Statistical physics theory of query learning by an ensemble of higher-order neural networks

Gustavo Deco* and Dragan Obradovic
Siemens AG, Corporate Research and Development, ZFE T SN 41, Otto-Hahn-Ring 6, 81739 Munich, Germany

(Received 13 March 1995)

Query learning aims to improve the generalization ability of a network that continuously learns by actively selecting nonredundant data, i.e., data that contain new information about the process. In this paper, we formulate the problem of query learning in the statistical mechanical framework. We define an information theoretic measure of the informativeness of the newly presented data in order to decide if the latter should be used for the model update or not. Only the data that carry new information about the underlying process are selected for learning. The informativeness of the new data is defined as the Kullback-Leibler distance between the likelihood of the *a posteriori* parameter distributions obtained before and after the inclusion of the new data point. In order to make the problem analytically solvable, we formulate the theory for the ensemble of higher-order neural networks, i.e., for the case of polynomial models. Comparison with other theoretical approaches is included. Simulations that validate the proposed theory are also included.

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

I. INTRODUCTION

In the last years, several research works [1–6] in the field of neural networks have addressed the interesting problem of active data selection, also known as "query learning." The idea behind query learning is to actively decide if a new data point is used for the model update or not, depending on the previously learned examples. This active selection is, therefore, of fundamental importance for the generalization capabilities of the model. The latter stems from the fact that it avoids overtraining in the input space regions where the excessive data is present. In order to further clarify the importance of data selection, let us suppose that due to the data acquisition characteristics most, but not all, of the observed data are clustered in a single region of the input space. Simultaneously, let us assume that the requirement for the derived model is to be valid in the whole input space. In this case, the network training will use all of its resources in the region where the data are concentrated and neglect the other regions where the few data are available. Another possible scenario is the case where off-line data are available. In the latter, in order to build a model with good interpolation capability over the whole input space, data points should be chosen ("experiment design" [7–10]) in such a way that the nonredundancy is avoided. The essential problem is, therefore, to define a measure of the new data informativeness for a given model architecture and a given set of previously seen example patterns.

Two principal approaches can be distinguished: the heuristic [3–5] and the approach derived from the minimization of an object function [1,2,6]. We concentrate in this paper on the second approach, i.e., we formulate an information-theory based objective function that measures the informativeness of a sequential data in the framework of the statistical mechanics. This novel informativeness measure of the incoming data is defined as the Kullback-Leibler distance between the probability of the output given the input and the past trained data (likelihood), and the same probability including the new data in the trained set. In this form the novelty is measured in the input-output space and not in the parameter space [2]. In order to formulate a theoretic approach to this problem, we have chosen the mechanical statistical ensemble formulation for supervised learning [11–15]. We apply the herein developed theory to polynomial models, i.e., higher-order neural networks. We do so due to the fact that in this case all integration can be performed analytically without approximation.

In Sec. II, we review the statistical mechanics approach to the ensemble of networks. Section III defines the novel informativeness measure while Sec. IV applies the latter to the case of polynomial models. Section V presents numerical results.

II. PROBABILITY INFERENCE WITH AN ENSEMBLE OF NETWORKS

Let us consider a feedforward neural network parametrized by a weight vector $\mathbf{w}$. We use the following notation: $\mathbf{x}$ for the N-dimensional input vector, $\mathbf{y}$ for the M-dimensional teacher output vector, and $\mathbf{f}(\mathbf{x},\mathbf{w})$ for the M-dimensional outputs of the network. The statistical physics approach models the input-output relation by considering an ensemble of neural networks. The goal of supervised learning given P examples

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

is to model the probability of predicting a new input-output pair $(\mathbf{x},\mathbf{y})$, for which we use the notation

*Electronic address: Gustavo.Deco@zfe.siemens.de
Tel.: +49 89 636 47373
Fax: +49 89 636 3320

$$p(\mathbf{y}/\mathbf{x},D^{(P)})\ . \tag{2}$$

Let us define the conditional probability $p(\mathbf{y}/\mathbf{x},\mathbf{w})$ as the likelihood of the pair $(\mathbf{y},\mathbf{x})$ for the network $\mathbf{w}$. In the ensemble theory, the model corresponds to a combination of individual networks from the ensemble. Mathematically we can express the prediction probability of the ensemble of neural networks by the equation

$$p(\mathbf{y}/\mathbf{x},D^{(P)})=\int p(\mathbf{w}/D^{(P)})p(\mathbf{y}/\mathbf{x},\mathbf{w})d\mathbf{w}\ , \tag{3}$$

where $p(\mathbf{w}/D^{(P)})$ is called *a posteriori* probability of the ensemble in the parameter space. It is clear that if we have this *a posteriori* probability, no learning process is necessary for defining the final model.

Let us first assume a model for the likelihood of a pair $(\mathbf{x},\mathbf{y})$ for the network $\mathbf{w}$. Levin, Tishby, and Solla [11] introduced a statistical description of the training process by postulating that the maximization of the likelihood should be equivalent to the minimization of the additive error. Therefore, a smooth and monotonic function ϕ should exist such that

$$\prod_{q=1}^{P} p(\mathbf{y}^{(q)}/\mathbf{x}^{(q)},\mathbf{w})=\phi\left[\sum_{q=1}^{P} e(\mathbf{y}^{(q)}/\mathbf{x}^{(q)},\mathbf{w})\right]\ , \tag{4}$$

where $e(\mathbf{y}^{(q)}/\mathbf{x}^{(q)},\mathbf{w})$ is a measure of the error of the network $\mathbf{w}$ for the pattern q. The only solution of Eq. (4) is given by Ref. [11]:

$$p(\mathbf{y}/\mathbf{x},\mathbf{w})=\frac{e^{-\beta e(\mathbf{y}/\mathbf{x},\mathbf{w})}}{z}\ , \tag{5}$$

with

$$z=\int e^{-\beta e(\mathbf{y}/\mathbf{x},\mathbf{w})}d\mathbf{y}\ . \tag{6}$$

Let us assume the quadratic error

$$e(\mathbf{y}/\mathbf{x},\mathbf{w})=\|\mathbf{y}-\mathbf{f}(\mathbf{x},\mathbf{w})\|^2\ , \tag{7}$$

which results in the Gaussian probability distribution of the input-output pair for the network $\mathbf{w}$:

$$p(\mathbf{y}/\mathbf{x},\mathbf{w})=\frac{e^{-\beta\|\mathbf{y}-\mathbf{f}(\mathbf{x},\mathbf{w})\|^2}}{(\sqrt{\pi/\beta})^M}\ . \tag{8}$$

The *a posteriori* probability in the parameter space $p(\mathbf{w}/D^{(P)})$ can be defined by means of the maximum entropy principle. The latter results in a probability distribution function whose entropy

$$-\int p(\mathbf{w}/D^{(P)})\ln[p(\mathbf{w}/D^{(P)})]d\mathbf{w} \tag{9}$$

is maximal. If the appropriate constraint is posed on the average ensemble error, the Gibbs distribution is obtained:

$$p(\mathbf{w}/D^{(P)})=\frac{e^{-\beta\sum_{q=1}^{P} e(\mathbf{y}^{(q)}/\mathbf{x}^{(q)},\mathbf{w})}}{Z(P)}\ , \tag{10}$$

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

$$Z(P)=\int e^{-\beta\sum_{q=1}^{P} e(\mathbf{y}^{(q)}/\mathbf{x}^{(q)},\mathbf{w})}d\mathbf{w}\ . \tag{11}$$

In the case where the quadratic error function [Eq. (7)] is used, the final model is given by

$$p(\mathbf{y}/\mathbf{x},D^{(P)})=\int\frac{\exp\left[-\beta\sum_{q=1}^{P}\|\mathbf{y}^{(q)}-\mathbf{f}(\mathbf{x}^{(q)},\mathbf{w})\|^2\right]}{Z(P)}\times\frac{e^{-\beta\|\mathbf{y}-\mathbf{f}(\mathbf{x},\mathbf{w})\|^2}}{z}d\mathbf{w}\ , \tag{12}$$

or, equivalently

$$p(\mathbf{y}/\mathbf{x},D^{(P)})=\frac{Z(P+1)}{zZ(P)}\ . \tag{13}$$

We see from Eq. (13) that the evaluation of the likelihood that the new point is measured accurately has been reduced to the calculation of the partition function Z. The latter is in most of the cases nonintegrable without approximations. We show in Sec. IV that the nonlinear models, which are linear in the parameters, allow exact integration of Eq. (11). Hence, it is possible to derive an analytical model of the ensemble of polynomials.

It is important to notice that the only free parameter is β, which is thermodynamic is associated with the inverse of the ensemble temperature. A special case is the error-free learning problem where the observables are noise-free and realizable. In this case, the result of Denker *et al.* [16] can be recovered in the lines $\beta\rightarrow\infty$, i.e., when the ensemble of networks yields a deterministic model.

III. DATA INFORMATIVENESS MEASURE

The informativeness measure $N(P)$ of the new data pair can now be defined as the Kullback-Leibler distance between the likelihoods of correctly representing the new output based on models with and without the new point, respectively,

$$N(P)=\sum_{i=1}^{P+1} K(p(\mathbf{y}/\mathbf{x}^{(i)},D^{(P)}),p(\mathbf{y}/\mathbf{x}^{(i)},D^{(P+1)}))\ , \tag{14}$$

where the Kullback-Leibler distance is defined as

$$K(p(\mathbf{y}/\mathbf{x}^{(i)},D^{(P)}),p(\mathbf{y}/\mathbf{x}^{(i)},D^{(P+1)}))=\int p(\mathbf{y}/\mathbf{x}^{(i)},D^{(P)})\ln\frac{p(\mathbf{y}/\mathbf{x}^{(i)},D^{(P)})}{p(\mathbf{y}/\mathbf{x}^{(i)},D^{(P+1)})}d\mathbf{y}\ . \tag{15}$$

MacKay [2] (see also Refs. [17 and 18]) proposes to measure the informativeness of a new data by just measuring its novelty in the parameter space, i.e., by the quantity

$$NW(P)=K(p(\mathbf{w}/D^{(P+1)}),p(\mathbf{w}/D^{(P)}))\ . \tag{16}$$

In other words, the novelty is measured only by the information gain in the parameter space. The relationship between these two measures is discussed at the end of Sec. IV.

IV. DATA INFORMATIVENESS IN HIGHER-ORDER NETWORK ENSEMBLES

The statistical mechanical theory presented in Sec. II is applied herein to the special case of higher-order neural networks, i.e., polynomial networks. In the case of modeling with an ensemble of polynomial, exact analytical results can be obtained due to the linearity of this model with respect to the parameters.

Let us define a higher-order neural network by the function $\mathbf{f}(\mathbf{x},\mathbf{w})$ given by

$$f_i(\mathbf{x}^{(q)},\mathbf{w}_i)=\mathbf{w}_i^T\cdot\mathbf{t},\quad 1\le i\le M \tag{17}$$

for each output i. $\mathbf{w}_i^T$ is the parameter vector for the output i and the vector $\mathbf{t}$ is defined by

$$(\mathbf{t}^{(q)})^T=(1,x_1^{(q)},\ldots,x_N^{(g)},x_1^{(q)}x_1^{(q)},\ldots,x_1^{(q)}x_N^{(g)},\ldots,x_N^{(g)}x_N^{(g)},\ldots,(x_N^{(g)})^R)\ , \tag{18}$$

where R is the order of the used polynomial network. The most important quantity to be calculated is the partition function, which can be written using Eqs. (11) and (7) as

$$Z(P)=\prod_{i=1}^{M}\int e^{-\beta\sum_{q=1}^{P}(y_i^{(q)}-f_i(\mathbf{x}^{(q)},\mathbf{w}_i))^2}d\mathbf{w}_i\ . \tag{19}$$

We use the Taylor expansion of the exponent around the point $\mathbf{w}_{P_i}$ defined as the point where the error

$$E_i^{(P)}=\sum_{q=1}^{P}(y_i^{(q)}-f_i(\mathbf{x}^{(q)},\mathbf{w}_i))^2 \tag{20}$$

has its minimum. Due to the linearity in the parameter, the minimum is a global minimum and can be easily found by using the Moore-Penrose inverse. At this point the gradient is equal to zero, i.e.,

$$\nabla E_i^{(P)}|_{\mathbf{w}_{P_i}}=0\ . \tag{21}$$

The Taylor expansion is then given by

$$E_i^{(P)}=E_i^{(P)}|_{\mathbf{w}_{P_i}}+\tfrac{1}{2}(\mathbf{w}_i-\mathbf{w}_{P_i})^T\nabla\nabla E_i^{(P)}|_{\mathbf{w}_{P_i}}(\mathbf{w}_i-\mathbf{w}_{P_i})\ , \tag{22}$$

which is exact. Replacing Eq. (22) in Eq. (19) the integral adopt then the Gaussian form and therefore can be easily calculated yielding

$$Z(P)=e^{-\beta E^{(P)}|_{\mathbf{w}_{P_i}}}(2\pi)^{MD/2}(\det(2\beta\Theta^{(P)}))^{-M/2}\ , \tag{23}$$

where $D=\dim(\mathbf{w}_i)=1+N+N^2+\cdots+N^R$ and the matrix $\Theta^{(P)}$ is defined by

$$\Theta^{(P)}=\tfrac{1}{2}\nabla\nabla E_j^{(P)}|_{\mathbf{w}_{P_i}}=\sum_{i=1}^{P}\mathbf{t}^{(i)}(\mathbf{t}^{(i)})^T \tag{24}$$

with

$$E^{(P)}=\sum_{i=1}^{M}E_i^{(P)}\ . \tag{25}$$

We need to calculate also $Z(P+1)$. To do that we use the Taylor expansion of $E_i^{(P+1)}$ around the point $\mathbf{w}_{P_i}$, which is defined above. We obtain

$$E_i^{(P+1)}=E_i^{(P+1)}|_{\mathbf{w}_{P_i}}+(\mathbf{w}_i-\mathbf{w}_{P_i})^T\nabla E_i|_{\mathbf{w}_{P_i}}+\tfrac{1}{2}(\mathbf{w}_i-\mathbf{w}_{P_i})^T\nabla\nabla E_i^{(P+1)}|_{\mathbf{w}_{P_i}}(\mathbf{w}_i-\mathbf{w}_{P_i}) \tag{26}$$

and after reordering

$$E_i^{(P+1)}=E_i^{(P+1)}|_{\mathbf{w}_{P_i}}+\tfrac{1}{2}(\mathbf{w}_i-\mathbf{w}_{P_i}+\mathbf{b}_i)^T\nabla\nabla E_i^{(P+1)}|_{\mathbf{w}_{P_i}}\times(\mathbf{w}_i-\mathbf{w}_{P_i}+\mathbf{b}_i)-\tfrac{1}{2}(\nabla E_i|_{\mathbf{w}_{P_i}})^T(\nabla\nabla E_i|_{\mathbf{w}_{P_i}})^{-T}(\nabla E_i|_{\mathbf{w}_{P_i}})\ , \tag{27}$$

where the error at the new point is

$$E_i=(y_i-f_i(\mathbf{x},\mathbf{w}_{P_i})) \tag{28}$$

and

$$\mathbf{b}_i=-(\nabla\nabla E_i|_{\mathbf{w}_{P_i}})^{-1}\nabla E_i|_{\mathbf{w}_{P_i}}\ . \tag{29}$$

After integration of a Gaussian function we obtain,

$$Z(P+1)=e^{-\beta E^{(P+1)}|_{\mathbf{w}_P}}(2\pi)^{MD/2}(\det(2\beta\Theta^{(P+1)}))^{-M/2}\times\left[e^{\beta E|_{\mathbf{w}_P}(t^T(\Theta^{(P+1)})^{-T}t)}\right]\ . \tag{30}$$

Using Eqs. (23), (30), and (13), it is possible to write the likelihood of a new data as

$$p(\mathbf{y}/\mathbf{x},D^{(P)})=\frac{1}{z}e^{-\beta\frac{E}{\det[I+\theta(\Theta^{(P)})^{-1}]}-\frac{M}{2}\ln\{\det[I+\theta(\Theta^{(P)})^{-1}]\}} \tag{31}$$

where

$$E=\sum_{i=1}^{M}(y_i-f_i(\mathbf{x},\mathbf{w}_{P_i}))^2 \tag{32}$$

is the error of the new point and

$$\theta=\mathbf{t}\mathbf{t}^T \tag{33}$$

with the vector $\mathbf{t}$ defined for the new point. In Eq. (31), we have used the fact that

$$\det[I+\theta(\Theta^{(P)})^{-1}]=\frac{1}{1-\mathbf{t}^T(\Theta^{(P+1)})^{-T}\mathbf{t}}\ , \tag{34}$$

which can be easily derived by using the relation of Fedorov [7]

$$\det(A+a\mathbf{c}\mathbf{c}^T)=\det(A)(1+a\mathbf{c}^TA^{-1}\mathbf{c})\ . \tag{35}$$

The unity matrix was noted by I.

Equation (31) defines the prediction probability of a new point given P examples. Using Eq. (31), we can calculate the informativeness of a new data analytically:

$$K(p(\mathbf{y}/\mathbf{x}^{(i)},D^{(P)}),p(\mathbf{y}/\mathbf{x}^{(i)},D^{(P+1)}))=\tfrac{1}{2}\ln\left[\frac{\det[I+\theta(\Theta^{(P+1)})^{-1}]}{\det[I+\theta(\Theta^{(P)})^{-1}]}\right]+\beta\frac{\|\mathbf{f}(\mathbf{x}^{(i)},\mathbf{w}_P)-\mathbf{f}(\mathbf{x}^{(i)},\mathbf{w}_{(P+1)})\|^2}{\det[I+\theta(\Theta^{(P+1)})^{-1}]}$$
$$+\tfrac{1}{2}\left[\frac{\det[I+\theta(\Theta^{(P)})^{-1}]}{\det[I+\theta(\Theta^{(P+1)})^{-1}]}-1\right]\ , \tag{36}$$

with θ defined for the new pattern i. Inserting the last equation in the definition of the novelty $N(P)$ of Eq. (14), we obtain an analytical measure of the informativeness of the new data. The novelty in the parameter space defined in Eq. (16) can be also analytically calculated yielding

$$NW(P)=\frac{M}{2}\ln[\det(I+\mathbf{t}^T(\Theta^{(P)})^{-1}\mathbf{t})]\ . \tag{37}$$

On the other hand, it follows from Eq. (31) that the maximum of the likelihood corresponds to the zeros error $E=0$. Hence, the output of the network is equal to

$$\mathbf{f}(\mathbf{x}^{(i)},\mathbf{w}_P) \tag{38}$$

and

$$p(\mathbf{y}/\mathbf{x},D^{(P)})=\frac{1}{z}e^{-(M/2)\ln[\det(I+\theta(\Theta^{(P)})^{-1})]}\ . \tag{39}$$

The quantity in the exponent can be rewritten by using the Fedorov equality (see Ref. [5]) as

$$\tfrac{1}{2}\ln[\det(I+\theta(\Theta^{(P)})^{-1})]$$
$$=\tfrac{1}{2}\ln[\det(I+\mathbf{t}^T(\Theta^{(P)})^{-1}\mathbf{t})]\ . \tag{40}$$

In fact, the exponent of Eq. (39) is a measure of the range of confidence for the new point and it is identical with the novelty $NW(P)$ measured in the weight space. So, we have derived the latter novelty in a different way. This deviation explicitly implies that the novelty in the parameter space is a good measure only if $E=0$. On the other hand, the regions where the novelty is important are the regions where E is normally big. Therefore, we conclude that a good measure of novelty should include both terms like those defined in Eqs. (14) and (15) and calculated for the polynomial case in Eq. (36).

V. NUMERICAL RESULTS

The informativeness of the new data is shown here in the following example. Originally, a polynomial of the order 6 is used to obtain 12 data points unevenly distributed in the region $[-1,0.8]$, which are compromised

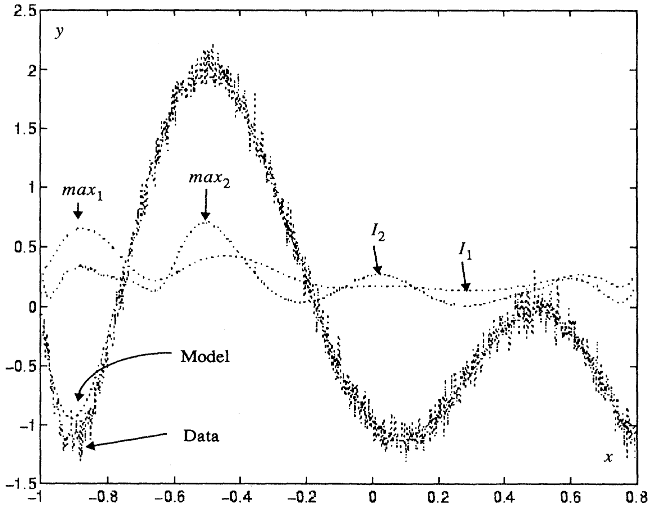

FIG. 1. New data informativeness measure. The figure depicts the training data pairs (x,y), the obtained sixth-order polynomial model, as well as the two measures of innovativeness. I_1, informativeness measure according to Eq. (36); I_2, informativeness measure according to Eq. (40).

with the additive Gaussian noise. The model is assumed to be a polynomial of the same order. Due to the small number of training data and the presence of noise, the obtained model parameters will differ from the real ones in spite of the correct polynomial order. Hence, an additional input-output data pair could have strong influence on the model accuracy, especially if it is in the region not covered by the original 12 points.

The informativeness measure introduced in this paper is calculated for the 900 points within the given interval $[-1,0.8]$ with $T=10$ and it is depicted in Fig. 1 as the curve I_1. For comparison, the measure of informativeness in the parameter space only is depicted as I_2. In addition, the same figure shows the noisy measurement of the original system whose informativeness was measured as well as the optimal model fit at the same points.

It is visible that both measures are peaked in the regions where the modeling error is large. Nevertheless, the positions of their maximum differ. This is an essential difference since the point with maximum informativeness would be selected and added to the training set in the experiment design process.

VI. CONCLUSIONS

In this paper, an information-theoretic based objective function that measures the informativeness of a sequential data in the framework of the statistical mechanical formulation of learning and generalization was formulated. This theory is especially suitable for the implementation of active data selection, known also as query learning. The aim of the latter is to improve the generalization ability of a network that continuously learns by actively selecting optimal nonredundant data, i.e., data that content new information for the model. In our model, only the data that carry new information are selected for learning. The novel informativeness measure of the incoming new data is defined as the Kullback-Leibler entropy between the probability of the output given the input and the past trained data (likelihood), and the same probability but including the new data in the trained data. We apply the developed theory for polynomial models, i.e., higher-order neural networks, due to the fact that in this case all integrals can be performed analytically without approximation.

[1] P. Sollich, Phys. Rev. E **49**, 4637 (1994).

[2] D. MacKay, Neural Comput. **4**, 590 (1992).

[3] E. Baum, IEEE Trans. Neural. Networks **2**, 5 (1991).

[4] J. N. Hwang, J. J. Choi, S. Oh, and R. Marks, Jr., IEEE Trans. Neural Networks **2**, 131 (1991).

[5] W. Kinzel and P. Rujan, Europhys. Lett. **13**, 473 (1990).

[6] 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.

[7] V. V. Fedorov, *Theory of Optimal Experiments* (Academic, New York, 1972).

[8] S. D. Silvey, *Optimal Design* (Chapman and Hall, London, 1980).

[9] P. Chaudhuri and P. A. Mykland, J. Am. Stat. Assoc. **88**, 538 (1993).

[10] J. Pilz, *Bayesian Estimation and Experimental Design in Linear Regression Models,* 2nd ed. (John Wiley, Chichester, 1991).

[11] E. Levin, N. Tishby, and S. Solla, Proc. IEEE **78**, 1568 (1990).

[12] N. Tishby, E. Levin, and S. Solla, in Proceedings of the International Joint Conference on Neural Networks (IEEE, Washington, DC, 1989), Vol. 2, p. 403.

[13] N. Tishby, in *SFI Studies in the Sciences of Complexity,* edited by D. Wolpert (Addison-Wesley, Reading, PA 1995), Vol. xx, p. 215.

[14] N. Tishby, in *From Statistical Physics to Statistical Inference and Back,* Vol. 428, *NATO Advanced Study Institute Series, Series C: Mathematical and Physical Sciences,* edited by P. Glassberger and J. P. Nadal (Plenum, New York, 1995), p. 205.

[15] R. Meir and F. Fontanari, Physica A **200**, 644 (1993).

[16] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackel, and J. Hopfield, Complex Syst. **1**, 877 (1987).

[17] D. MacKay, Neural Comput. **4**, 415 (1991).

[18] D. MacKay, Neural Comput. **4**, 448 (1992).