^{(P)})\cong&\left[\frac{1}{z}e^{\beta H'+(\beta/2)\vec{g}'^{T}(F'^{(P+1)})^{-1}\vec{g}'}\right]\\&\times\det{}^{1/2}[F'^{(P)}(F'^{(P+1)})^{-1}]\end{aligned} \tag{40}$$

and therefore the probabilities

$$p(\vec{y}/\vec{x},D^{(P)})\le p(\vec{y}/\vec{x},\vec{w}_P,\beta)\left[\frac{\det F'^{(P)}}{\det F'^{(P+1)}}\right]^{1/2} . \tag{41}$$

In the last two equations,

$$F'^{(P)}=-\sum_{i=1}^{P}\vec{\nabla}\vec{\nabla}\ln[p(\vec{e}^{(i)}/\vec{x}^{(i)},\vec{w}_P)] , \tag{42}$$

$$H'=\ln[p(\vec{e}/\vec{x},\vec{w}_P)] , \tag{43}$$

and

$$\vec{g}'=\vec{\nabla}H'|_{\vec{w}_P} . \tag{44}$$

Hence we have obtained an upper bound for the prediction probability of new data given P training data. The upper bound is essentially determined by the square root of the ratio between the determinants of the Fisher information without and with inclusion of the new point. The factor $p(\vec{y}/\vec{x},\vec{w}_P,\beta)$ is the probability of observing the new pair given by the best network (the one with parameter vector $\vec{w}_P$). The square root of the ratio between the determinants of the Fisher information provides us with a measure of how much the reliability of the best network should be reduced. We call the negative logarithm of this quantity the novelty measure $\mathcal{N}(P)$, i.e.,

$$\mathcal{N}(P)=-\frac{1}{2}\ln\left[\frac{\det F'^{(P)}}{\det F'^{(P+1)}}\right] . \tag{45}$$

This information is a consequence of the use of the ensemble approach.

In summary, the statistical-mechanics-based theory of unsupervised learning by redundancy reduction with a volume-conserving network has been formulated and used for the improvement of the standard ensemble theory of supervised learning by exploiting the duality between the two learning paradigms.

V. NOVELTY DETECTION—THE GAUSSIAN CASE

In this section, we focus on the modeling by an ensemble of feedforward neural networks. In this case the networks may be defined as multilayer perceptrons, e.g.,

$$f_i(\vec{x}^{(q)},\vec{w}_i)=\sigma(\vec{w}_i'\cdot\sigma(\vec{w}_i''\cdot\vec{x}^{(q)})) ,\quad \text{with } \vec{w}_i=\{\vec{w}_i',\vec{w}_i''\} , \tag{46}$$

where σ is a sigmoidal function. In general $(\vec{x}^{(q)},\vec{w}_i)$ may be arbitrary nonlinear functions of the inputs and of the parameters. We specialize to a Gaussian distribution so that

$$h(\vec{e})=\tfrac{1}{2}\ln(2\pi e)^M\det(C)=-E^{(P)} , \tag{47}$$

where C is the covariance matrix of the error components. Applying Hadamard's inequality [26]

$$\det(C)\le\prod_{i=1}^{M}C_{ii} \tag{48}$$

yields

$$h(\vec{e})\le\sum_{i=1}^{M}\ln C_{ii}\le\sum_{i=1}^{M}C_{ii} , \tag{49}$$

and therefore the constraint now is to minimize the sum of the variances so that

$$H''^{(P)}=-\sum_{q=1}^{P}\sum_{i=1}^{M}[y_i^{(q)}-f_i(\vec{x}^{(q)},\vec{w}_i)]^2 . \tag{50}$$

It is important to remark that in general the constraint introduced by the unsupervised view of supervised learning based on the maximum-likelihood principle is much tougher than that based on the quadratic error. In fact, the entropy calculated assuming a Gaussian distribution $h(\vec{e}\,')$ with a covariance matrix C identical to that of the real distribution is an upper bound of the real entropy [26], i.e.,

$$h(\vec{e})\le h(\vec{e}\,') , \tag{51}$$

which means that

$$\sum_{i=1}^{P}\ln[p(\vec{e}^{(i)}/\vec{x}^{(i)},\vec{w})]\le\sum_{i=1}^{P}\|\vec{y}^{(i)}-\vec{f}(\vec{x}^{(i)},\vec{w})\|^2 . \tag{52}$$

The right-hand-side of Eq. (52) is the constraint used in the ensemble theory of supervised learning. In other words, the formulation obtained by the unsupervised ensemble theory permits us to describe the supervised one in a more precise way.

MacKay [1] made an additional approximation which consists in neglecting those terms which contain the second order in the derivative of the network function, i.e.,

$$\vec{\nabla}\vec{\nabla}H''^{(P+1)}\cong\vec{\nabla}\vec{\nabla}H''^{(P)}+2\vec{g}\,\vec{g}^{T},\quad g_i=\Delta_i f_i . \tag{53}$$

With this approximation we obtain for the prediction likelihood

$$p(\vec{y}/\vec{x},D^{(P)})\cong\frac{1}{z}e^{-\beta\{\|\vec{y}-\vec{f}(\vec{x},\vec{w}_P)\|^2/\det[I+F''(F''^{(P)})^{-1}]\}-(1/2)\ln\{\det[I+F''(F''^{(P)})^{-1}]\}} , \tag{54}$$

where

$$F''^{(P)}=\sum_{q=1}^{P}\vec{\nabla}\vec{\nabla}\left[\sum_{i=1}^{M}[y_i^{(q)}-f_i(\vec{x}^{(q)},\vec{w}_{P_i})]^2\right] , \tag{55}$$

$$F''=\vec{\nabla}\vec{\nabla}\left[\sum_{i=1}^{M}[y_i-f_i(\vec{x},\vec{w}_{P_i})]^2\right] , \tag{56}$$

and $\vec{w}_P$ are the best parameters, i.e., where $\vec{\nabla}H''^{(P)}=0$. The novelty measure defined by the Kullback-Leibler entropy in parameter space,

$$\mathcal{N}(P)=\int p(\vec{w}/D^{(P+1)})\ln\left[\frac{p(\vec{w}/D^{(P+1)})}{p(\vec{w}/D^{(P)})}\right]d\vec{w} \tag{57}$$

can in MacKay's approximation also be calculated analytically yielding

$$\mathcal{N}(P)=\tfrac{1}{2}\ln\{\det[1+F''(F''^{(P)})^{-1}]\} . \tag{58}$$

This is identical to the second term of the exponent of (54) and is a special case of Eq. (45). In Eq. (54) we have used the fact that

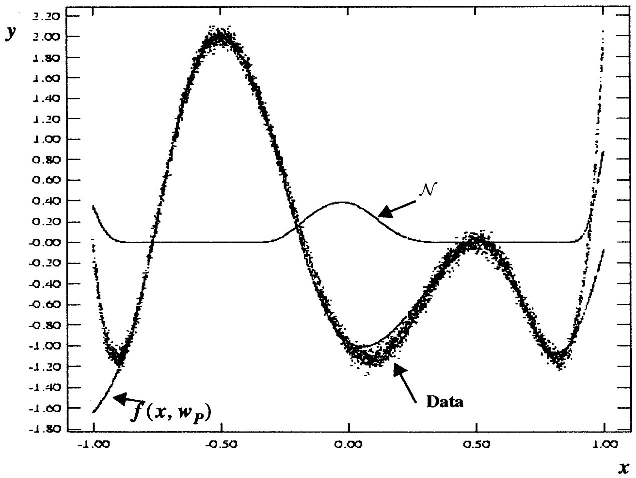

FIG. 3. Novelty detection with an ensemble of multilayer perceptrons.

$$\det[I+F''(F''^{(P)})^{-1}]=\frac{1}{[1-2\vec{g}^{\,T}(F''^{(P+1)})^{-T}\vec{g}\,]}\ , \qquad (59)$$

which can be verified by using the relation of Fedorov [21],

$$\det(A+a\vec{c}\,\vec{c}^{\,T})=\det(A)(1+a\vec{c}^{\,T}A^{-1}\vec{c}\,)\ . \qquad (60)$$

The unity matrix is denoted by I. Equation (54) yields the essential probability of the problem, namely, the prediction probability of a new point given P examples, in an analytic form. In the next section we apply the results obtained to the problem of query learning.

VI. NUMERICAL EXPERIMENTS

We apply the approximate model developed in the preceding section to the case of an ensemble of multilayer perceptrons with a single hidden layer and sigmoidal functions. The problem consists in modeling the function

$$y(x)=-1-3x+18x^2+4x^3-48x^4+32x^6+\nu \qquad (61)$$

by using only data in some regions and then checking the novelty of the data in all possible regions. In Eq. (61), ν stands for Gaussian noise with variance 0.05. The architecture used contains one input, ten hidden units, and one output. The temperature of the ensemble is chosen according to the noise as $\beta=0.1$. The training data are selected in the regions

$$-0.9\le x\le -0.3\ ,\quad 0.3\le x\le 0.9 \qquad (62)$$

and the novelty of the data is tested in the region $-1\le x\le 1$. Figure 3 displays the results of the simulations. The original data, the prediction of the ensemble model, and the novelty measures $\mathcal{N}(P)$ of Eq. (58) are plotted. In the regions not included in the training set, a large value of the novelty measure is observed, indicating the information content of the new data for the ensemble model.

VII. CONCLUSIONS

A statistical-mechanics-based model of unsupervised learning defined by redundancy reduction at the output components and entropy conservation from inputs to outputs has been derived. We have obtained an approximate expression for the probability distribution of the output components which is essentially determined by the probability distribution given by the best network and by the square root of the ratio between the determinants of the Fisher information without and with inclusion of the new point.

Furthermore, the problem of supervised learning has been posed as an unsupervised one. The ensemble theory derived for unsupervised learning then results in one for supervised learning by using ensemble theory based on the maximum-likelihood principle. An upper bound for the prediction probability of a new point not included in the training data is given. This upper bound is essentially determined by the ratio between the Fisher information given the training set and the one given a set which includes both the training data and the new point. This upper bound can be used as a mechanism to actively decide on the novelty of new data, and therefore it is a mechanism of query learning. Query learning aims to improve the generalization ability of a network that continuously learns by actively selecting optimal nonredundant data, i.e., data that contain new information for the model. An illustrative example has been given for the case where the training error possesses a Gaussian distribution. Needless to say, the key quantities of Sec. IV may be evaluated for non-Gaussian distributions as well, though with a larger numerical effort.

[1] D. MacKay, Neural Comput. **4**, 415 (1991).
[2] D. MacKay, Neural Comput. **4**, 448 (1992).
[3] H. White, Neural Comput. **1**, 425 (1989).
[4] E. Levin, N. Tishby, and S. Solla, Proc. IEEE **78**, 1568 (1990).
[5] N. Tishby, E. Levin, and S. Solla, in *Proceedings of the International Conference on Neural Networks* (IEEE, Washington, DC, 1989), Vol. 2, p. 403.
[6] N. Tishby, in *SFI Studies in the Sciences of Complexity,* edited by D. Wolpert (Addison-Wesley, Reading, MA, 1995), p. 215.
[7] N. Tishby, in *From Statistical Physics to Statistical Inference and Back,* Vol. 428 of *NATO Advanced Study Institute, Series C: Mathematical and Physical Sciences,* edited by P. Grassberger and J. P. Nadal (Kluwer, Dordrecht, 1995), p. 205.
[8] R. Meir and F. Fontanari, Physica A **200**, 644 (1993).
[9] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackel, and J. Hopfield, Complex Syst. **1**, 877 (1987).
[10] G. Deco and B. Schürmann, Phys. Rev. E **51**, 1780 (1995).
[11] G. Deco and W. Brauer, Neural Networks **8**, 525 (1995).
[12] H. Barlow, *National Physical Laboratory Symposium N.*

10, The Mechanization of Thought Processes (Her Majesty's Stationery Office, London, 1959).
[13] H. Barlow, Neural Comput. **1**, 295 (1989).
[14] J. P. Nadal and N. Parga, Neural Comput. **6**, 491 (1994).
[15] P. Sollich, Phys. Rev. E **49**, 4637 (1994).
[16] D. MacKay, Neural Comput. **4**, 590 (1992).
[17] E. Baum, IEEE Trans. Neural Networks, **2**, 5 (1991).
[18] J. N. Hwang, J. J. Choi, S. Oh, and R. Marks II, IEEE Trans. Neural Networks **2**, 131 (1991).
[19] W. Kinzel and P. Rujan, Europhys. Lett. **13**, 473 (1990).
[20] H. S. Seung, M. Opper, and H. Sompolinsky, in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (ACM, New York, 1992), p. 287.
[21] V. V. Fedorov, *Theory of Optimal Experiments* (Academic, New York, 1972).
[22] S. D. Silvey, *Optimal Design* (Chapman and Hall, London, 1980).
[23] P. Chaudhuri and P. A. Myklund, J. Am. Stat. Assoc. **88**, 538 (1993).
[24] J. Pilz, *Bayesian Estimation and Experimental Design in Linear Regression Models,* 2nd ed. (Wiley, Chichester, 1991).
[25] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[26] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[27] C. Beck and F. Schlögl, *Thermodynamics of Chaotic Systems,* Cambridge Nonlinear Science Series (Cambridge University Press, Cambridge, England, 1993).