Learning and Generalization in a Two-Layer Neural Network: The Role of the Vapnik-Chervonenkis Dimension

Manfred Opper

Physikalisches Institut, Universität Würzburg, Am Hubland, D-97074 Wurzburg, Germany

(Received 12 November 1993)

Bounds for the generalization ability of neural networks based on Vapnik-Chervonenkis (VC) theory are compared with statistical mechanics results for the case of the parity machine. For fixed phase space dimension, the VC dimension grows arbitrarily by increasing the number K of hidden units. Generalization is impossible up to a critical number of training examples that grows with the VC dimension. The asymptotic decrease of the generalization error ε_G comes out independent of K and the VC bounds strongly overestimate ε_G. This shows that phase space dimension and VC dimension can play independent and different roles for the generalization process.

PACS numbers: 87.10.+e, 05.90.+m

Statistical mechanics has been successfully applied to the performance of feedforward neural networks. Based on Gardner's phase space approach [1], mainly two important questions have been treated in the last years. (1) How many *random* input/output relations can be learned by a network? This is *the storage capacity problem.* (2) How well can a network infer an unknown classification rule from examples? This defines *the generalization* problem (for a review, see [2-4]). So far, within this approach, no general relation between the two topics has been established. However, an important connection between the *worst case* generalization ability and a quantity that has a close relation to the storage capacity has been discovered by computer scientists [5,6]. This is the so-called *Vapnik-Chervonenkis* (VC) dimension $d_{\rm VC}$, which is the size of the largest set of input patterns for which *all* $2^{d_{\rm VC}}$ combinations of binary output labels can be learned by the network.

A remarkable combinatorial theorem, *Sauer's lemma* [7], yields an upper bound for the number of possible output combinations in terms of $d_{\rm VC}$. In the thermodynamic limit $d_{\rm VC} \to \infty$, this theorem implies that if the number m of input patterns exceeds $2d_{\rm VC}$, then only an *exponentially small* fraction of all possible 2^m combinations can be realized. Thus, for large $d_{\rm VC}$, the probability that a *random set* of output labels can be stored is vanishingly small, if $m > 2d_{\rm VC}$. Thus, for the storage capacity relative to the number N of weights, one finds $\alpha_c < 2d_{\rm VC}/N$.

If the output labels are not random, but are provided by a teacher's classification rule, that can be realized by the network, then, the results of [5,6] state that good inference of the rule is possible, if m, the size of the training set sufficiently exceeds $d_{\rm VC}$. A recent calculation for a perceptron [8] has shown that in the worst case, nontrivial generalization begins when the number of training examples exceeds the capacity. One can say, that "generalization begins, when learning ends" [9]. Networks with large capacities and VC dimensions seem to realize more complex mappings than those with smaller capacities.

The VC method provides exact bounds for the generalization error which hold for *any* network and learning algorithm that properly learn the training examples. In contrast to such worst case studies the methods of statistical physics naturally aim at an average case learning scenario. Here, one may think of a network that is generated by a stochastic (Gibbsian) training algorithm. Applied to networks with continuous weights that infer a learnable rule, the latter methods yielded generalization errors ε_G that decay asymptotically like $\propto \alpha^{-1}$, as $\alpha \to \infty$. Here $\alpha = m/N$ and N is the number of independent, adjustable network parameters (weights). N equals the dimension of the phase space [4,10,11]. Unlike the worst case results, for Gibbs learning generalization begins mostly right at $\alpha = 0$. Nevertheless, one could expect that also here the VC dimension plays a characteristic role for the decay of the generalization error. Actually, by a combination of VC methods and information theoretic ideas, recently, the rigorous and general upper bound

$$\varepsilon_G \le 2(m/d_{\rm VC})^{-1} \tag{1}$$

for the performance of the Gibbs algorithm was derived in [12]. If $d_{\rm VC}$ scales like N, then (1) compares well with the statistical mechanics results. In fact, for a single layer perceptron, one has $d_{\rm VC} = N$ [13], and the asymptotic result $\varepsilon \simeq 0.62\alpha^{-1}$ for large α and a spherical input distribution is not much overestimated by the general bound (1).

In this Letter I will give an example where this is not the case. I will study a model where the VC dimension and the dimension of phase space can differ substantially. This happens for a *parity machine* of N adjustable weights and K hidden units. For the *tree architecture*, it is known [14] that the storage capacity α_c increases with the number of hidden units like $\simeq \ln(K)/\ln(2)$ as $K \to \infty$. Since $d_{\rm VC} > \frac{1}{2}N\alpha_c$, the VC dimension will, for fixed number N of weights, grow at least logarithmically in K.

The question arises whether then the VC dimension is the *main relevant parameter* that determines the shape of the learning curve. To answer this question, I will present a calculation of the generalization error for general K. The case $K=2$ has been investigated recently in [15].

The parity machine with a tree architecture is composed of K independent subperceptrons, corresponding to nonoverlapping receptive fields, each with a coupling vector $\mathbf{w}_j$ of N/K weights. For a picture, see Fig. 1 of Ref. [14]. The jth branch of the tree receives a number of N/K inputs, abbreviated by the vector $\mathbf{x}_j$. The output σ of the parity machine is computed as the product of the K hidden units via

$$\sigma=\prod_{j=1}^{K}\operatorname{sgn}(\mathbf{w}_j\cdot\mathbf{x}_j)\,.$$

For the generalization problem, I assume that the network is trained on m examples, where the output labels $\sigma^m=\{\sigma_1,\dots,\sigma_m\}$ are provided by a *teacher* parity machine with couplings $\mathbf{w}^t=\{\mathbf{w}_1^t,\dots,\mathbf{w}_K^t\}$. In the thermodynamic limit, the learning curve of this neural network can be calculated from an entropy, utilizing the replica method. This calculation turns out to be rather simple in the *Bayesian framework of learning* [4,10,12,16] where an average over the classification rule of the teacher $\mathbf{w}^t$ is included.

As in [1], one defines the phase space volume $V(\sigma^m)$ of all networks which correctly learn a training set of m classification labels. I assume further that the total phase space has a volume normalized to 1, implying $\sum_{\sigma^m}V(\sigma^m)=1$. The entropy per weight, averaged over inputs and rules, is

$$S=N^{-1}\langle\langle\ln V(\sigma^m)\rangle_{\mathbf{w}^t}\rangle_{\mathbf{x}}\,.$$

The average over $\mathbf{w}^t$ gives a probability to each combination of outputs σ^m that equals $V(\sigma^m)$ [10]. Thus, replacing the average over the rules by a sum over outputs, one finds a simple expression [4,16,17] in the language of replicas:

$$\begin{aligned}S&=N^{-1}\sum_{\sigma^m}\langle V(\sigma^m)\ln V(\sigma^m)\rangle_{\mathbf{x}}\\&=N^{-1}\lim_{n\to1}\frac{\partial}{\partial n}\ln\sum_{\sigma^m}\langle V^n(\sigma^m)\rangle_{\mathbf{x}}\,.\end{aligned}\qquad(2)$$

By averaging over the rules $\mathbf{w}^t$, teachers and students appear in a completely symmetric way: Both are just randomly drawn networks that are consistent with all examples. This is reflected in the $n\to1$ limit of the replica formula (2), where the teacher is an additional replicon [17]. From this symmetry, for a spherical distribution of inputs, *a priori* only a single order parameter enters the subsequent calculations [18], which describes both the overlap between teacher and student and between two typical student networks. In replica symmetry [19], one obtains a similar expression as in [14] for the entropy of the parity machine

$$S=\operatorname{extr}_{[q]}\left\{\frac{1}{2}\ln(1-q)+\frac{q}{2}+\alpha\lim_{n\to1}\frac{\partial}{\partial n}M_n\right\},\qquad(3)$$

with

$$M_n=2\int_{-\infty}^{\infty}\prod_{l=1}^{K}Dt_i\left[\operatorname{Tr}_{[\sigma_i=\pm1]}\Theta\left(\prod_{i=1}^{K}\sigma_i\right)\prod_{i=1}^{K}H(\sigma_i\gamma t_i)\right]^n.\qquad(4)$$

$Dt=(dt/\sqrt{2\pi})e^{-t^2/2}$ and $H(x)=\int_x^\infty Dt$. Finally, $\gamma=(q/1-q)^{1/2}$, and $q=(K/N)\mathbf{w}_j^a\cdot\mathbf{w}_j^b$ is the overlap between two replicas a and b of the same subperceptron.

The expression (4) as it stands is not very suitable for practical calculations when K is sufficiently large. I will derive a more convenient expression *before* taking the limit $n\to1$.

Obviously the Θ function in (4) only contributes if for an even number of i, $\sigma_i=-1$. Fixing the set $\{t_i\}$ for a moment, it is helpful to think of $p_i^-=H(-\gamma t_i)$ as a probability that $\sigma_i=-1$ and $p_i^+=H(\gamma t_i)=1-H(-\gamma t_i)$ as the corresponding probability for $\sigma_i=+1$. Then clearly

$$\operatorname{Tr}_{[\sigma_i=\pm1]}\Theta\left(\prod_{i=1}^{K}\sigma_i\right)\prod_{i=1}^{K}H(\sigma_i\gamma t_i)=\Pr(\text{even no. of }\sigma_i=-1)\,.$$

By a direct expansion, one finds

$$\prod_{i=1}^{K}[2H(\gamma t_i)-1]=\prod_{i=1}^{K}(p_i^+-p_i^-)=\Pr(\text{even no. of }\sigma_i=-1)-\Pr(\text{odd no. of }\sigma_i=-1)\,.$$

Thus, one obtains

$$M_n=2\int_{-\infty}^{\infty}\prod_{l=1}^{K}Dt_i2^{-n}\left(1+\prod_{i=1}^{K}[2H(\gamma t_i)-1]\right)^n=2\sum_{l=0}^{\infty}\binom{n}{l}2^{-n}\left[\int_{-\infty}^{\infty}Dt[2H(\gamma t)-1]^l\right]^K,$$

where for integer n, $\binom{n}{l}=0$ for all $l>n$. Continuing to noninteger n, one uses $\lim_{n\to1}\partial/\partial n\binom{n}{l}=(-)^l1/l(l-1)$, for $l\geq2$. By the symmetry of the H function, the terms with odd l vanish and the entropy (4) is found from maximizing [20] the expression

$$S(q)=\frac{1}{2}\ln(1-q)+\frac{1}{2}+\alpha\left(-\ln(2)+\sum_{l=1}^{\infty}\frac{1}{2l(2l-1)}\left[\int_{-\infty}^{\infty}Dt[2H(\gamma t)-1]^{2l}\right]^K\right).\qquad(5)$$

The *generalization error* ε_G as a function of q is obtained by summing over all events, where an *odd* number of student subperceptrons disagree with their teacher counterparts. The probability for a disagreement on each of the independent subperceptrons is $\delta=(1/\pi)\arccos(q)$. Hence,

$$\varepsilon_G=\tfrac{1}{2}[1-(1-2\delta)^K]. \qquad (6)$$

For large α, the order parameter q will be close to 1 and the asymptotic behavior of ε_G can be found by expanding (5) to first order in $1-[1-2H(\gamma t)]^{2l}$. This leads to the result

$$\varepsilon_G\simeq\frac{2}{\sqrt{\pi}}\left\{\int_{-\infty}^{\infty}dt\,H(t)\ln[H(t)]\right\}^{-1}\alpha^{-1}=0.62\alpha^{-1},$$

independent of K. It is the same as for the simple perceptron ($K=1$). This might suggest that the generalization ability of the network is not much affected by the number of hidden units and thus by the storage capacity and VC dimension.

However, as has been shown for the case $K=2$ [15], the learning curve of the parity machine exhibits a nonsmooth behavior as a function of α: Because of internal symmetries, the state with $q=0$ ($\varepsilon_G=\frac{1}{2}$) is thermodynamically stable for sufficiently small α. For $q\rightarrow 0$, the entropy (5) is

$$S(q)\simeq-\frac{q^2}{4}+\alpha\left[-\ln(2)+\frac{1}{2}\left(\frac{2q}{\pi}\right)^K\right].$$

For $K=2$, the state with $q=0$ is locally stable, if $\alpha<\pi^2/8$ and a *second order* transition to nontrivial generalization ($\varepsilon_G>\frac{1}{2}$) occurs for $\alpha_0(2)=\pi^2/8$. For $K>2$, $q=0$ is locally stable for *all* α, suggesting a first order transition. To have generalization with nonzero q, the corresponding entropy $S(q)$ must exceed the value $S(q=0)=-\alpha\ln(2)$. The learning curves for $K=1,2,3,6$ are displayed in Fig. 1. ε_G begins to decrease at a critical value $\alpha_0(K)$, which grows with the number K of hidden units.

To understand this growth quantitatively, I will discuss the case when K is large. Then, ε_G is different from $\frac{1}{2}$ only, if δ in (6) is close to zero, i.e., for $q\approx 1$. To get a proper scaling for $K\rightarrow\infty$, I assume that $\delta\simeq K^{-1}$, such that $\delta K=\lambda$ is a new finite order parameter [21]. This scaling yields the generalization error

$$\varepsilon_G=\tfrac{1}{2}(1-e^{-2\lambda}).$$

With $\delta\simeq\sqrt{2}/\pi\sqrt{1-q}$, the ansatz implies the following asymptotic scaling of the entropy (5):

$$S\simeq-\ln(K)+\ln(\lambda)+\alpha\left[-\ln(2)+\sum_{l=1}^{\infty}\frac{1}{2l(2l-1)}\exp\left[-\frac{\lambda\sqrt{\pi}}{2}\int_{-\infty}^{\infty}dt\{1-[2H(\gamma t)-1]^{2l}\}\right]\right]. \qquad (7)$$

The large negative entropy contribution $-\ln(K)$ makes the state $\lambda=q=0$ globally stable for any finite α as $K\rightarrow\infty$. On the other hand, a locally stable state with nonzero λ fulfills $\partial S/\partial\lambda=0$, which is clearly independent of K. For large α, this equation yields $\lambda\propto\alpha^{-1}$. The corresponding entropy is to leading order

$$S\simeq-\ln(K)-\ln(\alpha). \qquad (8)$$

This solution becomes globally stable, if the entropy (8) exceeds the value $S(q=0)=-\alpha\ln(2)$. Thus, for large K, the transition can be found from the equation

$$\alpha_0=\frac{\ln(\alpha_0)+\ln(K)}{\ln(2)}. \qquad (9)$$

Equation (9) gives also a quite reasonable approximation for smaller K. For example, one finds $\alpha_0(3)\approx 3.31$ and $\alpha_0(6)\approx 4.87$ compared to the exact results 3.21 and 4.95.

The solution of Eq. (9) is displayed in Fig. 2 together

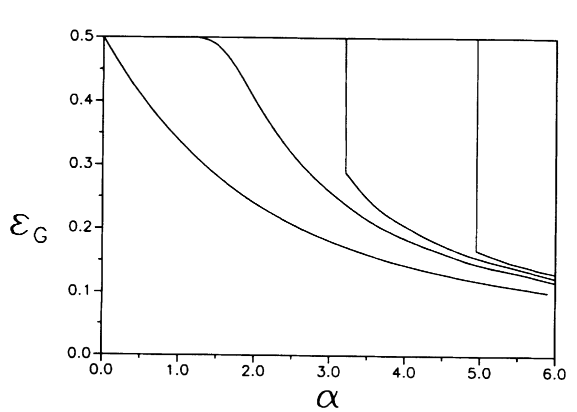

FIG. 1. Generalization errors for, from left, $K=1$ (perceptron) and $K=2,3,6$.

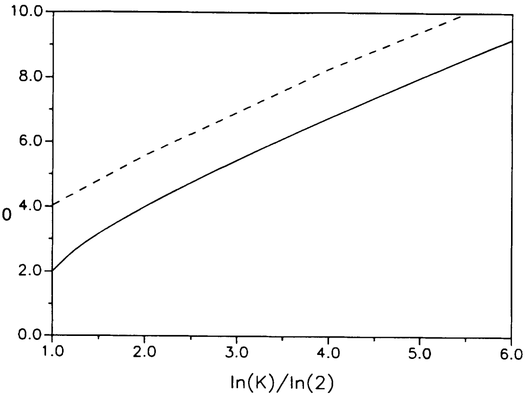

FIG. 2. Transition point $\alpha_0(K)$ to nontrivial generalization (full line) from the approximation (9), and storage capacity $\alpha_c(K)$ (dashed line, from [14]).

with the storage capacity taken from [14]. As can be seen, both quantities grow roughly with the same slope and for $K \rightarrow \infty$, one gets $\alpha_0 \simeq \ln(K)/\ln(2) \simeq \alpha_c$. Here, one finds a behavior for the average case Gibbs learning that resembles the prediction of VC theory for the worst case scenario: Generalization becomes possible only above capacity or VC dimension.

This result shows that the dimension of phase space and the VC dimension can play rather different and independent roles for the generalization performance of a multilayer net. While the VC dimension of the parity machine roughly determines the minimal number of examples that is needed to achieve nontrivial generalization, the asymptotic scaling of the learning curve only depends on the dimension of phase space.

It is interesting to compare this behavior with a result found for the *committee* machine with nonoverlapping receptive fields. As shown in [11], for $K \rightarrow \infty$ the learning curve converges to a limiting function that is smooth and decreasing. This may indicate that the capacity [22-24] and VC dimension for this machine does not diverge for $K \rightarrow \infty$ but rather converges to a finite value.

[1] E. Gardner, J. Phys. A **21**, 257 (1988).

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

[3] T. L. H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).

[4] M. Opper and W. Kinzel, "Physics of Neural Networks," edited by J. S. van Hemmen, E. Domany, and K. Schulten (Springer-Verlag, Berlin, to be published).

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

[6] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, J. Assoc. Comp. Mach. **36**, 929 (1989).

[7] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data* (Springer-Verlag, New York, 1982).

[8] A. Engel and C. Van den Broeck, Phys. Rev. Lett. **71**, 1772 (1993).

[9] T. Cover, in *Proceedings of the IVth Annual Workshop on Computational Learning Theory (COLT91), Santa Cruz, 1991* (Morgan Kaufmann, San Mateo, 1991).

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

[11] H. Schwarze and J. Hertz, Europhys. Lett. **20**, 375 (1992).

[12] D. Haussler, M. Kearns, and R. Schapire, in *Proceedings of the IVth Annual Workshop on Computational Learning Theory (COLT91), Santa Cruz, 1991* (Ref. [9]), pp. 61-74.

[13] T. M. Cover, IEEE Trans. El. Comp. **14**, 326 (1965).

[14] E. Barkai, D. Hansel, and I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990).

[15] D. Hansel, G. Mato, and C. Meunier, Europhys. Lett. **20**, 471 (1992).

[16] M. Opper and D. Haussler, in *Proceedings of the IVth Annual Workshop on Computational Learning Theory (COLT91), Santa Cruz, 1991* (Ref. [9]), pp. 75-87.

[17] A limit $n \rightarrow 0$ in Eq. (2) would lead to the entropy for the capacity problem.

[18] The case of a fixed teacher can be treated similarly, with the same result: A spherical distribution of inputs at each branch of the tree does not prefer a direction in the space of teachers. Thus the input averaged entropy cannot depend on the $\mathbf{w}_l^t$'s and must then be equal to the average over teachers.

[19] Replica symmetry is believed to be correct for networks with continuous weights when the rule is learnable.

[20] In contrast to the replica limit $n=0$, where one looks for local minima, in the present case ($n=1$), states with *maximal entropy* must be found.

[21] In contrast to the committee machine with tree architecture at large K, where the number of disagreeing subperceptrons becomes *Gaussian* distributed [11], this scaling corresponds to the *Poisson limit.*

[22] G. J. Mitchison and R. M. Durbin, Biol. Cybern. **60**, 345 (1989), derived the asymptotic upper bound $\alpha_c \leq \ln(K)/\ln(2)$, as $K \rightarrow \infty$, for the capacity.

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

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