Generalization in a two-layer neural network

Kukjin Kang, Jong-Hoon Oh

Department of Physics, Pohang Institute of Science and Technology, Pohang, Kyongbuk, Korea

Chulan Kwon, Youngah Park

Department of Physics, Myong Ji University, Yongin, Kyonggi, Korea

(January 10, 1993)

Abstract

Learning of a fully connected two-layer neural networks with N input nodes, M hidden nodes and a single output node is studied using the annealed approximation. We study the generalization curve, i.e. the average generalization error as a function of the number of the examples. When the number of examples is the order of N, the generalization error is rapidly decreasing and the system is in a permutation symmetric(PS) phase. As the number of examples P grows to the order of MN the generalization error converges to a constant value. Finally the system undergoes a first order phase transition to a perfect learning and the permutation symmetry breaks. The computer simulations show a good agreement with analytic results.

87.10.+e, 05.50.+s, 64.60.Cn

Following the pioneering works of Gardner [1,2], there have been many efforts [3–6] to apply the methods of statistical mechanics to understand the feed-forward neural networks such as a perceptron [7]. Two main issues were studied in this context. Gardner studied the storage capacity of the single layer perceptron and there are several works in the same line [1–3]. The other problem is the learning curve, which shows a variation of the average error as a function of the number of training examples. For the single layer perceptron, this problem was studied extensively using the statistical mechanics [4–6,8,9] and also using the mathematical approach [10]. However, it is well known that the perceptron architecture can only solve a linear threshold problem. Many of the real-world problems are approached using the networks with a hidden layer and the error back-propagation algorithm [11]. Usually, direct application of the error-back-propagation algorithm has problems such as a large computing time required for learning and the existence of many local minima which prevents proper learning. Also estimation of an appropriate size of training example set for a valid generalization is an important issue. Study of generalization curve can give many useful information for such problems. However, existence of the hidden layer makes learning mechanism much more complex than that of single layer perceptron and not much is known about the generalization of multi-layer networks. Tree like 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 present a study of the generalization curve 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. Specially, we set all the weights in the second layer to 1. This architecture is usually called the committee machine. Suppose we have a network with arbitrary binary weights, we can always map this network into a committee machine by changing the sign of weights in the second layer which are -1 and the signs of 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 calculated the generalization curve for this machine using the annealed approximation. When the number of examples is the order of N, the generalization error is rapidly decreasing and the system is in a PS phase which we will explain below. When the number of examples P grows the order of MN the generalization error converges to a constant value and does not decrease until a first order phase transition to a perfect learning is reached. Above the transition the permutation symmetry breaks. The Monte Carlo simulations of the networks with various transfer functions well agree with the annealed calculations.

Using the permutation symmetry property, we are able to calculate the generalization curve using the similar method used for a perceptron learning. The network maps input vector $\boldsymbol{S}^l = \{S_i^l, \ldots, S_N^l\}$ to output σ as

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

$\boldsymbol{W} = \{W_{ji}\}$ 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 layer neuron and the output neuron respectively. A teacher network has the same architecture as that of the student. The weights of the teacher are given by $\boldsymbol{W}^0 = \{W_{ji}^0\}$. The training procedure is a Glauber dynamics which leads at long times to a Gibbs distribution of the weights. Energy of the system is defined as follows.

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

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

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 $\langle\langle \cdots \rangle\rangle$ denotes the quenched average over the examples and $\langle \cdots \rangle_T$ is the thermal average. An input S_i^l is chosen according to a Gaussian distribution with variance unity.

In this letter, we will rely on the annealed approximation. Actually from our preliminary result using the replica trick, we have a confidence that most of the qualitative behavior of the network is explained within the annealed calculation. Also it was known from the study of single layer perceptron that 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,

$$-\beta F = \langle\langle \ln Z \rangle\rangle \simeq \ln \langle\langle Z \rangle\rangle. \tag{4}$$

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.

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

An interesting property of this fully connected machine is that by exchanging positions of two hidden nodes we can construct another network which performs an equivalent task as the original network does. We might expect that the student has a close correlation with one of the networks constructed by the permutation of the teacher network. We assume this asymmetry makes diagonal elements of the matrix have different value with the rest of the matrix elements.

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

Also we assume the teacher does not have any correlation between sites, i.e. $N^{-1}\sum_i^N W^0_{ji} W^0_{ki} = \delta_{jk}$. In the thermodynamic limit where M and N go to infinity, we can use the saddle point analysis for the free energy. Then the saddle point solution gives the order parameters Q, R_0, R_1. As the order of magnitude of the number of examples varies, the free energy and the order parameters scale differently. We consider three different cases, where P is of the order of N, MN and in the intermediate region. We present the results for the cases where transfer functions are sign(x), x. Our method can also be applied to other transfer functions.

1. $P \sim O(N)$: The free energy can be divided into the entropy part G_0 and $G_{an} = \ln\langle\langle e^{-\beta\epsilon}\rangle\rangle$,

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

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

We find an important result,

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

This condition corresponds to the so called permutation symmetry pointed out by Barkai et al. [15] To show this, we calculated G_0 and G_{an} up to the order of $1/M$. Note that the permutation of the hidden nodes of a given teacher yields many networks which give equivalent outputs. Each of the permuted networks becomes an equivalent teacher. Let us consider the energy surface in the phase space $\{\boldsymbol{W}\}$. Each teacher is at a minimum of the energy surface. For a small number of examples P and a high temperature 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 in a equi-distance from all of the permuted teachers. This picture coincide with the permutation symmetry described by Eq. (8). The learning rate is also relatively fast. For a large P and a low T, there appear many thermally disconnected valleys around the permuted teachers. In this case, the permutation symmetry is broken, which will be discussed in the case 2.

In this PS phase, the free energy and the generalization error ϵ_g is given as

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

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

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

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.

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

When $\lambda_1 = \lambda_2^2$, ϵ_0 becomes zero. 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 i-th 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 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 non-zero for a non-linear $g_1(x)$, which means learning is not fully accomplished in the region $P \sim N$.

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

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

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

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

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

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_{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 of Eq. (12).

This corresponds to the limit where α goes to infinity in the case 1. The other is the permutation symmetry breaking(PSB) solution where R_1 is non-zero. When either $g_1(x)$ or $g_2(x)$ is the sign function, R_1 is one and ϵ_g is zero. This solution can be interpreted as a perfect learning state. When both transfer functions are continuous, we have small ϵ_g decaying exponentially for large α'. There is 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 energy of the two solutions. For $g_1(x) = \text{sign}(x)$ and $g_2(x) = x$, $\alpha'_c(T) \propto -1/\ln T$ at low T. For $g_1(x) = g_2(x) = \text{sign}(x)$, $\alpha'_c(T) - \alpha'_c(0) \propto e^{-2/T}$ and $\alpha'_c(0) \simeq 1.3$. It can be shown that the PS solution exists in the whole α'-T plane. Therefore it is very difficult to observe the transition from the simulation if the network is a fairly large size. For a network with smaller size, $M = N = 10$, we observed the transition from the the Monte Carlo simulation.

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

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

For large M, R_1 goes to zero, which leads to the PS solution discussed in the case 2.

The learning curve from the annealed calculation is compared with the Monte Carlo simulation. Fig. 1 shows the learning curve for the case $g_1(x) = g_2(x) = \text{sign}(x)$. Fig. 2 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 a closed form as above. We find the saddle point and the generalization error numerically. The simulation results show a pretty good agreement with the annealed calculations. The limiting value ϵ_0 of the generalization error is much smaller for the transfer function $\tanh(x)$. With a smother hidden layer transfer function, we obtain a smaller ϵ_0 and a larger α'_c.

There are some efforts to derive an asymptotic behavior of the learning curve by analyzing the scaling of the volume of solution space [17,18]. These approach is usually useful for a network with continuous weights. The generalization error for a machine with continuous weights is inversely proportional to the ratio of training examples and the number of weights in the network. If we apply this theorem 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 is not considered in this theorem reduces the effective number of weights to the order of N. This explain a rather fast decay of the error with a power law $1/\alpha$ in the region $P \sim N$. As the number of example P grows to the order of MN, the permutation symmetry breaks and the asymptotic behavior of the learning curve in the PSB phase slows down to the scale $1/\alpha'$. This expectation does agree with the result by Schwarze and Hertz [19] and also with our calculation not reported here. For binary weights we are dealing with here, the asymptotic behavior is similar in the PS phase. A combination of M binary weights behaves like a continuous weight. In the PSB region, binary weights affect the asymptotic behavior significantly. The asymptotic decay after the transition is not observed for most of the networks with sigmoidal transfer functions.

Summarizing our results, learning from examples in this fully connected two-layer network has an intrinsic limitation due to the existence of the metastable PS state above the thermodynamic transition. If the system has moderately large size, the system is trapped to the the metastable state in most cases. The minimum generalization error we have in the PS state critically depends on the form of the transfer function. Many of the efforts to improve the back-propagation network were concentrated on the development of better learning algorithm with fast convergence without consideration of the energy surface structure of the system. Our result shows that a more effective way for improvement is to analyze the phase space structure and to find an optimal architecture of the network with appropriate transfer functions. It may be useful to find an energy function which has less local minimum than the usual quadratic form [20]. Other approach where the network is constructed by adding hidden nodes during the training may be helpful to avoid the local minima [21,22].

We appreciate H. Schwarze for sending the preprint prior to publication. During the progress of this work [23] we came to know that they were doing the similar calculation for

the case where the transfer functions are sign function. 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 is also supported by the Basic Science Research Institute of POSTECH. K. K. and J. H. O. appreciate financial support from RIST.

REFERENCES

[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, in press.

[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, preprint (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, preprint (1992).

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

[19] H. Schwarze and J. Hertz, preprint (1992).

[20] E. Eisenstein and I. Kanter, preprint (1992).

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

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

[23] K. Kang and J.-H. Oh, in the presentation at *The 18th IUPAP International Conference on Statistical Physics*, Berlin, August 2-8, 1992.

FIGURES

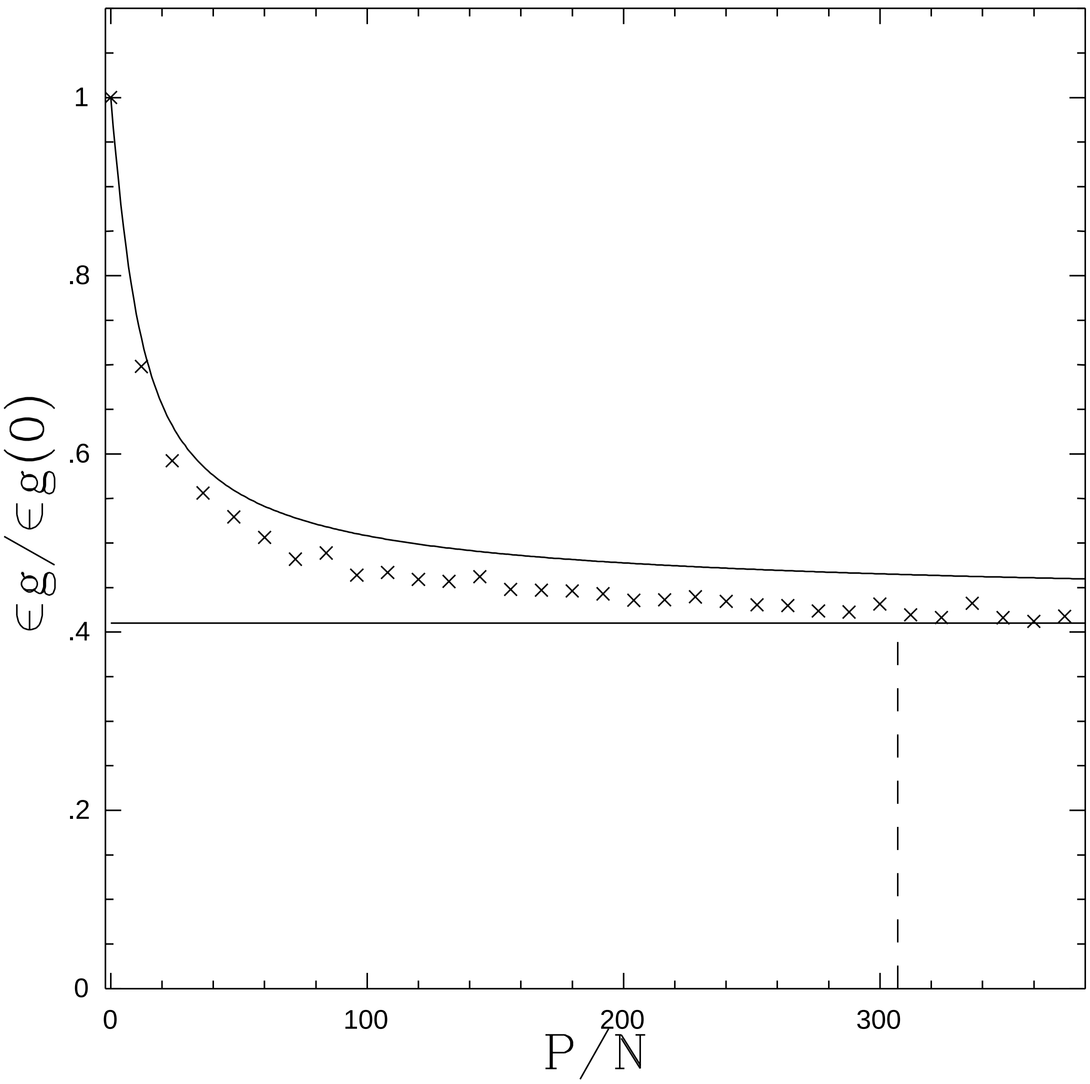

FIG. 1. Generalization curve for a network with the transfer functions $g_1(x) = g_2(x) = \mathrm{sign}(x)$, $N = M = 31$, and the temperature $T = 5$. The solid line is the analytic plot from the annealed approximation and the horizontal line denotes ϵ_0. The first order transition at $\alpha'_c \simeq 9.9$ is shown with the vertical dashed line. Dots show the result from the Monte Carlo simulation.

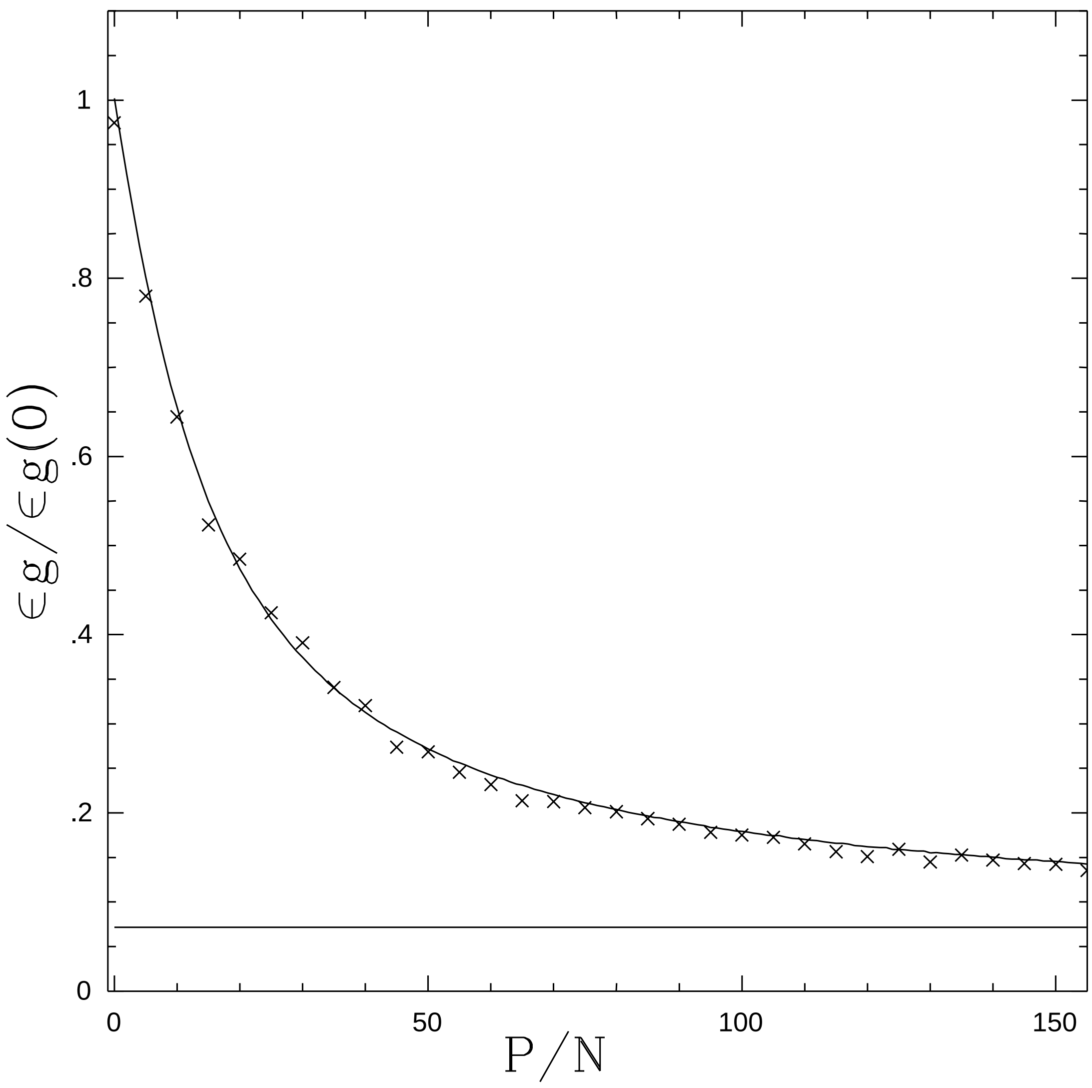

FIG. 2. Generalization curve for a network with the transfer functions $g_1(x) = g_2(x) = \tanh x$, $N = M = 31$, and the temperature $T = 5$. The solid line is the analytic plot from the annealed approximation and the horizontal line denotes ϵ_0. Dots show the result from the Monte Carlo simulation. $\alpha_c' \simeq 207$ is too large to be shown in this graph.