Statistical physics estimates for the complexity of feedforward neural networks

Manfred Opper
Institut für Theoretische Physik, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany
(Received 20 September 1994)

Using simple information theoretic inequalities, a lower bound to the Vapnik-Chervonenkis (VC) complexity of neural networks is investigated. This bound is expressed by the average entropy used in the statistical mechanics approach to the network's generalization problem. Within the annealed theory, exact bounds to the VC dimension or the storage capacity can be calculated explicitly, without using the replica method. For the parity machine, the estimates of capacities match known upper bounds asymptotically, when the number of hidden units grows large.

PACS number(s): 87.10.+e, 05.90.+m

I. INTRODUCTION

Understanding the complexity of a learning process in artificial neural networks has become a fascinating topic in both computer science and statistical physics. Theoretically, supervised learning of a task is often modeled by a teacher network, which implements the concept to be learned and provides ideally classified training examples. During the learning process, a perfect response to the examples is achieved by the adjustment of the student network's weights. The complexity of a learning task manifests itself in the number of training examples that are necessary for the student to generalize from the learned examples, i.e., to give the correct response to unknown inputs with high probability. Using methods from statistical mechanics, exact calculations of learning curves, which display the average generalization error as a function of the size of the training set, were obtained in recent years for a variety of network models (for a review, see [1–3]). These results are believed to represent a typical learning behavior rather than the worst case. In fact, most of the derivations are based on the assumption of specific, natural input distributions and on averages over ensembles of learning networks [4].

A different approach to machine learning, based on uniform convergence results of mathematical statistics [5,6], is favored in theoretical computer science. Here exact bounds for the probability of correct generalization were obtained [7,8]. These hold for arbitrary input distributions and arbitrary networks which are consistent with the training data. The bounds are formulated in terms of a general complexity measure for the teacher networks that implement the rule. This is the so called Vapnik-Chervonenkis (VC) dimension [5,6], a quantity which is related to the teacher network's capacity. In the worst case, rules with a higher VC dimension are harder to learn than those with a lower VC dimension.

Is there a similar connection between capacity and generalization in the average case learning scenario treated by statistical mechanics? In a previous paper [9], I investigated this question for a special type of two-layer network, the parity machine. This work, like most statistical mechanics approaches to network learning, relied on the application and validity of the replica method. For multilayer networks, replica calculations become rather involved and less transparent. In particular, the estimation of capacities is a nontrivial task by the strong effects of broken replica symmetry.

In this paper I will study the aforementioned question using a technically simpler, but nevertheless exact approach. Combining information theoretic ideas developed by Hausler, Kearns, and Schapire [10] with well known mean field methods of statistical mechanics, exact lower bounds to the VC complexity can be obtained from the annealed entropy of the generalization problem.

As a result, for parity networks, exact lower bounds to their capacities are found, which asymptotically match known upper bounds. In such cases, the correct scaling of capacities, previously calculated by a replica symmetry breaking ansatz, is obtained in a much simpler, transparent way.

The paper is organized as follows: In Sec. II, the VC dimension and a corresponding entropy is introduced. Section III gives a lower bound to the VC entropy which is related to the generalization problem of the network. The simpler annealed theory is discussed in Sec. IV and applied to the estimation of VC dimensions or capacities of feedforward networks in Sec. V. Finally, Sec. VI contains a discussion of the results.

II. VC COMPLEXITY

As the simplest scenario of network learning, I consider the classification of a random input x (usually being a vector of features which is fed into the net) into two classes labeled by $\sigma\epsilon\{-1,+1\}$. For any set of m inputs $x^m=\{x_1,\ldots,x_m\}$, maximally 2^m classifications $\sigma^m=\{\sigma_1,\ldots,\sigma_m\}$ are possible. However, if m is sufficiently large, then a network of given architecture will realize only a much smaller number $\mathcal{N}(x^m)\ll 2^m$ of classifications upon varying its weights. Networks which have a larger variety of outputs seem to implement more complex rules, which are harder to learn, than those realizing a smaller $\mathcal{N}$. In fact, based on results of Vapnik and Chervonenkis [5,6], Blumer *et al.* [8] showed that, in the worst case, nontrivial generalization can be achieved only

if the number of training examples is comparable to the size d_{VC} of the largest set of inputs for which all $2^{d_{VC}}$ classifications can be realized by the teacher net. d_{VC} is the Vapnik-Chervonenkis (VC) dimension. For a perceptron, in the thermodynamic limit, this fact has been demonstrated explicitly in [11]. For a discussion of the VC approach in the statistical mechanics context, also see [12].

A remarkable combinatorial theorem, often called Sauer's lemma, bounds the growth of $\mathcal{N}(x^m)$ in terms of d_{VC}. This theorem, which was proved by Vapnik in the early 60s, [5,6], states that

$$\mathcal{N}(x^m) \le \sum_{i=0}^{d_{VC}} \binom{m}{i} \tag{1}$$

if $m \ge d_{VC}$. Specializing to the thermodynamic limit, where the number N of network parameters is large, the scaling m, N, $d_{VC} \to \infty$, keeping $\alpha = m/N$ and $\alpha_{VC} = d_{VC}/N$ fixed, seems natural. In this limit, the sum (1) can be simplified to give a bound for the VC entropy:

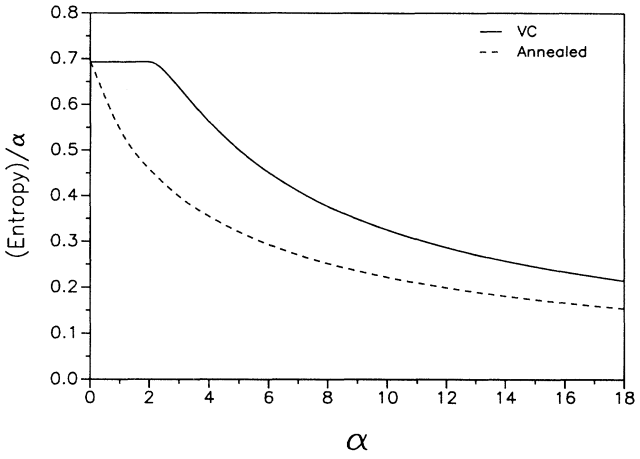

FIG. 1. VC entropy S_{VC}/α (solid line) and annealed entropy S_{ann}/α (dashed line) for the single layer perceptron with continuous weights.

$$S_{VC}(\alpha) = \frac{1}{N}\ln\mathcal{N}(x^m) \le \begin{cases} \alpha \ln(2) & \text{for } \alpha \le 2\alpha_{VC} \\ -\alpha\left[\dfrac{\alpha_{VC}}{\alpha}\ln\left[\dfrac{\alpha_{VC}}{\alpha}\right] + \left[1-\dfrac{\alpha_{VC}}{\alpha}\right]\ln\left[1-\dfrac{\alpha_{VC}}{\alpha}\right]\right] & \text{for } \alpha > 2\alpha_{VC}\,. \end{cases} \tag{2}$$

This function shows an interesting threshold phenomenon (see Fig. 1). If $\alpha > 2\alpha_{VC}$, then only an exponentially small fraction of all $2^{\alpha N}$ classifications can be realized. With a probability approaching 1 in the thermodynamic limit, a random choice of output labels cannot be implemented by the net, when $\alpha > 2\alpha_{VC}$. This result relates the storage capacity α_c of the network, via $\alpha_c \le 2\alpha_{VC}$, to its VC dimension. As is well known, the thermodynamic limit bound (2) is saturated [13] for the case of the single layer perceptron, independent of the set of inputs x^m [14]. In this case $d_{VC} = N$, $\alpha_{VC} = 1$, and the capacity equals α_c equals 2.

III. A LOWER BOUND TO THE VC ENTROPY

A lower bound to d_{VC} can be found from lower bounding S_{VC}. Such a bound is constructed by interpreting

$$NS_{VC} = -\sum_{\sigma^m} P(\sigma^m)\ln P(\sigma^m)$$

as the entropy for a discrete probability distribution $P(\sigma^m)$ of the output labels, where each realizable combination of outputs σ^m has equal probability $P(\sigma^m) = 1/\mathcal{N}(x^m)$. Thus NS_{VC} is the maximum entropy over all distributions on realizable σ^m. Any other such distribution will have an entropy that is a lower bound to S_{VC}.

I will now consider a class of distributions $P(\sigma^m)$, that naturally appears in the statistical mechanics of the generalization problem.

Let $V(\sigma^m, x^m)$ be the phase space volume [4] of all networks, that implement the set of m input-output pairs (x_k, σ_k) correctly. Further, let the indicator function $\theta_w(\sigma_k; x_k)$ be 1 if, for given (vector of) network weights w, the correct output is σ_k. If σ_k is the wrong output, then $\theta_w(\sigma_k; x_k) = 0$. Then one can write

$$P(\sigma^m) = V(\sigma^m, x^m) = \int dw\, p(w) \prod_{k=1}^{m} \theta_w(\sigma_k; x_k)\,. \tag{3}$$

$p(\omega)$ denotes a prior measure in the space of network parameters, with $\int p(\omega)dw = 1$. Obviously, $V(\sigma^m, x^m) = 0$ if the classification σ^m cannot be realized by the net. For a single layer perceptron with weight vector $\mathbf{w}$ and input vector $\mathbf{x}$, $\theta_w(\sigma_k; x_k)$ equals the Heaviside step function $\Theta(\sigma_k \mathbf{w}\cdot\mathbf{x}_k)$. Since V is non-negative and normalized to $\sum_{\sigma^m} V(\sigma^m, x^m) = 1$, it defines a probability. One concludes that the inequality

$$-\frac{1}{N}\sum_{\sigma^m} V(\sigma^m, x^m)\ln V(\sigma^m, x^m) \le \frac{1}{N}\ln\mathcal{N}(x^m) = S_{VC} \tag{4}$$

is valid for any prior distribution on the couplings and any set of inputs. The entropy (4) has a natural interpretation in the context of information theory [15]. $-\ln(V(\sigma^m, x^m)) = -\ln(P(\sigma^m))$ is the information that is gained from observing the labels σ^m. Thus the entropy equals just the average gain of information. This fact has been utilized by Haussler, Kearns, and Schapire [10] in order to bound the cumulative information gain during learning by means of the VC entropy. In a related context, this bound was used in [16].

The entropy of the discrete random variable σ^m in (4) has a second important interpretation in the statistical mechanics of generalization. Assume that a teacher net-

work with weights w is chosen at random from the prior distribution $p(w)$. Then the probability $P(\sigma^m)$ that the classification labels σ^m are the actual outputs of the teacher net is just $V(\sigma^m,x^m)$. Assume next that a student network is chosen at random from the so called version space, the space of all networks which correctly learn the training set (σ^m,x^m). This is precisely the Bayesian learning scenario discussed in [10,17]. The probability density for the random choice of the student is simply

$$p_m(\omega)=\begin{cases}\dfrac{p(w)}{V(\sigma^m,x^m)} & \text{for } w \text{ inside the version space}\\ 0 \text{ for } w \text{ outside the version space.}\end{cases} \tag{5}$$

The (differential) relative entropy for this probability density is then

$$-\int dw\, p_m(\omega)\ln\frac{p_m(w)}{p(w)}=\ln V(\sigma^m,x^m)\ .$$

Thus $-NS=\sum_{\sigma^m}V(\sigma^m,x^m)\ln V(\sigma^m,x^m)$ is equal to the differential entropy, averaged over the teachers, for an ensemble of student networks that learn ideally classified examples perfectly. Note the minus sign in the relation between discrete and differential entropies. Following Gardner's pioneering work, all order parameters which characterize a typical student in the thermodynamic limit can be calculated from the average of $\ln V(\sigma^m,x^m)$ over the examples. Thus, by performing the quenched average of $-S$ over the distribution of inputs, we obtain a solution to the generalization problem.

IV. ANNEALED ENTROPY

Entropies for the generalization problem have frequently been calculated using the replica method of statistical mechanics. In the thermodynamic limit, for a set of inputs that are drawn independently from the same distribution, the typical value of the entropy is given by the quenched average S_q of (4) over the input distribution. To avoid the difficulties of the replica method, I will use a further lower bound to (4), which is technically simpler. This bound is given by the annealed approximation [1] to S_q, which probably was applied for the first time to the generalization problem by Gardner and Derrida in [18]. Applying Jensen's inequality to the convex function $-\ln(\ )$, and using (3), we find

$$\langle S_{\rm VC}\rangle \geq S_q \geq S_{\rm ann}=-\frac{1}{N}\int dw^{(1)}p(w^{(1)})\ln\int dw^{(2)}p(w^{(2)})\left\langle\sum_\sigma \theta_{w^{(1)}}(\sigma;x)\theta_{w^{(2)}}(\sigma;x)\right\rangle^m$$
$$=-\frac{1}{N}\int dw^{(1)}p(w^{(1)})\ln\int dw^{(2)}p(w^{(2)})e^{m\ln[1-\varepsilon(w^{(1)},w^{(2)})]}\ , \tag{6}$$

where the brackets $\langle\ \rangle$ denote the expectation over the distribution of inputs and

$$\varepsilon(w^{(1)},w^{(2)})=\sum_\sigma\langle\theta_{w^{(1)}}(\sigma;x)\theta_{w^{(2)}}(-\sigma;x)\rangle \tag{7}$$

is the probability that network $w^{(1)}$ and network $w^{(2)}$ disagree on a random input. This is just the generalization error for network $w^{(2)}$, learning a task that is defined by network $w^{(1)}$. Equation (6) is the main result of this paper. It directly relates the VC entropy, measuring the complexity of a network, to the annealed approximation, which is a simple tool to describe a network's ability to generalize from learned examples.

V. APPLICATIONS

For many network problems, the high dimensional integrals in the annealed entropy (6) can be evaluated in the thermodynamic limit by the saddle-point method.

Of current interest are large networks with one layer of N input units and a second layer of K hidden units. For simplicity, I restrict myself to an architecture with nonoverlapping receptive fields that receive inputs abbreviated by the vectors $\mathbf{x}_j, j=1,\ldots,K$. For a picture, see [19]. Then the network consists of K subperceptrons with weight vectors $\mathbf{w}_j$ each having N/K couplings. Thus a hidden neuron computes an individual output given by $\mathrm{sgn}(\mathbf{w}_j\cdot\mathbf{x}_j)$. In addition, I assume that in the last layer, the output neuron computes a prewired Boolean function of the hidden outputs. A variety of examples for such networks can be found in [20]. This class of networks also includes the single layer perceptron for $K=1$ as a special case.

As is well known [21], for a spherical distribution of inputs and a flat prior distribution of coupling vectors $\mathbf{w}_l$, the order parameter $q=(K/N)\mathbf{w}_j^{(1)}\cdot\mathbf{w}_j^{(2)}$, completely determines the annealed entropy and generalization error (7). q measures the overlap between the subperceptrons of the networks 1 and 2 at the same hidden unit j, which, by symmetry, does not depend on j. The annealed entropy can then be written as

$$S_{\rm ann}=\min_q f(q) \tag{8}$$

where

$$f(q)=-\tfrac{1}{2}\ln(1-q^2)-\alpha\ln[1-\varepsilon(q)] \tag{9}$$

if the network weights are continuous, or

$$f(q)=\ln(2)+\frac{1-q}{2}\ln\left[\frac{1-q}{2}\right]+\frac{1+q}{2}\ln\left[\frac{1+q}{2}\right]-\alpha\ln[1-\varepsilon(q)] \quad (10)$$

for the case of binary weights, i.e., $\mathbf{w}_j\epsilon\{-1,+1\}^{N/K}$.

Before evaluating these expressions for specific models, let me briefly discuss a general strategy for obtaining bounds on the VC dimension when the networks have discrete weights which allow for totally $\mathcal{M}$ different network states. In such case, there is a critical number of examples α_g, above which only a single network, the teacher, gives the correct classifications to all training examples. Hence for the typical phase volume we have $V_{\rm typ}=1/\mathcal{M}$, $m\geq N\alpha_g$. Thus the quenched entropy freezes into the maximal value $S_q=\ln\mathcal{M}$ when $\alpha\geq\alpha_g$. Thus a lower bound to the VC dimension is obtained by solving the equation

$$S_{\rm VC}(\alpha_g)=\ln\mathcal{M}\ ,$$

with respect to $\alpha_{\rm VC}$. For a perceptron with N binary couplings, one has $\mathcal{M}=2^N$. Using the replica trick, the quenched result $\alpha_g=1.24$ was obtained in [22], giving the bound $\alpha_{\rm VC}\geq 0.306$. The simpler annealed entropy [18,22] which predicts the larger value $\alpha_g=1.44$ leads to the worse bound of $\alpha_{\rm VC}\geq 0.268$. The exact storage capacity of $\alpha_c=0.83$ obtained from a one step replica symmetry breaking (RSB) ansatz [23] yields the better bound $\alpha_{\rm VC}\geq\alpha_c/2=0.415$.

For continuous couplings we must use another type of argument. For such networks, the solution of the saddle-point equation (8) often yields a smooth behavior of $S_{\rm ann}$ as a function of α. In the limit $\alpha\to\infty$, the annealed theory usually leads to the scaling $\varepsilon\simeq c/\alpha$ and, using $\partial S_{\rm ann}/\partial\alpha\simeq\varepsilon$ the corresponding scaling $S_{\rm ann}\simeq c\ln(\alpha)$ for the annealed entropy is obtained. Here c is a model dependent numerical constant. Comparing with (2), the asymptotic behavior $S_{\rm VC}\simeq\alpha_{\rm VC}\ln(\alpha)$ yields the bound

$$\alpha_{\rm VC}\geq c\ .$$

Applying this result to the single layer perceptron, i.e., $K=1$, the order parameter equation (8) and (9) has to be solved using $\varepsilon(q)=(1/\pi)\arccos(q)$. The asymptotic behavior [1] gives $c=1$, so that the bound $\alpha_{\rm VC}\geq 1$ is obtained. Thus, for the perceptron, the bound is actually tight.

Let us next consider the committee machine, defined by the output $\sigma=\mathrm{sgn}[\sum_{l=1}^{K}\mathrm{sgn}(\mathbf{w}_l\cdot\mathbf{x}_l)]$. Here I will only discuss the limit $K\to\infty$, for which a nice expression for the generalization error was derived in [21]. It has the form $\varepsilon_{\rm com}(q)=\varepsilon[(2/\pi)\arcsin(q)]$. Here [21] found that $\varepsilon\simeq 2/\alpha$ within the annealed theory. Hence $c=2$, which implies the bound $\alpha_{\rm VC}\geq 2$. This result seems to underestimate the VC dimension drastically. The storage capacity obtained from a RSB calculation [24] was found to diverge for $K\to\infty$, thereby leading to a diverging VC dimension.

A third type of estimate for the complexity can be obtained for networks which show memorization without generalization when the size of the training set is lower than a certain critical value. In this case, our results seem to be more promising. We study the so called parity machine [25,9], which is defined by the output $\sigma=\prod_{l=1}^{K}\mathrm{sgn}(\mathbf{w}_l\cdot\mathbf{x}_l)$. Two parity machines produce different outputs to a given input only when both disagree at an odd number of subperceptrons. Since the probability for disagreement on a single perceptron is $\varepsilon_{\rm sing}(q)$, one finds the expression

$$\begin{aligned}\varepsilon_{\rm par}(q)&=\sum_{n\ \rm even}\binom{K}{n}\varepsilon_{\rm sing}^n(q)[1-\varepsilon_{\rm sing}(q)]^{K-n}\\&=\tfrac{1}{2}\{1-[1-2\varepsilon_{\rm sing}(q)]^K\}\\&=\frac{1}{2}\left[1-\left[1-\frac{2}{\pi}\arccos(q)\right]^K\right]\end{aligned} \quad (11)$$

for the generalization error. Using (9) with (11) we find that $q=0$ is always a locally stable minimum (see Fig. 2) of $f(q)$ for all α when $K\geq 3$. For $K=2$, this holds when $\alpha<\pi^2/8$. This fact has a simple physical explanation [25]: The output of the parity machine is invariant against the symmetry operation which transforms $\mathbf{w}_l$ into $-\mathbf{w}_l$ for an even number of subperceptrons l. As long as, for fixed teacher network $w^{(1)}$, all student nets $w^{(2)}$, which are related by this transformation, belong to the same ergodic component, then one has obviously $q=(K/N)\mathbf{w}_l^{(1)}\cdot\mathbf{w}_l^{(2)}=0$ as the globally stable state. This leads to the network's inability to generalize, i.e., $\varepsilon_{\rm par}=\frac{1}{2}$. When the correlation between network 1 and network 2 vanishes, both annealed and quenched entropies become equal and assume the value $S_q=S_{\rm ann}=\alpha\ln(2)$. A comparison with the VC entropy (2) shows that in this case all possible classifications can be realized by the machine. When α exceeds a critical value α_0, a second local minimum (see Fig. 2) of $f(q)$ becomes the global

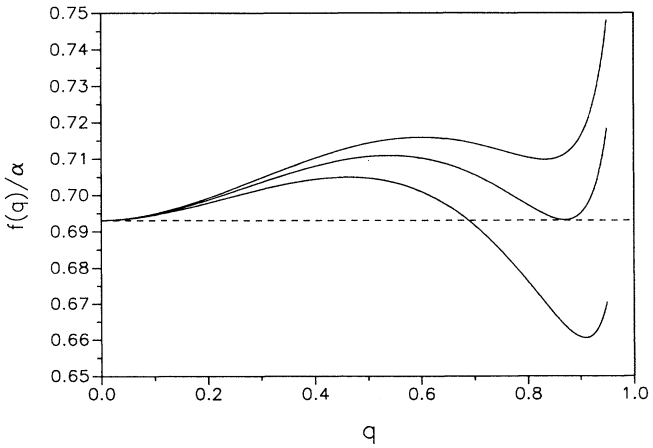

FIG. 2. $f(q)/\alpha$ [Eq. (9)] for the parity machine with $K=3$ hidden units and $\alpha=3.0$ (lower curve), $\alpha=\alpha_0=2.67$ (middle curve), and $\alpha=2.5$ (upper curve). The dashed line gives the value $\ln(2)$.

minimum, leading to a first order transition to $q>0$, if $K \geq 3$. Then, for the entropy, we have $S_{\text{ann}} < \alpha \ln(2)$. Hence α_0 gives an exact lower bound to the capacity α_c of the parity machine, i.e.,

$$\alpha_0 \leq \alpha_c \leq 2\alpha_{\text{VC}} .$$

This result holds in general: The number of examples, below which there is no correlation between learning networks (implying the inability to generalize), is always a lower bound to the capacity. Note that, for this result, no use of Sauer's lemma was made.

Solving the saddle-point equation for the parity machine yields $\alpha_0(K)$, as in the lower curve of Fig. 3. The second curve from below is an approximation to the exact quenched result for α_0, which was found in [9], Eq. (9). The third curve yields the storage capacity as obtained in [19] from a first step RSB. Finally, the upper curve is an exact upper bound to the capacity derived by Mitchison and Durbin [26]. All curves display the same asymptotic scaling $\alpha_0(K) \simeq \ln(K)/\ln(2)$, for K large. For the annealed lower bound, these asymptotics can be understood as follows. Large K also implies large α_0, where q jumps to a value close to 1. Expanding (11) for $q \to 1$ yields

$$\varepsilon \simeq \frac{K}{\pi} \arccos(q) \simeq K\sqrt{2}/\pi\sqrt{1-q} .$$

By solving the saddle-point equation using these asymptotics, we find $S_{\text{ann}} \simeq \ln(K) + \ln(\alpha)$. At $\alpha = \alpha_0$, this has to match with $\alpha_0 \ln(2)$, from which we obtain the correct slope of the curves in Fig. 3.

Both upper and lower curves represent exact results which do not rely on the application of the replica-technique. Hence, we have found an independent proof that the replica estimate of α_c obtained from RSB gives the correct scaling for the capacity. A different way of obtaining a lower bound to α_c was given in [27]. There the capacity was estimated for an explicit construction algorithm for the parity machine.

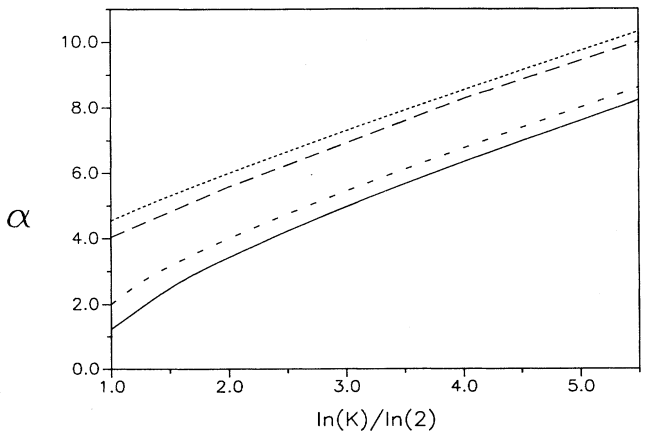

FIG. 3. Estimates for the capacity of a parity machine. The curves are from below: Exact lower bound from the annealed theory, lower bound from the quenched theory [9], result of the first step RSB calculation of [19], and exact upper bound of [26].

A similar analysis is possible for the parity machine with binary weights. In that case, from (10), $q=0$ is always a local minimum of $f(q)$ with $f(q=0)=\alpha \ln(2)$. There is only one second minimum for $q=1$, with $f(q=1)=\ln(2)$. Thus the annealed theory predicts that, for all K, a transition from trivial generalization $\varepsilon=\frac{1}{2}$ to perfect generalization $\varepsilon=0$ takes place at $\alpha_0=1$. Hence a lower bound to the capacity is always given by $\alpha=1$. On the other hand, $\alpha=1$ is also an upper bound, which may be seen from simple information theoretic arguments, or by calculating the annealed average of the phase space volume of weight vectors that store a set of random input-output pairs correctly. In agreement with the replica result of [28], we conclude that, for all K, the capacity is $\alpha_c=1$.

As the final example of a model which shows memorization without generalization, let me mention the so called reversed wedge perceptron [29], which is defined by the output $\sigma=\text{sgn}[h(h-\kappa)(h+\kappa)]$, with $h=(1/\sqrt{N})\mathbf{w}\cdot\mathbf{x}$. As has been shown in [30], for the case of binary weights and the special value $\kappa=\sqrt{2\ln 2}$, the state $q=0$ is the global minimum of the annealed entropy up to $\alpha=1$. This again yields $\alpha_c=1$, in agreement with the replica theory of [30].

VI. DISCUSSION

The simple bounds discussed in this paper relate the VC entropy, which measures the number of possible output configurations realizable by a network, and the annealed entropy of the average case generalization problem. It is interesting that the quality of the bounds differs drastically between various network models. For networks, like the parity machine, where a sharp transition from generalization inability to nontrivial generalization takes place, the estimated capacities are close to the values calculated from the replica theory. In such cases, it is true that "generalization begins, when learning ends" [31]. On the other hand, for networks like the committee machine, the learning curves remain smooth even when the number of hidden units grows arbitrary large. In such a case, the lower bound to the VC dimension remains finite. In contrast, the RSB estimates for the committee machine predict a capacity and dimension that diverge with K. From a mathematical viewpoint, both inequalities in (6) may not be very tight. However, for learnable problems, the transition from the quenched to the annealed average does not destroy the qualitative features of the learning curves for both parity and committee machines [21,9]. Thus we can expect that the large deviation between the estimates of capacities results from the first inequality in (6). For the VC entropy, all realizable output combinations are treated as equally probable. On the other hand, the phase space volumes (3) defined in Gardner's approach to learning usually will fluctuate over many orders of magnitude. The typical phase space volume calculated by the quenched entropy can then be very different from the VC estimate. Thus,

provided the RSB results for capacities are correct, we must conclude that the connection between VC dimension and the ability to generalize may be, at least in some average case setting, rather weak.

ACKNOWLEDGMENT

I am greatly indebted to David Haussler for many stimulating discussions on these topics.

[1] H. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).
[2] T. L. H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).
[3] 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).
[4] E. Gardner, J. Phys. A **21**, 257 (1988).
[5] V. N. Vapnik and A. Ya. Chervonenkis, Theor. Prob. Appl. **16**, 264 (1971).
[6] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data* (Springer-Verlag, New York, 1982).
[7] E. Baum and D. Haussler, Neural Comput. **1**, 151 (1989).
[8] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, J. Assoc. Comput. Mach. **36**, 929 (1989).
[9] M. Opper, Phys. Rev. Lett. **72**, 2113 (1994).
[10] D. Haussler, M. Kearns, and R. Schapire, in *IVth Annual Workshop on Computational Learning Theory (COLT91), Santa Cruz, 1991,* edited by Leslie G. Valiant and Manfred K. Warmuth (Morgan Kaufmann, San Mateo, 1991), pp. 61–74.
[11] A. Engel and C. Van den Broeck, Phys. Rev. Lett. **71**, 1772 (1993).
[12] J. Parrondo and C. Van den Broeck, J. Phys. A **26**, 2211 (1993); **26**, 7663 (1993).
[13] T. M. Cover, IEEE Trans. Electron. Comput. **14**, 326 (1965).
[14] For finite N, there is a difference between the bound (1) and the exact result, which can be neglected for the VC entropy in the thermodynamic limit.
[15] T. Cover and Joy A. Thomas, *Elements of Information Theory,* Wiley Series in Telecommunications (Wiley, New York, 1991).
[16] J. P. Nadal and N. Parga, Network **4**, 295 (1993).
[17] M. Opper and D. Haussler, Phys. Rev. Lett. **66**, 2677 (1991).
[18] E. Gardner and B. Derrida, J. Phys. A **22**, 1983 (1989).
[19] E. Barkai, D. Hansel, and I. Kanter; Phys. Rev. Lett. **65**, 2312 (1990).
[20] A. Priel, M. Blatt, T. Grossman, E. Domany, and I. Kanter, Phys. Rev. E **50**, 577 (1994).
[21] H. Schwarze and J. Hertz, Europhys. Lett. **20**, 375 (1992).
[22] G. Gyorgyi, Phys. Rev. A **41**, 7097 (1990).
[23] W. Krauth and M. Mézard, J. Phys. France **50**, 3057 (1989).
[24] E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992).
[25] D. Hansel, G. Mato, and C. Meunier, Europhys. Lett. **20**, 471 (1992).
[26] 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 $K \to \infty$, for the capacity.
[27] M. Biehl and M. Opper, Phys. Rev. A **44**, 6888 (1991).
[28] E. Barkai and I. Kanter, Europhys. Lett. **14**, 107 (1991).
[29] T. Watkin and A. Rau, Phys. Rev. A **45**, 4102 (1992).
[30] G. Bex, R. Serneels, and C. Van den Broeck (unpublished).
[31] T. Cover (unpublished).