Generalization in a two-layer neural network with multiple outputs

Kukjin Kang and Jong-Hoon Oh
Department of Physics, Pohang Institute of Science and Technology, Hyoja San 31, Pohang, Kyongbuk 790-784, Korea

Chulan Kwon and Youngah Park
Department of Physics, Myong Ji University, Yongin, Kyonggi 449-728, Korea

(Received 2 October 1995; revised manuscript received 22 March 1996)

We study generalization in a fully connected two-layer neural network with multiple output nodes. Similar to the learning of fully connected committee machine, the learning is characterized by a discontinuous phase transition between the permutation symmetric phase and the permutation symmetry breaking phase. We find that the learning curve in the permutation symmetric phase is universal, irrespective of the number of output nodes. The first-order phase transition point, i.e., the critical number of examples required for perfect learning, is inversely proportional to the number of outputs. The replica calculation shows good agreement with Monte Carlo simulation. [S1063-651X(96)11308-8]

PACS number(s): 87.10+e, 05.50+q, 64.60.Cn

Learning from examples in layered neural networks has been a common interest of statistical mechanics and other related areas such as computer science and mathematical statistics for the last few years [1]. Following the statistical mechanics formulation of Gardner [2,3], there have been many efforts to study learning from examples in feed-forward neural networks such as the perceptron [4–8]. Whereas most of the mathematical approaches gave general asymptotic bounds [9,10], statistical mechanics was able to predict a precise learning curve for a specific model. Specially, an interesting first-order phase transition was found in the single-layer perceptron with binary weights. This first-order phase transition was interpreted as a sudden learning of the perceptron related to the discrete phase space structure [5,6,8]. Recently, there has been also some progress in the study of the generalization in multilayered neural networks from a statistical physics perspective. Special attention was paid to a two-layer network called a committee machine, and a discontinuous phase transition originating from a different mechanism was found [11–15].

Most of the studies of generalization have concerned learning of a network with a single output node. In particular, theoretical studies concentrate on learning a dichotomy rule. On the other hand, neural networks used for real world applications, such as classification tasks, usually need multiple output nodes. Understanding the effect of multiple output in a multilayer perceptron would be a meaningful step, which extends the relevance of the theory from toy models to more realistic neural networks. It can reduce the gap between theories and practice.

Here, we present a study of generalization in a fully connected two-layer perceptron with multiple output nodes. Consider a two-layer neural network with N input nodes, M hidden nodes, and K output nodes. In the fully connected architecture, each input node is connected to all the hidden nodes and each hidden node is connected to all the output nodes. In this study, we consider binary weights and take the limit where $N \gg M \gg 1$.

In our previous work [14], we studied generalization in a fully connected committee machine with binary weights, which corresponds to the case $K=1$ in this work. When the number of examples is of the order of N, the system is in the permutation symmetric (PS) phase. In the PS phase, the generalization error decreases more rapidly than expected from the asymptotic behavior of the upper bounds predicted by the Vapnik-Chervonenkis approach and others [9,10]. As the number of examples, P, grows to the order of MN the generalization error converges to a constant value. When the number of examples reaches a critical value, the system undergoes a first-order phase transition driven by permutation symmetry breaking (PSB). Above the transition point, the system immediately falls into a state of perfect learning.

The motivation of this work is to study whether this picture is also relevant in a network with multiple output nodes. The discontinuous learning curve is also observed in the neural network with multiple output nodes. Interestingly the learning curve in the permutation symmetric phase is the same irrespective of the number of output nodes. However, the first-order transition occurs for a smaller number of examples, inversely proportional to the number of outputs. These results are obtained from the replica calculation and the Monte Carlo simulation.

We consider a student network learning a realizable rule from the examples provided by a teacher with the same architecture. The network maps input vectors, $\boldsymbol{S}^l=(S_i^l,\ldots,S_N^l)$ to output vectors, $\boldsymbol{\sigma}=(\sigma_1,\ldots,\sigma_K)$ as

$$\sigma_k(\boldsymbol{W};\boldsymbol{S}^l)=g_2\left[M^{-\frac{1}{2}}\sum_j^M W_{kj}^{(2)} g_1\left(N^{-\frac{1}{2}}\sum_i^N W_{ji}^{(1)} S_i^l\right)\right]. \quad (1)$$

Here, $\boldsymbol{W}=\{W_{ji}^{(1)},W_{kj}^{(2)}\}$ is a set of synaptic weights whose element $W_{ji}^{(1)}$ denotes the first-layer weight between the ith input node and the jth hidden node, and $W_{kj}^{(2)}$ denotes the second-layer weight between the jth hidden node and the kth output node. The transfer functions of the hidden nodes and the output nodes are $g_1(x)$ and $g_2(x)$, respectively. In this paper, we will consider the case $g_1(x)=g_2(x)=\mathrm{sgn}(x)$. The weights of the teacher are given by $\boldsymbol{W}^0=\{W_{ji}^{0(1)},W_{kj}^{0(2)}\}$.

The energy of the system is defined as a sum of errors over output nodes and examples:

$$E=\sum_{l=1}^{P} \epsilon(\boldsymbol{W};\boldsymbol{S}^l), \tag{2}$$

$$\epsilon(\boldsymbol{W};\boldsymbol{S}^l)=\sum_{k=1}^{K} \Theta(-\sigma_k(\boldsymbol{W}^0;\boldsymbol{S}^l)\sigma_k(\boldsymbol{W};\boldsymbol{S}^l)), \tag{3}$$

where $\Theta(x)$ is the Heaviside step function. The training procedure is assumed to be a stochastic process that leads to a Gibbs distribution of the weights after a long time. The equilibrium probability distribution of weights is given by

$$\mathcal{P}(\boldsymbol{W})=Z^{-1}\exp[-\beta E(\boldsymbol{W})], \tag{4}$$

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

$$Z=\int d\mu\ (\boldsymbol{W})\exp[-\beta E(\boldsymbol{W})]. \tag{5}$$

We use the Monte Carlo method to simulate this training algorithm.

The performance of the network is measured by the generalization function $\boldsymbol{\epsilon}(\boldsymbol{W})=(1/K)\int d\boldsymbol{S}\epsilon(\boldsymbol{W};\boldsymbol{S})$, where $\int d\boldsymbol{S}$ represents an average over the whole space of inputs. The generalization error $\boldsymbol{\epsilon}_g$ is defined by $\boldsymbol{\epsilon}_g=\langle\langle\langle\epsilon(\boldsymbol{W})\rangle_T\rangle\rangle$ where $\langle\langle\ \rangle\rangle$ denotes the quenched average over the examples and $\langle\ \rangle_T$ denotes the thermal average over the distribution of Eq. (4).

The replica partition function can be written as

$$\langle\langle Z^n\rangle\rangle=\mathrm{Tr}_{\boldsymbol{W}}\exp[P\mathcal{G}_r], \tag{6}$$

where

$$\mathcal{G}_r=-\ln\int d\mu\ (\boldsymbol{S})\exp\left(-\beta\sum_{\sigma=1}^{n} \epsilon(\boldsymbol{W}^\sigma;\boldsymbol{S})\right). \tag{7}$$

Assuming a large M, we obtain $\mathcal{G}_r$ up to the zeroth order in $1/M$,

$$\begin{aligned}\exp[\mathcal{G}_r]=&\int\prod_{\sigma,k}\frac{du_k^\sigma d\hat{u}_k^\sigma}{2\pi}\int\prod_k\frac{dv_k d\hat{v}_k}{2\pi}\exp\left[-\frac{\beta}{2}\sum_{\sigma,k}\{g_2(u_k^\sigma)-g_2(v_k)\}^2+i\sum_{\sigma,k}u_k^\sigma\hat{u}_k^\sigma+i\sum_k v_k\hat{v}_k-\frac{1}{2}\sum_k\hat{v}_k^2\right]\\&\times\exp\left[-\frac{1}{2}\sum_{k,k'}\sum_{\sigma,\rho}\hat{u}_k^\sigma\hat{u}_{k'}^\rho\frac{1}{M}\sum_{jj'}W_{kj}^{\sigma(2)}W_{k'j'}^{\rho(2)}\frac{2}{\pi}\sin^{-1}\left(\frac{1}{N}\sum_i W_{ji}^{\sigma(1)}W_{j'i}^{\rho(1)}\right)\right.\\&\left.-\sum_{k,k'}\sum_\sigma\hat{u}_k^\sigma\hat{v}_{k'}\frac{1}{M}\sum_{jj'}W_{kj}^{\sigma(2)}W_{k',j'}^{0(2)}\frac{2}{\pi}\sin^{-1}\left(\frac{1}{N}\sum_i W_{ji}^{\sigma(1)}W_{j'i}^{0(1)}\right)\right],\end{aligned} \tag{8}$$

where σ,ρ are replica indices. We assume that the weights of the teacher are uncorrelated such that $(1/N)\Sigma_i W_{ji}^0 W_{j'i}^0=\delta_{jj'}$.

Analyzing the energy function, we find two symmetries in this network, i.e., the gauge symmetry and the permutation symmetry. Using these symmetries, we can rearrange the configuration of the weights so that many degenerate configurations can be described by a single representation. In the following, we will describe the rearrangement scheme.

The gauge symmetry comes from the fact that the transfer function is odd. Keeping the outputs unchanged, we can simultaneously flip the sign of the second layer weight connected to a particular hidden node, and the sign of every first-layer weight connected to the same hidden node. Let us choose a reference output node, say $k=1$. Then we perform a transformation:

$$\begin{aligned}&W_{kj}^{\sigma(2)}W_{1j}^{\sigma(2)}\rightarrow W_{kj}^{\sigma(2)},\\&W_{ji}^{\sigma(1)}W_{1j}^{\sigma(2)}\rightarrow W_{ji}^{\sigma(1)}.\end{aligned} \tag{9}$$

Under this gauge transformation, a network with binary weights and a single output node leads to the committee machine.

The permutation symmetry relates to permutation of the hidden nodes. A permutation does not change the output of the network. The next step of rearrangement uses the permutation symmetry. As an example, consider a network with two output nodes. When M is large enough, roughly half the weights $W_{2j}^{(2)}$ are $+1$ and the rest are -1. Then, we rearrange the order of the hidden nodes using the permutation symmetry. We divide the weights into two groups, one with positive sign and the other with negative sign. Now we have the systematically arranged configuration:

$$\begin{aligned}&\boldsymbol{W}_1^{\sigma(2)}=(\ldots,W_{1j}^{(2)},\ldots)=(1,\ldots,1,1,\ldots,1),\\&\boldsymbol{W}_2^{\sigma(2)}=(\ldots,W_{2j}^{(2)},\ldots)=(1,\ldots,1,-1,\ldots,-1).\end{aligned} \tag{10}$$

The first line is the result of the gauge transformation. If we have a third output we can repeat a similar permutation in each block of $W_{2j}^{(2)}$ and the weights are arranged as

$$\boldsymbol{W}_3^{\sigma(2)}=(1,\ldots,1,-1,\ldots,-1,1,\ldots,1,-1,\ldots,-1). \tag{11}$$

In this way, the hidden nodes and the second-layer weights decompose into blocks. We can generalize this procedure for

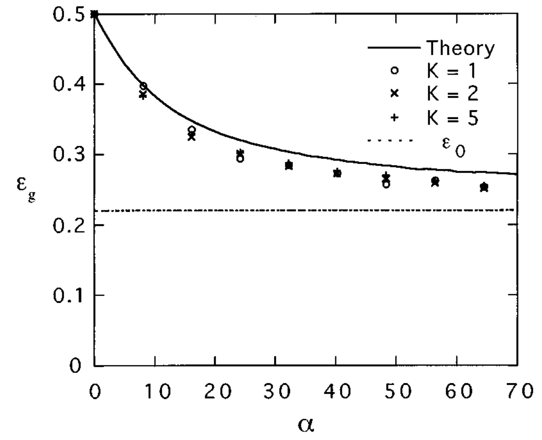

FIG. 1. The generalization curve of two-layer networks when $P\sim O(N)$. $M=N=31$ and $K=1,2$, 5, respectively, and $T=2.5$. The solid line is the analytic plot obtained by the replica trick and the horizontal line denotes $\epsilon_0=\lim_{\alpha\to\infty}\epsilon_g$.

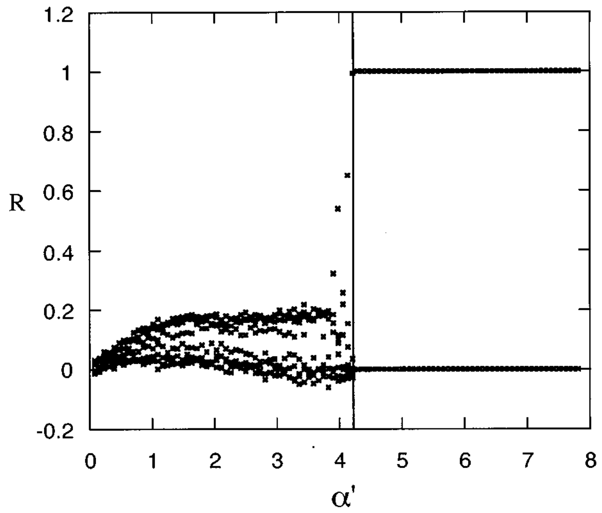

FIG. 2. Snapshot of the matrix elements $R_{jj'}$ for a network $N=16$, $M=8$, $K=2$, and temperature $T=2.5$. The vertical line denotes the theoretical phase transition point $\alpha'_c(K=2)\simeq 8.46/2$.

K output nodes, as long as the number of hidden nodes in each block is large enough, i.e., $M_K=M/2^{K-1}\gg 1$.

The phase transition by permutation symmetry breaking plays an essential role in learning of the committee machine. The permutation of the hidden nodes of a given teacher yields many different teachers with the same input-output relations. Many alternative teachers also can be realized by permuting the hidden nodes.

Let us consider the energy surface in the phase space of $\{\boldsymbol{W}\}$. Each teacher is at a minimum of the energy surface. For a small P, all the teachers belong to a single thermally connected region in the phase space. A student does not know from which teacher to learn, and the student is roughly equidistant from all the transformed teachers. We will call this the PS phase of the network. As P becomes of order MN, many thermally disconnected valleys appear around the permuted teachers. This phase is called the PSB phase.

Now the order parameters,

$$Q^{\sigma\rho}_{jj'}=\left\langle\left\langle\left\langle\frac{1}{N}\sum_i W^{\sigma(1)}_{ji}W^{\rho(1)}_{j'i}\right\rangle_T\right\rangle\right\rangle \tag{12}$$

and

$$R^{\sigma}_{jj'}=\left\langle\left\langle\left\langle\frac{1}{N}\sum_i W^{\sigma(1)}_{ji}W^{0(1)}_{j'i}\right\rangle_T\right\rangle\right\rangle, \tag{13}$$

are defined in the rearranged representation. We assume a replica symmetric ansatz:

$$Q^{\sigma\rho}_{jj'}=\begin{cases}(1-\delta_{jj'})C_{jj'}+\delta_{jj'}q, & \sigma\neq\rho,\\ (1-\delta_{jj'})Q_{jj'}+\delta_{jj'}, & \sigma=\rho,\end{cases}$$

$$R^{\sigma}_{jj'}=(1-\delta_{jj'})R_{jj'}+\delta_{jj'}r.$$

Here, the diagonal order parameters q and r measure the preference for a particular implementation of the teacher. The off diagonal matrices $\boldsymbol{C}(C_{jj'})$, $\boldsymbol{Q}(Q_{jj'})$, and $\boldsymbol{R}(R_{jj'})$ represent the order parameters between different hidden nodes. In the PS phase, the student cannot recognize a particular teacher. The role of each hidden node is not specialized, so the off-diagonal order parameters play an important role. The diagonal order parameters dominate in the PSB phase, where the student is similar to a particular teacher. Each hidden node specializes its role.

We expect that the matrices $\boldsymbol{C}$, $\boldsymbol{Q}$, and $\boldsymbol{R}$ form block matrices. The dimension of the blocks is assumed to be $M_K\times M_K$. It comes from the fact that permutation is allowed among the hidden nodes in each block of size M_K. Each block has a constant matrix element in the limit $M_K\to\infty$. Typical forms of the matrices for two outputs and three outputs are shown below:

a	b
b	a

a	b	c	d
b	a	d	c
c	d	a	b
d	c	b	a

As for the committee machine, the overlap order parameters $\boldsymbol{C},\boldsymbol{Q},\boldsymbol{R}$ are of order $1/M$. We can expand $\sin^{-1}(\,)$ for $j\neq j'$ in Eq. (8). To leading order in $1/M$ the free energy is expressed in terms of

$$\begin{aligned}&\frac{1}{M}\boldsymbol{W}^{\sigma(2)}_k\boldsymbol{C}\boldsymbol{W}^{\rho(2)}_{k'}=\delta_{kk'}\lambda_{Ck},\\ &\frac{1}{M}\boldsymbol{W}^{\sigma(2)}_k\boldsymbol{Q}\boldsymbol{W}^{\sigma(2)}_{k'}=\delta_{kk'}\lambda_{Qk},\\ &\frac{1}{M}\boldsymbol{W}^{\sigma(2)}_k\boldsymbol{R}\boldsymbol{W}^{0(2)}_{k'}=\delta_{kk'}\lambda_{Rk},\end{aligned} \tag{14}$$

where $\lambda_{Ck},\lambda_{Qk}$, and λ_{Rk} are the kth eigenvalues of $\boldsymbol{C}$ $\boldsymbol{Q}$, and $\boldsymbol{R}$, respectively. The rearranged second-layer weights are eigenvectors of the order parameter matrices. Therefore only the case with $k=k'$ contributes. Now the free energy is a function of the eigenvalues rather than the matrix elements themselves. Note that the eigenvalues are of order 1 while the matrix elements are of order $1/M$.

Now we take the $n\rightarrow 0$ limit and find q,r and the eigenvalues $\lambda_{Ck},\lambda_{Qk}$, and λ_{Rk} by the saddle point approximation in the thermodynamic limit. As in Ref. [14], we consider two different regimes, where P is $O(N)$ and $O(MN)$, respectively.

(i) $P\sim O(N)$. Here only the PS phase exists and $r=q=0$. The free energy of the system can be written as:

$$-\beta F=KN(G_0+\alpha G_r), \tag{15}$$

$$G_0=-\frac{1}{2}\lambda_Q+\frac{1}{2}\frac{\lambda_C-\lambda_R^2}{1+\lambda_Q-\lambda_C}+\frac{1}{2}\ln(1+\lambda_Q-\lambda_C), \tag{16}$$

$$G_r=2\alpha\int_{-\infty}^{\infty}Dx\ H(ax)\ln[e^{-\beta}+(1-e^{-\beta})H(bx)], \tag{17}$$

with

$$a=\frac{(2/\pi)\lambda_R}{\sqrt{(2/\pi)\lambda_C-[(2/\pi)\lambda_R]^2}},$$

$$b=\sqrt{\frac{(2/\pi)\lambda_C}{1+(2/\pi)(\lambda_Q-\lambda_C)}}, \tag{18}$$

where $\int Dx=\int dx\ e^{-x^2/2}$ and $H(u)=\int_u^\infty Dx$. The generalization error is given by:

$$\epsilon_g=\frac{1}{\pi}\cos^{-1}\left(\frac{(2/\pi)\lambda_R}{\sqrt{1+(2/\pi)\lambda_Q}}\right). \tag{19}$$

Note that the k dependence of the $\lambda_{Ck},\lambda_{Qk}$, and λ_{Rk} is removed since the saddle point equation is the same for all k. Minimizing the free energy with respect to λ_C,λ_Q, and λ_R, we find the generalization error. The free energy has the same form n as for the fully connected committee machine except for the multiplicative factor K. For $K=1$, the eigenvalues $\lambda_C,\lambda_Q,\lambda_R$ reduce to the corresponding matrix elements for the committee machine [14,15]. It explains the surprising result that the learning curve is the same as that of the committee machine irrespective of the number of output nodes.

This interesting result is confirmed by the Monte Carlo simulation. Figure 1 shows the learning curve from the numerical simulation along with the analytic result obtained by the replica calculation. In this simulation we use networks with different numbers of outputs, $K=1,\ 2$, and 5, respectively. The number of input and hidden nodes is the same, i.e., $N=M=31$. When we plot $\epsilon_g(\alpha)$, all the learning curves with different numbers of outputs collapse to a single curve that also agrees well with the replica calculation.

(ii) $P\sim O(MN)$. We introduce a scaling for the free energy, and the free energy is written as

$$-\beta F=MN(G_0+K\alpha' G_r), \tag{20}$$

where $\alpha'=P/MN$,

$$G_0=-\frac{1}{2}(1-q)\hat{q}-R\hat{R}+\int Dz\ln\cosh(\sqrt{\hat{q}}z+\hat{r}), \tag{21}$$

$$G_r=2\alpha\int_{-\infty}^{\infty}Dx\ H(ax)\ln[e^{-\beta}+(1-e^{-\beta})H(bx)], \tag{22}$$

with

$$a=\frac{(2/\pi)(\sin^{-1}r+\lambda_R)}{[(2/\pi)[\sin^{-1}q-q+(r+\lambda_R)^2]-(2/\pi)^2(\sin^{-1}r+\lambda_R)^2]^{1/2}},$$

$$b=\sqrt{(2/\pi)[\sin^{-1}q-q+(r+\lambda_R)^2]}1-2/\pi-(2/\pi)(\sin^{-1}q-q). \tag{23}$$

The generalization error is given by

$$\epsilon_g=\frac{1}{\pi}\cos^{-1}\left[\frac{(2/\pi)(\sin^{-1}r+\lambda_R)}{\sqrt{1-(2/\pi)+(2/\pi)(r+\lambda_R)^2}}\right]. \tag{24}$$

In the above expressions, λ_Q and λ_C are eliminated by the saddle point equation,

$$1+\lambda_Q=q+\lambda_C=(r+\lambda_R)^2. \tag{25}$$

There are two solutions that minimize the free energy. One is the PS solution where $q=r=0$, and $\lambda_C,\lambda_Q,\lambda_R$ are nonzero. We find that this solution describes the limit $\alpha\rightarrow\infty$ for $P\sim O(N)$. The generalization error is also a nonzero constant, which coincides with the asymptotic value of ϵ_g in the limit $\alpha\rightarrow\infty$ for $P\sim O(N)$. Increasing the number of examples, the system reaches the PSB phase by a first-order phase transition. The PSB solution is given by $q=r=1$ and $\lambda_C=\lambda_Q=\lambda_R=0$. This means that the student becomes an exact copy of one of many equivalent teachers made by the transformations explained above. The generalization error vanishes in the PSB phase. When we compare the expression for the free energy with that of the committee machine, we find that α' is replaced by $K\alpha'$. The new transition point α_c' therefore scales with the number of the outputs as

$$\alpha_c'(K)=\frac{1}{K}\alpha_c'(K=1). \tag{26}$$

Observing the behavior of the order parameter matrices in the simulation is a good way to check the phase transition. Figure 2 show a snapshot of the matrix elements $R_{jj'}$ measured from the simulation of the network with two output nodes. The theory predicts that the matrix elements should split into two different values $R_{jj'} \sim O(1/M)$ and $R_{jj'}=0$ in the PS phase, and $R_{jj'}=1$ and 0 in the PSB phase. The observed values of the matrix element $R_{jj'}$ show the expected picture. The theoretical phase transition point $\alpha_c'(K=2)=1/2\alpha_c'(K=1)$ shown by the vertical line also agrees with the simulation.

Our study can be extended to other situations as was possible in the study of the committee machine, for example, to the case of continuous weights in the input layer as in Ref. [15], and to the case of the sigmoid transfer function as in Ref. [14]. We expect that the learning curve in the PS phase will be the same for different K and the scaling of the phase transition point also will be described by Eq. (26) in these cases. The asymptotic behavior in the PSB phase may be different. It is now tempting to ask whether the symmetry breaking and the phase transition are relevant to the back propagation learning of the multilayer perceptron. With fully continuous weights the situation is more complicated, so analytic calculation based on the replica method may not be feasible. Existence of the first-order transition is also questionable. However, we believe that the symmetry breaking still plays an important role in characterizing the nature of the learning curve. We note that recent large scale simulations of learning curves show qualitatively similar behavior to that shown in this work [16,17].

J.H.O. and C.K. was supported in part by KOSEF through Grant No. 941-0200-030-2 and the Korea Ministry of Education through the POSTECH Basic Science Research Institute. C. K. and Y. P. were also supported by the Center for Theoretical Physics of Seoul National University and the nondirected fund from the Korea Research Foundation, 1993.

[1] For reviews, see M. Opper and W. Kinzel, in Physics of Neural Networks, edited by J. L. Van Hemmen, E. Domany, and K. Schulten (Springer-Verlag, Berlin, in press).

[2] E. Gardner, Europhys. Lett. **4**, 481 (1987); J. Phys. A **21**, 257 (1988).

[3] E. Gardner and B. Derrida, J. Phys. A **22**, 1983 (1989).

[4] H. Sompolinsky, N. Tishby, and H. S. Seung, Phys. Rev. Lett. **65**, 1683 (1990).

[5] H. S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).

[6] G. Györgyi, Phys. Rev. Lett. **64**, 2957 (1990); Phys. Rev. A **41**, 7097 (1990).

[7] C. Kwon, Y. Park, and J.-H. Oh, Phys. Rev. E **47**, 3707 (1993).

[8] S. Ha, K. Kang, J.-H. Oh, C. Kwon, and Y. Park, in *Proceeding of 1993 International Joint Conference on Neural Network* (IEEE, Nagoya, 1993), Vol. 2, pp. 1723–1726.

[9] E. Baum and D. Haussler, Neural Comput. **1**, 151 (1989).

[10] S. Amari, Neural Networks **6**, 161 (1993).

[11] E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992).

[12] A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius, Phys. Rev. A **45**, 7590 (1992).

[13] G. Mato and N. Parga, J. Phys. A **25**, 5047 (1992).

[14] K. Kang, J.-H. Oh, C. Kwon, and Y. Park, Phys. Rev. E **48**, 4805 (1993).

[15] H. Schwarze and J. Hertz, J. Phys. A **26**, 4919 (1993).

[16] C. Lee and J.-H. Oh (unpublished).

[17] K. Mueller, M. Finke, N. Murata, K. Schulten, and S. Amari, in *Neural Networks: The Statistical Mechanics Perspective*, edited by J.-H. Oh, C. Kwon, and S. Cho (World Scientific, Singapore, 1995), p. 73.