PHYSICAL REVIEW E VOLUME 48, NUMBER 6 DECEMBER 1993

Generalization in a two-layer neural network

Kukjin Kang and Jong-Hoon Oh
Department of Physics, Pohang Institute of Science and Technology, Pohang, Kyongbuk, Korea

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

(Received 12 July 1993)

Generalization in a fully connected two-layer neural network with N input nodes, M hidden nodes, a single output node, and binary weights is studied in the annealed approximation. When the number of examples is the order of N, the generalization error approaches a plateau and the system is in a permutation symmetric phase. When the number of examples is of the order of MN, the system undergoes a first-order phase transition to perfect generalization and the permutation symmetry breaks. Results of the computer simulation show good agreement with analytic calculation.

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

Following the pioneering works of Gardner [1,2], there have been many studies [3–6] to understand the properties of feed-forward neural networks using statistical mechanics. Gardner studied the storage capacity of the single-layer perceptron [7] and there have been several related works [1–3]. Another interesting topic would be learning from examples. The estimation of an appropriate size of training example set for a valid generalization is an important issue in this problem. This problem was extensively studied for the single-layer perceptron [4–6,8–10]. However, it is well known that the perceptron architecture can solve only a linear threshold problem. Meanwhile, many of the real-world problems are approached using the networks with hidden layers [11]. In the presence of hidden layers, the learning mechanism is much more complicated and not much is known about the generalization of multilayer networks. Treelike architecture where an input node is connected to only one hidden node was studied by several groups [12–14]. Recently there have been some efforts to get the storage capacity of fully connected two-layer networks [15,16].

Here we study the generalization of a fully connected two-layer network which is believed to perform a fairly complex task. Consider a two-layer feed-forward network with N input nodes, M hidden nodes, and a single output node. Every input node is connected to all of the hidden nodes by the binary weights. Specifically, all the weights in the second layer are set to unity. This architecture is usually called a committee machine. We can always map any network with binary weights into a committee machine by changing all the negative weights in the second layer to $+1$ and at the same time flipping the signs of the all the weights in the first layer connected to those negative weights. When the transfer function is an odd function, this new network performs exactly the same function as the original one. We calculate the generalization curve for this machine using the annealed approximation. We also perform Monte Carlo simulations, showing good agreement with the analytic calculations.

The network maps an input vector $\boldsymbol{S}^l = \{S_i^l, \dots, S_N^l\}$ to output σ given by

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

$\boldsymbol{W} = \{W_{ji}\}$ is a set of the synaptic weights whose element W_{ji} is a weight from the ith input node to the jth hidden node. $g_1(x)$ and $g_2(x)$ are transfer functions of the hidden nodes and the output node, respectively. A teacher network has the same architecture as that of the student. The weights of the teacher are given by $\boldsymbol{W}^0 = \{W_{ji}^0\}$. We assume a stochastic training algorithm which leads at long times to a Gibbs distribution of the weights. Energy of the system is defined as follows:

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

$$\epsilon(\boldsymbol{W};\boldsymbol{S}^l) = \frac{1}{2}[\sigma(\boldsymbol{W}^0;\boldsymbol{S}^l) - \sigma(\boldsymbol{W};\boldsymbol{S}^l)]^2. \quad (3)$$

The performance of the network is measured by the generalization function $\epsilon(\boldsymbol{W}) = \int d\boldsymbol{S}\,\epsilon(\boldsymbol{W};\boldsymbol{S})$, where $\int d\boldsymbol{S}$ represents an average over the whole space of inputs. The generalization error ϵ_g is defined by $\epsilon_g = \langle\langle\langle\epsilon(\boldsymbol{W})\rangle_T\rangle\rangle$, where $\langle\langle\,\rangle\rangle$ denotes the quenched average over the examples and $\langle\,\rangle_T$ is the thermal average. An input S_i^l is chosen according to a Gaussian distribution with variance unity.

In this paper, we will rely on the annealed approximation. Actually from our preliminary result using the replica trick, we have confidence that most of the qualitative behavior of the network is explained within the annealed calculation. As the studies of the single-layer perceptron have shown, the annealed calculation gives a pretty good quantitative prediction of the generalization curve [5]. So, we replace the quenched average of the free energy with the annealed average,

$$-\beta F = \langle\langle \ln Z\rangle\rangle \simeq \ln\langle\langle Z\rangle\rangle. \quad (4)$$

Here $\beta = 1/T$ in which the temperature T parametrizes the level of stochastic noise.

The free energy depends on the overlap order parameter matrices $\boldsymbol{R}, \boldsymbol{Q}$ defined as follows:

$$\begin{aligned} R_{jk} &= \frac{1}{N}\sum_{i}^{N} W_{ji}W_{ki}^{0}, \\ Q_{jk} &= \frac{1}{N}\sum_{i}^{N} W_{ji}W_{ki}. \end{aligned} \tag{5}$$

An interesting property of this fully connected machine is that by exchanging positions of hidden nodes we can construct another teacher which performs an equivalent task as the original teacher. In successive learning process the student approaches one of of many teachers constructed by permutation, which gives a particular structure of the matrix elements in Eq. (5). There are many possible forms of matrices according to various teachers. We assume the matrix elements are given by

$$\begin{aligned} Q_{jk} &= \delta_{jk} + (1-\delta_{jk})Q, \\ R_{jk} &= \delta_{jk}R_1 + (1-\delta_{jk})R_0. \end{aligned} \tag{6}$$

This means that the student happens to approach the original teacher. Also we assume the teacher does not have any correlation between sites, i.e., $N^{-1}\sum_i^N W_{ji}^0 W_{ki}^0 = \delta_{jk}$. In the limit where M and N go to infinity, we can use the saddle point analysis for the free energy. Then the saddle point condition gives the order parameters Q, R_0, R_1. We consider three different cases, where P is of the order of N, MN, and in the intermediate region. The free energy and the order parameters scale differently in each case. We consider $\mathrm{sgn}(x)$ and x as transfer functions. Our method can also be applied to other transfer functions.

(*i*) $P \sim O(N)$. Following Seung, Sompolinsky, Tishby, and Seung [5], we divide the free energy into two part G_0 and G_{an} [5],

$$-\beta F = N(G_0 + \alpha G_{\mathrm{an}}). \tag{7}$$

Here P scales as $\alpha = P/N$. G_0 and G_{an} are of the order of unity.

We calculated G_0 and G_{an} up to the order of $1/M$. Then we find an important result,

$$R_0 = R_1 \sim O\left(\frac{1}{M}\right). \tag{8}$$

This condition corresponds to the so-called permutation symmetry pointed out by Barkai, Hansel, and Sompolinsky [15]. As discussed above, the permutation of the hidden nodes of a given teacher yields many different teachers. Let us consider the energy surface in the phase space $\{W\}$. Each teacher is at a minimum of the energy surface. For a small P and a high T, all the teachers belong to a single thermally connected region in the phase space. In this case, a student does not know from which teacher to learn, and the student is roughly equidistant from all of the permuted teachers. This picture coincides with the permutation symmetry described by Eq. (8). The learning rate is also relatively fast. As P becomes the order of MN, there appear many thermally disconnected valleys around the permuted teachers. In this case, the permutation symmetry (PS) is broken, which will be discussed in case (ii).

In this PS phase, the free energy and the generalization error ϵ_g are given by

$$G_0 = -\frac{1}{2}\bar{Q} + \frac{1}{2}\ln(1 - \bar{R}_1^2 + \bar{Q}), \tag{9}$$

$$G_{\mathrm{an}} = \begin{cases} -\frac{1}{2}\ln(1 + 2\beta\epsilon_g, & g_2(x) = x \\ \ln\left(1 + \frac{e^{-2\beta}-1}{2}\epsilon_g\right), & g_2(x) = \mathrm{sgn}(x), \end{cases} \tag{10}$$

$$\epsilon_g = \begin{cases} \lambda_1 - \bar{R}_1\lambda_2^2 + \frac{\lambda_2^2}{2}\bar{Q}, & g_2(x) = x \\ \frac{2}{\pi}\cos^{-1}\left(\frac{\lambda_2^2 \bar{R}_1}{\sqrt{\lambda_1^2 + \lambda_1\lambda_2^2\bar{Q}}}\right), & g_2(x) = \mathrm{sgn}(x). \end{cases} \tag{11}$$

We define $\lambda_1 = \int_{-\infty}^{\infty} Dx\,[g_1(x)]^2$, $\lambda_2 = \int_{-\infty}^{\infty} Dx\,x\,g_1(x)$, where $Dx = \frac{dx}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$. We rescaled the order parameters as $\bar{R}_1 = MR_1$ and $\bar{Q} = MQ$. When α goes to infinity, the generalization error converges to a limit ϵ_0, given by

$$\epsilon_0 = \begin{cases} \lambda_1 - \lambda_2^2, & g_2(x) = x \\ \frac{2}{\pi}\cos^{-1}\left(\frac{\lambda_2}{\sqrt{\lambda_1}}\right), & g_2(x) = \mathrm{sgn}(x). \end{cases} \tag{12}$$

Note that ϵ_0 becomes zero when $\lambda_1 = \lambda_2^2$. A network with $g_1(x) = x$ belongs to this case. Here the permutation symmetry holds for all values of P. A network with a linear transfer function in the first layer maps to a single-layer perceptron with continuous weights. Effective weight of this perceptron from the ith input node is $M^{-1/2}\sum_j^M W_{ji}$ for a student and $M^{-1/2}\sum_j^M W_{ji}^0$ for a teacher. In the limit where M goes to infinity, the effective weights become continuous. The effective weights of all the permuted teachers are the same. The generalization error decays with an asymptotic form $\epsilon_g \propto 1/\alpha$. This explains why learning of the two-layer network with a linear transfer function in the hidden layer shows the same asymptotic behavior as that of a single-layer perceptron with continuous weights [5]. Generally, ϵ_0 is nonzero for a nonlinear $g_1(x)$, which means learning is not fully accomplished in the region $P \sim O(N)$.

(*ii*) $P \sim O(MN)$. In this case we introduce a new scaling for the free energy,

$$-\beta F = MN(G_0 + \alpha' G_{\mathrm{an}}), \tag{13}$$

where $\alpha' = P/MN$. G_0 and ϵ_g are given by

$$G_0 = -\frac{1}{2}(1 + R_1)\ln(1 + R_1) - \frac{1}{2}(1 - R_1)\ln(1 - R_1), \tag{14}$$

$$\epsilon_g = \begin{cases} \lambda_1 - \lambda_3 - (1 - R_1)\lambda_2^2, & g_2(x) = x \\ \frac{2}{\pi}\cos^{-1}\left[\sqrt{\frac{(\lambda_3 - R_1\lambda_2^2)^2 + (\lambda_1 - \lambda_2^2)\lambda_2^2}{\lambda_1(\lambda_1 - \lambda_2^2)}}\right], & g_2(x) = \mathrm{sgn}(x), \end{cases} \tag{15}$$

where $\lambda_3 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} DxDy\, g_1(\sqrt{1-R_1^2}x + R_1 y) g_1(y)$. $G_{\rm an}$ has the same form as in Eq. (10). The other order parameters are eliminated by the saddle point condition.

For a network with a nonlinear transfer function in the hidden layer, there are two solutions for R_1. One is the PS solution, where $R_1 = 0$ and ϵ_g is equal to ϵ_0 in Eq. (12). This corresponds to the limit where α goes to infinity in case (i). The other is the permutation symmetry breaking (PSB) solution where R_1 is nonzero. When either $g_1(x)$ or $g_2(x)$ is the signum function, R_1 is one and ϵ_g is zero. This solution describes a perfect learning state, which is also observed in the perceptron with binary weights [4,5]. When both transfer functions are continuous, ϵ_g decays exponentially from a small value for large α'.

There occurs a first order phase transition from the PS phase to the PSB one. The transition line $\alpha'_c(T)$ is determined by comparing the free energies of the two solutions. For $g_1(x) = \text{sgn}(x)$ and $g_2(x) = x$, $\alpha'_c(T) \propto -1/\ln T$ at low T. For $g_1(x) = g_2(x) = \text{sgn}(x)$, $\alpha'_c(T) - \alpha'_c(0) \propto e^{-2/T}$ and $\alpha'_c(0) \simeq 3.0$. It can be shown that the PS solution always exists everywhere in the α'-T plane. Therefore it is difficult to observe the transition in the simulation when the network is of a fairly large size. For a network of a small size, e.g., $M = N = 11$, we observe the transition in Monte Carlo simulation, which is shown in Fig. 1. This figure shows a change between the two phases in the distribution of matrix elements of the order parameter $\boldsymbol{R}$ defined in Eq. (5). In the PS phase the matrix elements are small, as can be seen in Eq. (8). On the other hand, the matrix elements shows a discrete spectrum in the PSB phase. Here the students does not always exactly coincide with one of the permuted teacher. This indicates that the structure of the PSB phase in this small network may not bed as simple as our annealed approximation for large M describes. In order to explain this PSB phase more precisely, we should also consider replica symmetry breaking, which is beyond of the scope of this paper.

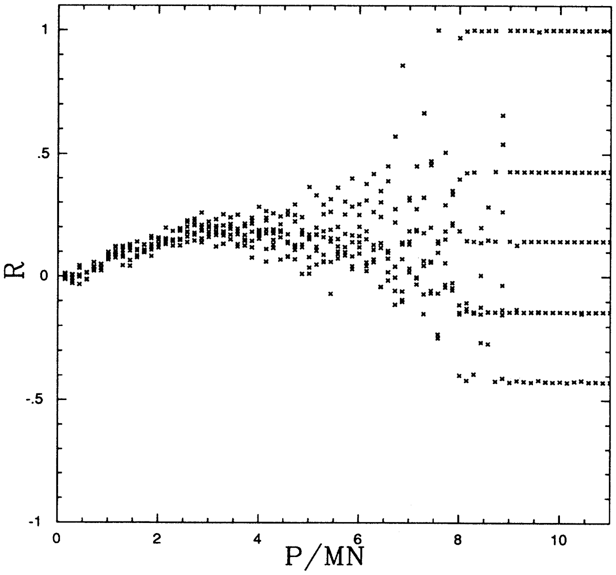

FIG. 1. Snapshot of the matrix elements R_{jk} for a network with the transfer functions $g_1(x) = g_2(x) = \text{sgn}(x)$, $N = M = 11$, and the temperature $T = 5$.

(iii) Intermediate region: $P \sim M^\delta N$. In this region, we study the network whose transfer function is the signum function for both layers. Note that the order parameters $\bar{R}_1$ and $\bar{Q}$ diverge as P/N goes to infinity. We examine this divergence in detail for $0 < \delta < 1$. As a result, we find R_1 and Q scale as

$$\begin{array}{l} R_1 = R_0 \propto M^{-\frac{4-\delta}{4}} \\ Q \propto M^{-\frac{2-\delta}{2}} \end{array} . \tag{16}$$

For large M, R_1 goes to zero, which leads to the PS solution discussed in case (ii).

The learning curve from the annealed calculation is compared with the Monte Carlo simulation. Figure 2 shows the learning curve for the case $g_1(x) = g_2(x) = \text{sgn}(x)$. Figure 3 shows the learning curve for the case $g_1(x) = g_2(x) = \tanh x$. For the latter case, the free energy and the generalization error do not have closed expressions as above. We find the saddle point and the generalization error numerically. The simulation results show pretty good agreement with the annealed calculations. The limiting value ϵ_0 of the generalization error is much smaller for the transfer function $\tanh(x)$ than for $\text{sgn}(x)$. We obtain smaller ϵ_0 and larger α'_c for smoother transfer functions.

There have been several efforts to derive an asymptotic behavior of the learning curve by analyzing the scaling of

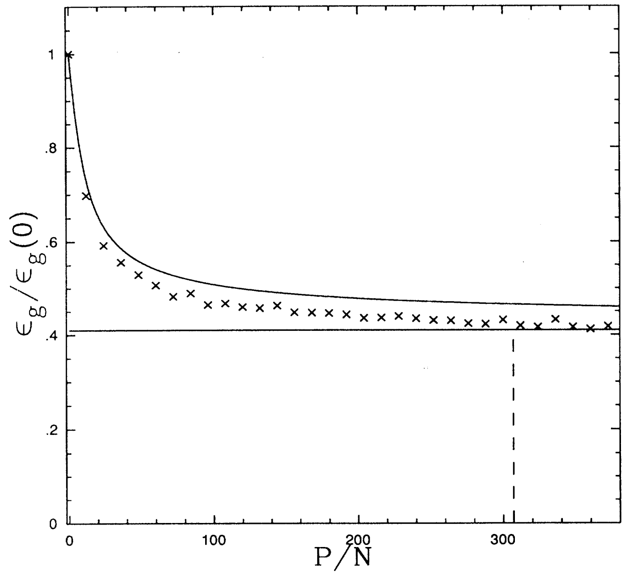

FIG. 2. Generalization curve for a network with the transfer functions $g_1(x) = g_2(x) = \text{sgn}(x)$, $N = M = 31$, and the temperature $T = 5$. The solid line is the analytic plot from the annealed approximation and the horizontal line denotes ϵ_0. The first order transition at $\alpha'_c \simeq 9.9$ is shown with the vertical dashed line. Dots show the result from the Monte Carlo simulation.

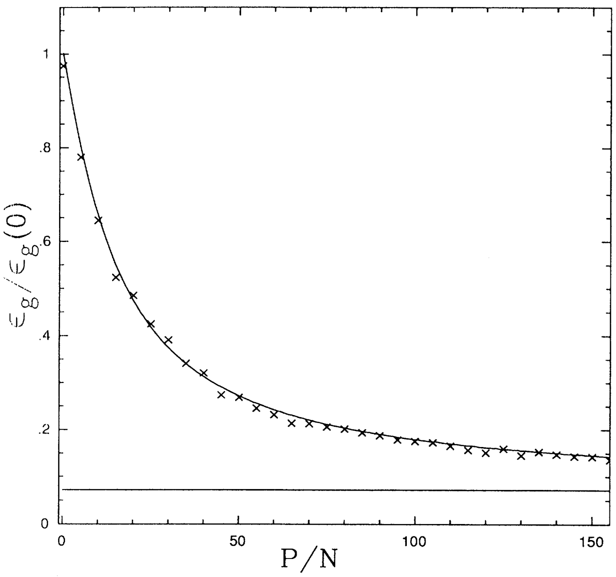

FIG. 3. Generalization curve for a network with the transfer functions $g_1(x) = g_2(x) = \tanh x$, $N = M = 31$, and the temperature $T = 5$. The solid line is the analytic plot from the annealed approximation and the horizontal line denotes ϵ_0. Dots show the result from the Monte Carlo simulation. $\alpha'_c \simeq 207$ is too large to be shown in this graph.

the volume of solution space [17,18]. These approaches are usually useful for a network with continuous weights. The generalization error for a network with continuous weights is inversely proportional to the ratio of training examples and the number of weights in the network. If we apply this to the two-layer machine with continuous weights, the generalization error should decay with a $1/\alpha'$ form. However, in the PS phase the permutation symmetry condition, which was not considered in these approaches, reduces the effective number of weights to the order of N. This explains a rather fast decay of the error with a power law $1/\alpha$ in the region $P \sim O(N)$. For $P \sim O(MN)$, the permutation symmetry breaks and the generalization error decays asymptotically as $1/\alpha'$. This expectation agrees with the result by Schwarze and Hertz [19] and also with our calculation not reported here. For the binary weights that we are dealing with in this paper, this asymptotic behavior is also found for $P \sim O(N)$, i.e., $\epsilon_g - \epsilon_0 \propto 1/\alpha$. For $P \sim O(MN)$, the network undergoes a discontinuous transition to the perfect learning at α'_c. There is no asymptotic decay in this case.

Summarizing our results, learning from examples in this fully connected two-layer network is difficult due to the existence of the metastable PS state where the PSB state is stable. If the system is moderately large, it is trapped in the metastable state in most cases. The minimum generalization error we have in the PS state critically depends on transfer functions. Many of the efforts to improve the back-propagation network have been devoted to the development of a better learning algorithm for fast convergence without considering the shape of the energy surface in the weight space. Our result shows that an alternative way for improvement is to analyze the energy surface structure and to find an optimal architecture of the network by considering various transfer functions and the size of the hidden unit. It may be useful to find an energy function which can avoid local minima other than the usual quadratic form [20]. Other approaches to construct a network by adding hidden nodes during training may also be helpful to avoid local minima [21,22].

Recently we came to know that H. Schwarze and J. Hertz were doing a similar calculation for the case where the transfer functions are sign function [19].

We are grateful to H. Schwarze for sending us a copy of his work prior to publication. This work was supported in part by the KOSEF through the Center for Advanced Materials Physics at POSTECH and the Center for Theoretical Physics of Seoul National University. It was also supported by the Basic Science Research Institute of POSTECH. K.K. and J.H.O. appreciate financial support from RIST.

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

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

[3] W. Krauth and M. Mézard, J. Phys. (Paris) **50**, 3057 (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] M. L. Minsky and S. Papert, *Perceptron* (MIT Press, Cambridge, MA, 1969).

[8] K. Kang, J.-H. Oh, C. Kwon, Y. Park, and H. S. Song, J. Kor. Phys. Soc. **25**, 270 (1992).

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

[10] S. Amari, N. Fujita, and S. Shinomoto, Neural Comput. **4**, 605 (1992).

[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, edited by D. E. Rumelhart and J. L. McClelland (MIT Press, Cambridge, MA, 1986), Vol. 2, pp. 318–362.

[12] M. Opper, and D. Haussler, Phys. Rev. Lett. **66**, 2677 (1991).

[13] H. Schwarze, M. Opper, and W. Kinzel, Phys. Rev. A **46**, R6185 (1992).

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

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

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

[17] S. Amari (unpublished).

[18] H. S. Seung, M. Opper, and H. Sompolinsky, in *Proceedings of the Fifth ACM Workshop on Computational Learning Theory* (ACM, New York, 1992), pp. 287–294.

[19] H. Schwarze and J. Hertz, Europhys. Lett. **21**, 785 (1993).

[20] E. Eisenstein and I. Kanter, Europhys. Lett. **21**, 501 (1993).

[21] P. Rujan and M. Marchand, Complex Syst. **3**, 229 (1989).

[22] M. Biehl and M. Opper, Phys. Rev. A **44**, 6888 (1991).