PHYSICAL REVIEW A VOLUME 46, NUMBER 10 15 NOVEMBER 1992

Generalization in a two-layer neural network

Holm Schwarze, Manfred Opper, and Wolfgang Kinzel
Institut für Theoretische Physik, Justus-Liebig-Universität Giessen, 6300 Giessen, Federal Republic of Germany

(Received 26 August 1992)

Statistical mechanics is applied to study the generalization properties of a two-layer neural network trained to implement a linearly separable problem. For a stochastic learning algorithm the generalization error as a function of the training set size is calculated exactly. The network with three hidden units experiences two first-order phase transitions due to an asymmetric freezing of the hidden units. Compared to a simple perceptron the committee machine is found to generalize worse.

PACS number(s): 87.10.+e, 05.20.−y, 02.50.+s

An interesting feature of neural networks is their ability to solve classification tasks by learning from examples [1]. After a set of correctly classified input-output pairs has been presented during a training phase they are able to generalize, i.e., to predict correct classifications even for novel input data.

Methods from statistical mechanics have been applied to quantify the generalization performance of neural networks [2–13]. Especially for the simplest network, the single-layer perceptron with one input layer and a single output unit, this approach was successful [7–13]. However, perceptrons are known to classify only linearly separable problems exactly [14], and additional layers of hidden units have to be added to implement more general tasks. On the other hand, it has been shown that just one hidden layer is sufficient to realize any Boolean function [2].

Usually in real applications the complexity of a classification task is not known *a priori*. So the actual size of a network required to obtain satisfying performance is a crucial parameter of interest. A lot of work has been done to understand how simple networks will perform on complicated, not completely learnable tasks. A common approach to the solution of a problem of unknown complexity is, however, to initially choose a large network which is believed to be able to learn the task. This choice may not be the most efficient possible solution to the problem, and it is important to understand how well a complex network will perform on an easy task. In this Rapid Communication we calculate the generalization error of such a problem exactly, namely, the generalization efficiency of a multilayer network trained to solve an "easy," linearly separable problem. As an example we consider a two-layer network with an odd number h of hidden units in addition to an input layer consisting of N units and a single output unit. Here we report results for binary couplings only, because this restriction allows a relatively easy comparison with computer simulations. It is important to note that without loss of generality the hidden-to-output couplings can be fixed to +1 in the binary case, hence the network determines the majority decision of the hidden units. A network of this type is called "committee machine" [15].

The network output

$$\sigma(\mathbf{S},\{\mathbf{J}^a\}) = \mathrm{sgn}\left[\sum_{a=1}^{h} \mathrm{sgn}(\mathbf{J}^a\cdot\mathbf{S})\right]$$

is a function of the input vector $\mathbf{S}=(S_1,\ldots,S_N)$ and the input-to-hidden coupling vectors $\mathbf{J}^a=(J_1^a,\ldots,J_N^a)$, $a\in\{1,\ldots,h\}$, where we have defined the scalar product $\mathbf{X}\cdot\mathbf{Y}=\sum_{j=1}^{N}X_jY_j$. The couplings are taken to be binary, $J_j^a\in\{-1,+1\}$, throughout this paper, so each hidden unit can be regarded as the output of a Boolean perceptron with discrete couplings and zero threshold. We study the case, where the complex "student," namely, our committee machine, learns from examples given by a simple "teacher," namely, a perceptron with binary weights $\mathbf{B}\in\{-1,+1\}^N$ and an output given by

$$\sigma_0(\mathbf{S}) = \mathrm{sgn}(\mathbf{B}\cdot\mathbf{S})\,. \tag{1}$$

The student network is trained by using a set of $P=\alpha N$ input vectors ξ^μ, $\mu\in\{1,\ldots,P\}$, and the corresponding desired outputs $\sigma_0(\xi^\mu)\in\{-1,+1\}$ given by the teacher (1).

A common learning strategy minimizes the training error, here defined as the number of incorrectly classified training examples:

$$E(\{\mathbf{J}^a\}) = \sum_{\mu=1}^{P} \Theta[-\sigma(\xi^\mu,\{\mathbf{J}^a\})\sigma_0(\xi^\mu)]\,,$$

where $\Theta[x]$ is the unit step function. The minimum of the training error, however, does not necessarily correspond to a minimal generalization error

$$\epsilon(\{\mathbf{J}^a\}) = \overline{\Theta[-\sigma(\mathbf{S},\{\mathbf{J}^a\})\sigma_0(\mathbf{S})]}^{\{\mathbf{S}\}} = \frac{1}{P}\overline{E(\{\mathbf{J}^a\})}^{\{\mathbf{S}\}}\,,$$

i.e., the probability that a randomly chosen input $\mathbf{S}$ is misclassified [12].

It is often useful to explore large regions in phase space of couplings by considering a stochastic learning algorithm. In equilibrium such a Monte Carlo process yields a Gibbs distribution of couplings $\exp[-\beta E(\{\mathbf{J}^a\})]$ characterized by a temperature $T=1/\beta$ measuring the amount of noise during training. In the limit of vanishing noise, $T\rightarrow 0$, this is equivalent to minimizing the training error $E(\{\mathbf{J}^a\})$.

Following an approach proposed by [12] we consider the high-temperature limit ($T,\alpha \rightarrow \infty$, $\alpha\beta$ finite). This limit has several advantages. First, an exact solution is easily derived. Second, for sufficiently large temperatures the Monte Carlo process gives a practical learning algorithm which is not the case for zero temperatures, i.e., learning without errors. In addition, the high-temperature limit has been shown in [12] to be a useful approximation revealing most of the qualitative features of the learning model without noise.

Following [12] the average over the distribution of training examples simply leads to the equilibrium distribution

$$P(\{\mathbf{J}^a\}) = \frac{1}{Z}\exp[-\alpha\beta N\epsilon(\{\mathbf{J}^a\})]$$

with

$$Z = \sum_{\{\mathbf{J}^a\}} \exp[-\alpha\beta N\epsilon(\{\mathbf{J}^a\})]\,. \tag{2}$$

The training error $E(\{\mathbf{J}^a\})$ has simply been replaced by its mean value $\alpha N\epsilon(\{\mathbf{J}^a\})$ neglecting any fluctuations around the mean value. Hence in this limit, training and generalization errors are equal. Nevertheless, most of the qualitative features of the learning model still exist. As will be shown below, the generalization error $\epsilon(\{\mathbf{J}^a\})$ for a given network can be expressed by two types of order parameters, the overlaps $R_a = (1/N)\mathbf{J}^a\cdot\mathbf{B}$ between the coupling vectors and the teacher perceptron, and the mutual overlaps $q_{a\beta} = (1/N)\mathbf{J}^a\cdot\mathbf{J}^\beta$ between coupling vectors of different hidden units, $\epsilon(\{\mathbf{J}^a\}) = \epsilon(\{R_a, q_{a\beta}\})$. In the thermodynamic limit ($N \rightarrow \infty$) (2) can be written as

$$Z = \int \prod_a dR_a \prod_{a\beta} dq_{a\beta} \exp[-\beta N f(\{R_a, q_{a\beta}\})]$$

with a free energy per input unit f, which may be written

$$\beta f(\{R_a, q_{a\beta}\}) = \tilde{\alpha}\epsilon(\{R_a, q_{a\beta}\}) - s(\{R_a, q_{a\beta}\}),\ \tilde{\alpha} = \frac{\alpha}{T}\,, \tag{3}$$

s being an entropy term. Note that in the free energy the product $\alpha\beta$ appears as the only temperature dependence. For sufficiently large training sets the effective temperature T/α of the system can be made small even in the high-temperature limit. This allows for nontrivial behavior due to a balance of the generalization error and the entropic term. Minimizing f with respect to the R's and q's yields the equilibrium value ϵ of the generalization error as a function of the effective load parameter $\tilde{\alpha}$.

The generalization error for a given network can be calculated by noting that the variables

$$x_0 = \frac{1}{\sqrt{N}}\mathbf{B}\cdot\mathbf{S},\ x_a = \frac{1}{\sqrt{N}}\mathbf{J}^a\cdot\mathbf{S},\ a \in 1,\ldots,h$$

for random uncorrelated inputs $S_i \in \{-1,+1\}$ are, by virtue of the multidimensional central limit theorem, correlated Gaussian variables with zero means and

$$\overline{x_0^2} = \overline{x_a^2} = 1,\ \overline{x_0 x_a} = \frac{1}{N}\mathbf{B}\cdot\mathbf{J}^a = R_a,\ \overline{x_a x_\beta} = \frac{1}{N}\mathbf{J}^a\cdot\mathbf{J}^\beta = q_{a\beta}\,.$$

To make further progress one is tempted to assume symmetry between the hidden units, $R_a \equiv R$, $q_{a\beta} \equiv q \,\forall a \neq \beta \in \{1,\ldots,h\}$. Surprisingly, this simple ansatz was not in accordance with our Monte Carlo simulations. There we observed a successive freezing of single hidden units into the teacher perceptron. If the majority of the hidden units is frozen, i.e., yields the correct output for every input $\mathbf{S}$, the committee machine gives the right answer regardless of the minority "opinion." Thus, the lowest energy is highly degenerate.

With this consideration in mind we make a more general, partially asymmetric ansatz, which is consistent with our computer simulations. We assume that c hidden units are frozen into the teacher, $R_a = 1 \forall a \in \{h-c+1,\ldots,h\}$, and assume symmetry between the remaining units, $R_a \equiv R$, $q_{a\beta} \equiv q \,\forall a \neq \beta \in \{1,\ldots,h-c\}$. The generalization error can now be written as

$$\epsilon(\{\mathbf{J}^a\}) = \text{Prob}\left[\text{sgn}(x_0)\sum_{a=1}^{h-c}\text{sgn}(x_a) < -c\right].$$

Calculating the probability according to the joint ($h-c+1$)-dimensional Gaussian distribution of the variables $x_0,\ldots,x_{h-c}$ yields the generalization error as function of R and q; we obtain

$$\epsilon(R,q) = 2\sum_{l=0}^{(h-1)/2-c}\binom{h-c}{l}\int_0^\infty Dx\int_{-\infty}^{+\infty} Dt\,\Phi^l\left(\frac{Rx - i(R^2-q)^{1/2}t}{(1-q)^{1/2}}\right)\Phi^{h-c-l}\left(-\frac{Rx - i(R^2-q)^{1/2}t}{(1-q)^{1/2}}\right)$$

with the notations $Dx = (dx/\sqrt{2}\pi)e^{-x^2/2}$, $\phi(x) = \int_{-x}^{\infty} Dt$.

The entropy $s(R,q)$ of all network configurations described by the parameters R and q can be calculated by considering the volume of phase space accessible to $\tilde{h} = h - c$ correlated coupling vectors $\mathbf{J}^a$

$$V(R,q) = \sum_{\{\mathbf{J}^a\}}\prod_{a=1}^{\tilde{h}}\delta\left(\frac{1}{N}\mathbf{J}^a\cdot\mathbf{B} - R\right)\prod_{(a,\beta)}\delta\left(\frac{1}{N}\mathbf{J}^a\cdot\mathbf{J}^\beta - q\right).$$

Using the integral representation of the δ function and applying the saddle-point method we obtain

$$s(R,q) = \frac{1}{N}\ln V(R,q)$$
$$= \tilde{h}R\hat{R} + \frac{\tilde{h}(\tilde{h}-1)}{2}q\hat{q} + \frac{\tilde{h}}{2}\hat{q} + \ln Z_0(\hat{R},\hat{q})$$

where $\hat{R},\hat{q}$ are determined by the saddle-point equations

$$\tilde{h}R = \frac{Z_1(\hat{R},\hat{q})}{Z_0(\hat{R},\hat{q})},\ \tilde{h}(\tilde{h}-1)q + \tilde{h} = \frac{Z_2(\hat{R},\hat{q})}{Z_0(\hat{R},\hat{q})}\,.$$

Here the abbreviation

$$Z_\nu(\hat{R},\hat{q}) = \sum_{l=0}^{\tilde{h}}\binom{\tilde{h}}{l}(\tilde{h}-2l)^\nu \exp\{(\tilde{h}-2l)\hat{R} - \tfrac{1}{2}(\tilde{h}-2l)^2\hat{q}\}$$

has been used.

We now apply this formalism to the committee machine with three hidden units. The equilibrium value of the generalization error as a function of $\tilde{\alpha}$ has to be determined at the global minimum of the free energy (3). The global

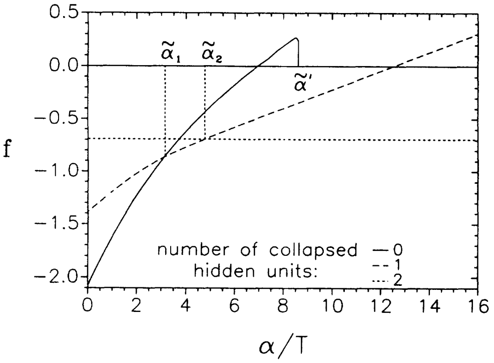

FIG. 1. High-temperature free energy for different numbers of frozen hidden units. First-order phase transitions appear at $\tilde{\alpha}_1 = 3.15$ and $\tilde{\alpha}_2 = 4.77$. The metastable ($c = 0$) state vanishes at $\tilde{\alpha}' = 8.60$.

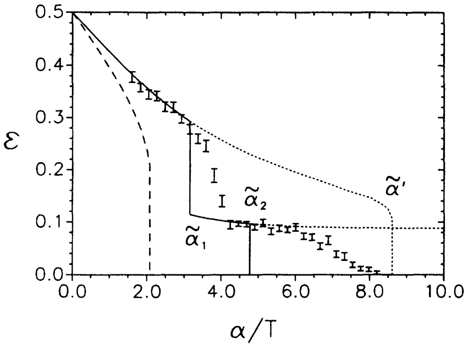

FIG. 2. High-temperature generalization error for the committee machine and Monte Carlo simulations (averaged over 50 samples). Dotted lines correspond to metastable states. The dashed line shows the high-T generalization error of a simple perceptron as calculated in [12].

minimum has been obtained by comparing the minimal values of the free energy for different numbers of frozen hidden units as shown in Fig. 1. With increasing size $\tilde{\alpha}$ of the training set we find two first-order phase transitions.

For small training sets, $\tilde{\alpha} < \tilde{\alpha}_1 = 3.15$, three symmetric hidden units ($c = 0$) correspond to the global minimum of $f(R,q)$; none of the hidden units is frozen. In the region $\tilde{\alpha}_1 < \tilde{\alpha} < \tilde{\alpha}_2 = 4.77$ the equilibrium situation is characterized by one hidden unit being frozen into the teacher perceptron ($c = 1$) but the remaining two units having an overlap $R < 1$ with the teacher. At the critical value $\tilde{\alpha} = \tilde{\alpha}_1$ the generalization error discontinuously drops to a lower but still finite value as shown in Fig. 2. The nonfrozen ($c = 0$) state remains metastable up to a value of $\tilde{\alpha} = \tilde{\alpha}' = 8.60$ where it completely vanishes. At the second critical value $\tilde{\alpha} = \tilde{\alpha}_2$ a second hidden unit freezes and the system reaches its ground state. The generalization error falls discontinuously to zero and the third unit has a vanishing overlap with the teacher corresponding to the maximal entropy.

Contrary to the nonfrozen state, the ($c = 1$) state does not disappear with increasing $\tilde{\alpha}$. The overlap R of the nonfrozen units to the teacher approaches the limiting value $R = \frac{1}{3}$ exponentially fast and the generalization error remains finite. From a practical point of view this implies the possibility that the learning process gets trapped in a metastable state even for large training sets.

Compared to a simple perceptron ($h = 1$), which has been studied recently in [12], the committee machine shows a lower generalization ability, as can be seen in Fig. 2. This behavior is in agreement with recent results [3,16] and reflects the fact that the target function (1) can be learned exactly by a simple perceptron. Adding hidden units to construct a committee machine increases the possible number of networks consistent with the training set and decreases the ability to find the correct generalization. This is in contrast to a strategy [13], where each unit is trained separately and the majority is taken afterwards.

We have performed Monte Carlo simulations for a committee machine with $N = 75$ input units and $h = 3$ hidden units at a training temperature of $T = 5$. As proposed in [17] the simulations were started with a small number of training examples. After the system was thermalized additional examples were added allowing the network to partially thermalize after each increasing of α. The generalization error, averaged over different sets of training examples, is shown in Fig. 2. The regions of a decreasing error are smeared out starting at the respective critical values due to finite system size and finite number of Monte Carlo steps. In these regions the generalization errors for different training sets show a double-peak structure indicating a first-order phase transition. Starting the simulation from high values of α and decreasing the number of training examples produces hysteresis loops at the two transition points $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$, respectively. Hence the numerical data are in agreement with the theory.

In summarizing our findings, the two-layer binary network experiences two first-order phase transitions with discontinuous decreases of the generalization error. The phase transitions occur due to a successive freezing of hidden units into the teacher perceptron implying a breaking of the symmetry between the hidden units.

In the general case of h hidden units we expect a cascade of phase transitions due to an increasing number of frozen coupling vectors. The generalization error is expected to drop to increasingly smaller values until the ground state is reached when more than half of the hidden units are perfectly aligned with the teacher perceptron.

Our results imply, that in order to obtain a satisfying generalization performance the size of the network should correspond to the complexity of the given task. A large network will suffer from a lower generalization ability due to a large number of degrees of freedom. The appearance of a number of first-order transitions furthermore causes the general problem of metastable states in which a learning algorithm may get trapped.

This work has been supported by the Volkswagenstiftung and the Höchstleistungs-Rechenzentrum Jülich.

[1] D. E. Rumelhart and J. L. McClellan, *Parallel Distributed Processing* (Bradford Books, Cambridge, 1986).

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

[3] P. Carnevali and S. Patarnello, Europhys. Lett. **4**, 1199 (1987).

[4] D. Haussler, N. Littlestone, and M. Warmuth, in *Proceedings of the Twenty-Ninth Annual IEEE Symposium on Foundations of Computer Science* (IEEE, Washington, DC, 1988), p. 100.

[5] E. B. Baum and D. Haussler, Neural Computation **1**, 151 (1989).

[6] E. Levin, N. Tishby, and S. A. Solla, in *Proceedings of the IEEE-Special Issue on Neural Networks* (IEEE, Washington, DC, 1989), Vol. 2.

[7] F. Vallet, J. Cailton, and P. Refregier, Europhys. Lett. **9**, 315 (1989).

[8] G. Györgyi and N. Tishby, in *Neural Networks and Spin Glasses*, edited by K. Thuemann and R. Köberle (World Scientific, Singapore, 1990).

[9] D. Hansel and H. Sompolinsky, Europhys. Lett. **11**, 687 (1990).

[10] M. Opper, W. Kinzel, J. Kleinz, and R. Nehl, J. Phys. A **23**, L581 (1990).

[11] G. Györgyi, Phys. Rev. A **41**, 7097 (1990).

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

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

[14] M. Minsky and S. Papert, *Perceptrons* (MIT Press, Cambridge, MA, 1969).

[15] N. J. Nilsson, *Learning Machines* (McGraw-Hill, New York, 1965).

[16] V. V. Anshelevich, B. R. Amirikian, A. V. Lukashin, and M. D. Frank-Kamenetskii, Biol. Cybern. **61**, 125 (1989).

[17] I. Kocher and R. Monasson (unpublished).