Vapnik-Chervonenkis dimension of neural networks with binary weights

Stephan Mertens*
Institut für Theoretische Physik, Otto-von-Guericke Universität, Postfach 4120, D-39016 Magdeburg, Germany

Andreas Engel†
Institut für Theoretische Physik, Otto-von-Guericke Universität, Postfach 4120, D-39016 Magdeburg, Germany
(Received 11 September 1996)

We investigate the Vapnik-Chervonenkis (VC) dimension of the perceptron and simple two-layer networks such as the committee and the parity machine with weights restricted to values ± 1. For binary inputs, the VC dimension is determined by atypical pattern sets, i.e., it cannot be found by replica analysis or numerical Monte Carlo sampling. For small systems, exhaustive enumerations yield exact results. For systems that are too large for enumerations, number theoretic arguments give lower bounds for the VC dimension. For the Ising perceptron, the VC dimension is probably larger than $N/2$. [S1063-651X(97)05203-3]

PACS number(s): 87.10.+e

I. INTRODUCTION

Presently investigations in different fields including mathematical statistics, computer science, and statistical mechanics aim at a deeper understanding of information processing in artificial neural networks. Every field has developed its own concepts, which, although related to each other, are naturally not identical. In order to use (and appreciate) progress made in another field it is hence important to know the different concepts and there mutual relation. The Vapnik-Chervonenkis (VC) dimension is one of the central quantities used in both mathematical statistics and computer science to characterize the performance of classifier systems [1,2]. In the case of feedforward neural networks it establishes connections between the storage and generalization abilities of these systems [3–5]. Unfortunately, for most architectures the precise value of the VC dimension is not known and only bounds exist [6].

The VC dimension was introduced to characterize certain *extreme* situations in machine learning. It is therefore very useful to derive bounds for the network performance by considering the worst possible case. Complementary investigations in statistical mechanics focus on the *typical* behavior described by appropriate averages. In simple situations as provided, e.g., by the spherical perceptron, it turns out that the typical and worst case behaviors are not dramatically different [7]. It is then comparatively easy to establish connections between results obtained in different fields.

In the present paper we discuss some peculiarities that are encountered when analyzing the VC dimension of neural networks with binary weights. Binary weights are the extreme case of discrete couplings with obvious advantages in biological and technical implementations. It turns out, however, that in this case the typical and the extreme behavior of the network can be rather different. Therefore, the relation between results obtained by different approaches is less obvious.

Let us also note that in the mathematical literature binary weights are usually assumed to take on the values 0 and 1. Physically minded people, on the other hand, prefer the values -1 and 1 reminescent of spin systems. As we will show, these two choices make a difference.

The paper is organized as follows. After giving the basic definitions in the next section we discuss some simple examples in Sec. III. In Sec. IV we give a short discussion of analytical methods using the replica trick to calculate the behavior of the typical growth function. Section V is devoted to the numerical investigation of the typical growth function for the binary perceptron and simple two-layer networks. In Sec. VI we derive bounds for the VC dimension of neural networks with binary couplings including simple multilayer systems. The bounds show that the VC dimension is determined by *atypical* situations. The VC dimension hence cannot be inferred from the properties of the typical growth function. We give arguments that the value of the VC dimension for networks with binary weights may depend on whether the input vectors are continuous, binary (0,1), or binary $(-1,1)$. Finally, Sec. VII contains our conclusions.

II. BASIC DEFINITIONS

The VC dimension d_{VC} is defined via the growth function $\Delta(p)$. Consider a set X of instances x and a set C of (binary) classifications c: $x \to \{-1,1\}$ that group all $x \in X$ into two classes labeled by 1 and -1, respectively. In the case of feedforward neural networks [8] with N input units and one output unit, X is the space of all possible input vectors $\boldsymbol{\xi} \in \mathbb{R}$ or $\boldsymbol{\xi} \in \{-1,+1\}^N$, the class is defined by the binary output $\sigma = \pm 1$, and C comprises all mappings that can be realized by different choices of the couplings $\mathbf{J}$ and thresholds $\boldsymbol{\theta}$ of the network. For any set $\{x^\mu\}$ of p different inputs $x^1, \ldots, x^p$ we determine the number $\Delta(x^1, \ldots, x^p)$ of different output vectors $\{\sigma_1, \ldots, \sigma_p\}$ that can be induced by using all the possible classifications $c \in C$. A pattern set is called *shattered* by the class C of classifications if $\Delta(x^1, \ldots, x^p)$ equals 2^p, the maximal possible number of different binary classifications of p inputs. Large values of $\Delta(x^1, \ldots, x^p)$ hence roughly correspond to a large diversity

*Electronic address: stephan.mertens@physik.uni-magdeburg.de
†Electronic address: andreas.engel@physik.uni-magdeburg.de

of mappings contained in the class C. The growth function $\Delta(p)$ is now defined by

$$\Delta(p)=\max_{\{x^\mu\}} \Delta(x^1,\ldots,x^p). \tag{1}$$

It is clear that $\Delta(p)$ cannot decrease with p. Moreover, for small p one expects that there is at least one shattered set of size p and hence $\Delta(p)=2^p$. On the other hand, this exponential increase of the growth function is unlikely to continue for all p. The value of p where it starts to slow down should give a hint on the complexity of the class C of binary classifications. In fact, the Sauer lemma [1,9] states that for all classes C of binary classifications there exists a natural number d_{VC} (which may be infinite) such that

$$\Delta(p)=2^p \quad \text{if} \quad p\leqslant d_{\mathrm{VC}},$$
$$\Delta(p)\leqslant \sum_{i=0}^{d_{\mathrm{VC}}} \binom{p}{i} \quad \text{if} \quad p\geqslant d_{\mathrm{VC}}. \tag{2}$$

d_{VC} is called the VC dimension of class C. Note that it will, in general, depend on the set X of instances to be classified. Hence, in the case of neural networks there can be different values of d_{VC} for the same class of networks depending on whether the input patterns are real or binary vectors.

Due to the max in Eq. (1) it is possible that the VC dimension is determined by a single very special pattern set. In many situations emphasis is, however, on the *typical* properties of the system. In order to characterize the typical storage and generalization abilities of a neural network a probability measure $\mathcal{P}$ on the input set X is introduced. One then asks for the properties of the *typical* growth function $\Delta^{\mathrm{typ}}(p)$, which at variance with Eq. (1) is defined as the most probable value of $\Delta(x^1,\ldots,x^p)$ with respect to the measure $\mathcal{P}$. In the relevant limit of large dimension N of the input space it is generally assumed that the distribution of $\Delta(x^1,\ldots,x^p)$ is sharply peaked around this value. In the same limit $N\to\infty$ methods from statistical mechanics can be used to investigate the properties of $\Delta^{\mathrm{typ}}(p)$. This limit is nontrivial if $\alpha=p/N=O(1)$ and results in $d_{\mathrm{VC}}=O(N)$. We will call $\alpha_{\mathrm{VC}}=\lim_{N\to\infty} d_{\mathrm{VC}}/N$ the VC capacity of the neural network [10]. In addition, we may define $d_{\mathrm{VC}}^{\mathrm{typ}}$ as the value of op at which $\Delta^{\mathrm{typ}}(p)$ starts to deviate from 2^p and $\alpha_{\mathrm{VC}}^{\mathrm{typ}}=\lim_{N\to\infty} d_{\mathrm{VC}}^{\mathrm{typ}}/N$. The storage threshold p_c is as usual defined by $\Delta^{\mathrm{typ}}(p_c)/2^{p_c}=1/2$ and $\alpha_c=\lim_{N\to\infty} p_c/N$ is the storage capacity.

Using Stirling's formula in Eq. (2) and replacing the sum by an integral, one can show that for large N the relative deviation of the upper bound from 2^p becomes $O(1)$ if $\alpha>2\alpha_{\mathrm{VC}}$ (see Sec. IV). Since we always have $\Delta^{\mathrm{typ}}(\alpha)\leqslant\Delta(\alpha)$ this implies

$$\alpha_c\leqslant 2\alpha_{\mathrm{VC}}. \tag{3}$$

In this paper we concentrate on three sets of classifiers: the Ising percpetron, the Ising committee tree and the Ising parity tree. The *Ising perceptron* realizes the classification $\boldsymbol{\xi}\mapsto\pm 1$ via

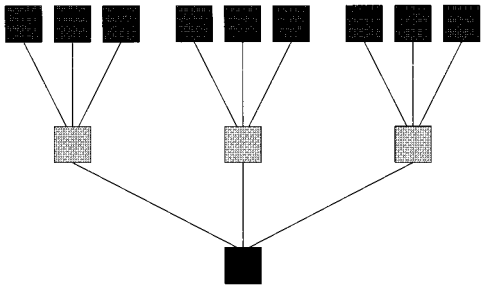

FIG. 1. Feedforward two-layer network with tree structure and nine input nodes, three hidden nodes, and one output node.

$$\sigma=\mathrm{sgn}\left(\sum_{i=1}^{N} J_i\xi_i-\theta\right), \tag{4}$$

with weight vector $\mathbf{J}\in\{\pm 1\}^N$. The perceptron is a prototype of what is usually termed single-layer feedforward networks: The N input values ξ_i are summed up and the resulting ''field'' is passed through a nonlinear function to yield a single output value σ. The computational capabilities of single-layer feedforward networks are rather limited. Hence one is interested in multilayer networks, where the output of single-layer networks is used as input for another single-layer network. The *Ising committee machine* and the *Ising parity machine* are examples of two-layer networks. In both machines, the input values ξ_i are mapped to K binary values τ_k by K Ising perceptrons. The τ_k are called the internal representation of the input. The internal representation is mapped onto the final output $\sigma=\pm 1$ by the so-called decoder function in the output layer. The decoder function is different in both machines. The committee-machine uses a perceptron with all weights $+1$,

$$\sigma=\mathrm{sgn}\left(\sum_{k=1}^{K}\tau_k\right)=\mathrm{sgn}\left[\sum_{k=1}^{K}\mathrm{sgn}\left(\sum_{i=1}^{N} J_i^{(k)}\xi_i-\theta\right)\right], \tag{5}$$

where $\mathbf{J^{(n)}}$ is the weight vector of the perceptron that ''feeds'' the kth hidden node. The restriction to all weights $+1$ in the output perceptron is not as severe as it may appear: The storage properties of this architecture are the same as for a machine where the output perceptron is an arbitrary Ising perceptron (see Appendix A).

The parity machine simply takes the parity of the internal representation

$$\sigma=\prod_{k=1}^{K}\tau_k=\prod_{k=1}^{K}\mathrm{sgn}\left(\sum_{i=1}^{N} J_i^{(k)}\xi_i-\theta\right). \tag{6}$$

In general, a hidden node can receive input from all input nodes. In this case we have NK weights to specify. If the input nodes are distributed among the hidden nodes such that no input node feeds more than one hidden node, the net has a tree structure (see Fig. 1). For simplicity we will assume that the input nodes are distributed evenly among the hidden nodes, i.e., each subperceptron has N/K weights.

III. SOME SIMPLE EXAMPLES

To begin with let us discuss some simple examples. In the case of the spherical perceptron defined by Eq. (4) but now with $\mathbf{J}\in\mathbb{R}$, $\Sigma_j J_j^2=N$, the exact results $d_{\mathrm{VC}}=\mathrm{N}+1$ and $\alpha_c=2$ have been obtained analytically [11]. Moreover, it is well known that the number of different realizable output vectors (dichotomies) is the same for all input pattern sets in general position [11]. Hence the max in Eq. (1) is realized by *almost all* possible inputs sets of length p and $\Delta^{\mathrm{typ}}(\alpha)=\Delta(\alpha)$. Furthermore, Eq. (3) is satisfied as equality.

A particular simple pattern set for which the result for the VC dimension can easily be verified is given by

$$\begin{aligned}\boldsymbol{\xi}^0&=(0,0,0,\ldots,0),\\ \boldsymbol{\xi}^1&=(1,0,0,\ldots,0),\\ \boldsymbol{\xi}^2&=(0,1,0,\ldots,0),\\ &\vdots\\ \boldsymbol{\xi}^N&=(0,0,\ldots,0,1).\end{aligned}\tag{7}$$

An arbitrary output vector $(\sigma_0,\sigma_1,\ldots,\sigma_N)$ can be realized for these inputs by using $J_j=\sigma_j$ and $\theta=-\sigma_0/2$.

Another interesting example is provided by a perceptron (4) for which the couplings are constrained to take the values $J_j=\{0,1\}$ only. Using the set of input patterns described above but omitting $\boldsymbol{\xi}^0$, an arbitrary output string $(\sigma_1,\ldots,\sigma_N)$ can be realized by using $J_j=(1+\sigma_j)/2$ and $\theta=-1/2$. Therefore $N\leqslant d_{\mathrm{VC}}\leqslant N+1$. On the other hand, it is known that the storage capacity of this perceptron is given by $\alpha_c=0.59$ [12]. This large difference between α_c and $2\alpha_{\mathrm{VC}}$ is due to the fact that the VC dimension is determined by a very special pattern set and that $\Delta^{\mathrm{typ}}(p)$ is much smaller than $\Delta(p)$. Hence the number of realizable output vectors is no longer the same for all input vectors in general position.

Finally, we consider the so-called Ising perceptron, again described by Eq. (4), but now with the constraint $J_j=\pm 1$ on the couplings. Since the couplings used above to show that the pattern set (7) is shattered by a spherical perceptron fulfill this constraint it is clear that the VC dimension of the Ising perceptron is for patterns $\boldsymbol{\xi}\in\mathbb{R}$ equal to $N+1$, exactly as for the spherical perceptron. For $\theta=0$ we get $d_{\mathrm{VC}}=N$ in both cases.

For binary input patterns $\xi_i=\pm 1$ we transform the pattern set (7) according to $\xi_i\to 2\xi_i-1$. Every output vector $(\sigma_0,\sigma_1,\ldots,\sigma_N)$ can then be realized by using $J_j=\sigma_j$ for $j=1,\ldots,N$ and $\theta=-\sigma_0-\Sigma_j\sigma_j$. Therefore the VC dimension is again $d_{\mathrm{VC}}=N+1$. However, since much of the interest in neural networks with discrete weights is due to their easy technical implementation it is not consistent to design an Ising perceptron with a threshold of order N. More interesting is the determination of the VC dimension of the Ising perceptron without (for N odd) or with a *binary* threshold $\theta=\pm 1$ (for N even) for binary patterns. This is a hard problem (see Sec. VI).

We note that the storage capacity of the Ising perceptron has been shown to be $\alpha_c=0.83$ [13]. Hence, also in this case we have $\alpha_c<2\alpha_{\mathrm{VC}}$ and the VC dimension is not determined by typical pattern sets. We also note that the storage capacity is believed to be the same for binary and Gaussian patterns [13–15]. As we will see in Sec. VI, it is unlikely that this holds also for the VC dimension.

IV. ANALYTICAL METHODS

Let us fix a particular set $\{\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^p\}$ of input patterns fed into a neural network with parameters $\mathbf{J}$. Different values of the parameters will result in different output strings $\{\sigma^1,\ldots,\sigma^p\}$ and hence the input patterns induce a partition of the parameter space into different cells labeled by the realized output sequences $\{\sigma^\mu\}$. The cells have a certain volume $V(\{\sigma^\mu\})$, which might be zero if the output string $\{\sigma^\mu\}$ cannot be realized. An interesting quantity is the number of cells of a given size

$$\mathcal{N}(V)=\mathrm{Tr}_{\{\sigma^\mu\}}\delta(V-V(\{\sigma^\mu\})),\tag{8}$$

which, of course, still depends on the particular set of input patterns $\{\boldsymbol{\xi}^\mu\}$. It is possible to calculate the typical value of $\mathcal{N}(V)$ for randomly chosen $\{\boldsymbol{\xi}^\mu\}$ using multifractal methods and an interesting variant of the replica trick [16]. This calculation has been explicitly performed for both the spherical and the Ising perceptron [17,18] and we give in this section a brief summary of the results relevant for the present paper.

For the perceptron (4) (with $\theta=0$ for simplicity) we have

$$V(\{\sigma^\mu\})=\int d\mu(J)\prod_\mu\theta(\sigma^\mu\mathbf{J}\cdot\boldsymbol{\xi}^\mu),\tag{9}$$

where $\int d\mu(J)=(2\pi e)^{-N/2}\int\Pi_j dJ_j\,\delta(\Sigma_j J_j^2-N)$ for the spherical perceptron and $\int d\mu(J)=2^{-N}\Sigma_{J_j=\pm 1}$ for the Ising case. The natural scale of V for $N\to\infty$ is then 2^{-N} and it is convenient to introduce $k(\{\sigma^\mu\})=-1/N\,\log_2 V(\{\sigma^\mu\})$ as a measure for the size of the cells. Similarily, the number of cells is exponential in N and we therefore use

$$c(k)=\frac{1}{N}\log_2\mathcal{N}(k)=\frac{1}{N}\log_2\mathrm{Tr}_{\{\sigma^\mu\}}\delta(k-k(\{\sigma^\mu\}))\tag{10}$$

to characterize the cell size distribution. Realizing that $c(k)$ is the microcanonical entropy of the spin system $\{\sigma^\mu\}$ with Hamiltonian $Nk(\{\sigma^\mu\})$ it can be calculated from the free energy

$$f(\beta)=-\frac{1}{\beta N}\log_2\mathrm{Tr}_{\{\sigma^\mu\}}2^{-\beta Nk(\{\sigma^\mu\})}\tag{11}$$

via Legendre transform

$$c(k)=\min_\beta\left[\beta k-\beta f(\beta)\right].\tag{12}$$

From the experience with related systems [19] one expects f (and therefore c) to be self-averaging with respect to the distribution of the input patterns $\boldsymbol{\xi}^\mu$. The average of $f(\beta)$ over the inputs can be performed using the replica trick. Within a special replica symmetric ansatz the calculation of $f(\beta)$ can be reduced to a saddle-point integral over one (for the spherical) or two (for the Ising case) order parameters, which are evaluated numerically [17].

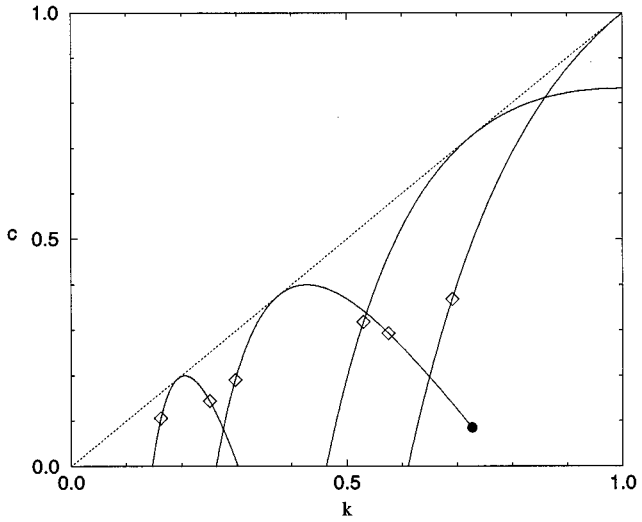

FIG. 2. Distribution of cell sizes $c(k)$ in the coupling space of an Ising perceptron with loading ratios $\alpha=0.2,0.4,0.833,1.245$ (from left to right). Inside the region given by the diamonds replica symmetry holds. The dot marks the divergences of negative moments.

Figure 2 shows some of the resulting curves for the Ising case. For $\alpha=0.2$ and 0.4 the corresponding curves for the spherical perceptron are rather similar. The typical cell size is given by $k_0=\arg\max c(k)$. Therefore $V_0=2^{-Nk_0}$ coincides with the typical phase-space volume as calculated by a standard Gardner approach [20]. On the other hand, $2^{Nc(k_0)}$ gives up to exponentially small countributions from other cell sizes the typical total number Δ^{typ} of cells as determined for the spherical perceptron by Cover [11]. From the explicit formulas one can show that for the spherical perceptron $c(k_0)=\alpha$ as long as $k_0<\infty$, i.e., $V_0>0$, and $c(k)<\alpha$ if $k_0=\infty$, i.e. $V_0=0$.

For the Ising perceptron there is a smallest possible cell size $k_{\max}=1$ where only one coupling remains. Hence $\Delta^{\mathrm{typ}}\sim 2^{Nc(k_0)}$ if $k_0\leqslant 1$ and $\Delta^{\mathrm{typ}}\sim 2^{Nc(1)}$ if $k_0>1$. The borderline $k_0=1$ is realized for $\alpha=0.83$, the well known value of α_c [13]. The calculation of the curves $c(k)$ therefore establishes the connection between the two complementary approaches by Cover and Gardner to determine the storage capacity of neural nets.

Since one has direct access to the number of realizable output sequences it is tempting to use this approach also to calculate the VC dimension analytically. Due to the averages over the input distribution necessary to accomplish the calculation we can at most hope to determine $\alpha_{\mathrm{VC}}^{\mathrm{typ}}$ in this way. As discussed above, $\alpha_{\mathrm{VC}}^{\mathrm{typ}}$ will only coincide with α_{VC} if the maximum in Eq. (1) is realized by a typical set of input patterns. To determine $\alpha_{\mathrm{VC}}^{\mathrm{typ}}$ we have to find the value of α at which the total number of cells Δ^{typ} starts to deviate from $2^{\alpha N}$. For $\alpha_{\mathrm{VC}}<\alpha<2\alpha_{\mathrm{VC}}$ an asymptotic analysis of the bound in Eq. (2) reveals that [8]

$$\frac{2^{\alpha N}-\sum_{i=0}^{\alpha_{\mathrm{VC}}N}\binom{\alpha N}{i}}{2^{\alpha N}}\to\frac{1}{2}\ \mathrm{erfc}\left[\sqrt{\frac{\alpha N}{2}}\left(\frac{2}{\alpha}-1\right)\right]. \quad (13)$$

Hence it may happen that the relative deviation is exponentially small in N. In principle, we are able to detect this deviation by using the *whole function* $c(k)$. However, for very small and very large k the calculation of $c(k)$ necessitates replica symmetry breaking [17], which renders the calculation practically impossible.

But there is another way to get some information on $\alpha_{\mathrm{VC}}^{\mathrm{typ}}$ from the $c(k)$ curves. It is clear from Eq. (11) that $f(\beta)$ will diverge for all $\beta<0$ if some of the cells are empty, i.e., if $k(\{\sigma^\mu\})=-\infty$. For $\alpha<\alpha_{\mathrm{VC}}^{\mathrm{typ}}$ this is possible only if the patterns are linearily dependent. For Gaussian patterns the probability for this to happen is zero and therefore no divergence of f for $\beta<0$ will show up for $\alpha<\alpha_{\mathrm{VC}}^{\mathrm{typ}}$. [For binary patterns the probability for two identical patterns is 2^{-N} and $f(\beta)$ should be divergent for $\beta<0$ for all α. This is, however, not found in the explicit calculation since by keeping only the first two moments of the pattern distribution in performing the ensemble average one effectively replaces the original distribution by a Gaussian one.] For $\alpha>\alpha_{\mathrm{VC}}^{\mathrm{typ}}$, however, there are *typically* some empty cells and $f(\beta)$ should be divergent for all $\beta<0$.

Within the replica symmetric approximation one finds this divergence of negative moments for both the spherical and the Ising perceptron at $\beta=(\alpha-1)/\alpha$ if $\alpha<1$ and $\beta=0^-$ if $\alpha\geqslant 1$ [17]. This suggests $\alpha_{\mathrm{VC}}^{\mathrm{typ}}=1$ for both cases. For the spherical perceptron this coincides with the known result. Moreover, the point $\beta=0,\alpha=1$ belongs to the region of local stability of the replica symmetric saddle point. For the Ising case the result must be wrong since $\alpha_{\mathrm{VC}}^{\mathrm{typ}}$ cannot be larger than $\alpha_c\approx 0.83$. Since also in this case the replica symmetric saddle point is locally stable at $\beta=0,\alpha=1$, it is very likely that there is a discontinuous transition to replica symmetry breaking as typical for this system [13]. It remains to be seen whether a solution in one step replica symmetry breaking can provide a more realistic value of $\alpha_{\mathrm{VC}}^{\mathrm{typ}}$.

In principle, it is possible, using the same techniques, to obtain expressions for the typical growth function of simple multilayer nets. However, the technical problems will increase and replica symmetry breaking is again likely to show up. We just note that a related analysis, namely, the characterization of the distribution of internal representations within the typical Gardner volume, has recently been performed [21–23] for the committee machine. From these investigations the storage capacity in the limit of a large number of hidden units could be obtained.

V. TYPICAL GROWTH FUNCTIONS

The typical growth function $\Delta^{\mathrm{typ}}(p)$ of a classifier system that is parametrized by N binary variables can be measured numerically by an algorithm that mixes Monte Carlo methods and exact enumeration [24]. The enumeration is required to determine $\Delta(\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^p)$, the number of different output vectors that are realizable for a given pattern set. To get this number, one has to calculate the output vectors of all 2^N classifiers. This exponential complexity limits the numerical calculations to small values of N.

To get $\Delta^{\mathrm{typ}}(p)$, we draw p random unbiased patterns $\boldsymbol{\xi}^\mu\in\{\pm 1\}^N$ and calculate $\Delta(\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^p)$. This is repeated again and again and the values of $\Delta(\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^p)$ are aver-

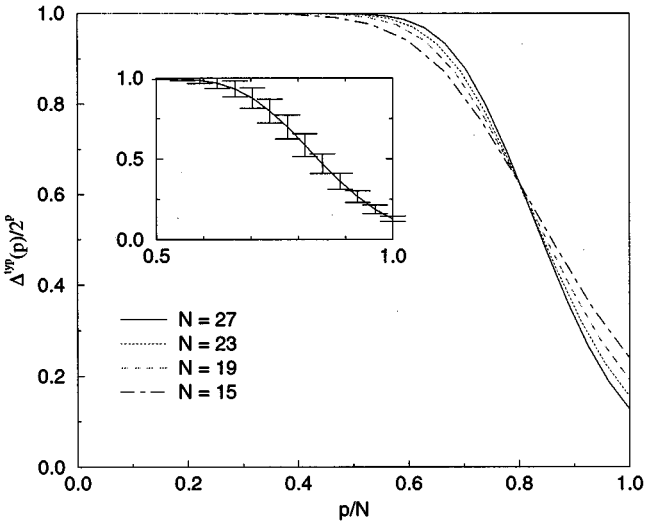

FIG. 3. Typical growth function of the Ising perceptron with binary patterns averaged over 1000 samples. The function is of course only defined at discrete values of p, but the continuous lines ease the readability. The inset displays the values for $N=27$ together with the error bars.

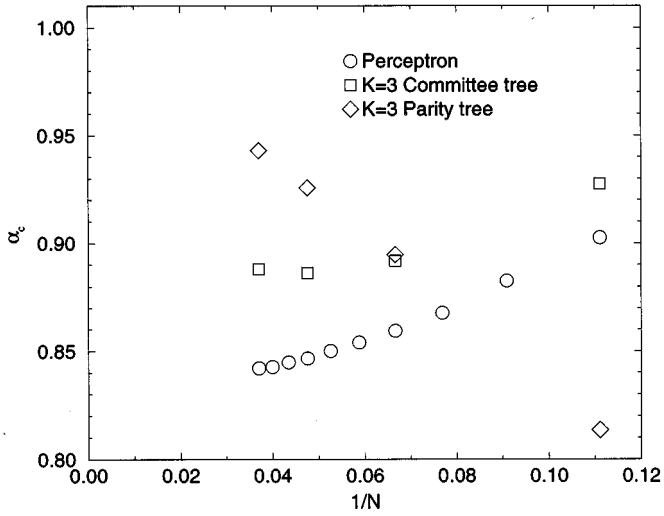

FIG. 4. Critical storage capacity α_c deduced from $\Delta^{typ}(p)$ for the Ising perceptron, the $K=3$ Ising committee tree, and the $K=3$ Ising parity tree.

aged to yield $\Delta^{typ}(p)$. The scale of Δ for large N is $O(2^N)$, so we average the logarithm

$$\ln\Delta^{typ}(p) = \langle \ln[\Delta(\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p)]\rangle_{\{\mu\}}. \tag{14}$$

Figure 3 shows $\Delta^{typ}(p)$ for the Ising perceptron with binary patterns. The curves display the expected behavior: $\Delta^{typ}(p)=2^p$ for small p and $\Delta^{typ}(p)\ll 2^p$ for larger values of p. The transition between these to regimes seems to become sharper with increasing N, but it is not clear whether we get a true step function in the limit $N\to\infty$. The corresponding curves for the committee and the parity tree look similar.

As a test we derive the critical storage capacity α_c from Fig. 3 by reading off the point where $\Delta^{typ}(p)=2^{p-1}$. Figure 4 shows α_c vs $1/N$ for the Ising perceptron and the committee and parity tree with $K=3$ each. The extrapolations to $N=\infty$ are in good agreement with the analytical results $\alpha_c=0.83$ for the Ising perceptron [13], $\alpha_c=0.92$ for the Ising committee tree with $K=3$ [25], and $\alpha_c=1$ for the Ising parity tree with $K\geqslant 2$ [26].

For the spherical perceptron $\Delta(\{\boldsymbol{\xi}^\mu\})$ is known to be the same for all pattern sets in general position. The inset of Fig. 3 displays that in the case of the Ising perceptron the average over the patterns introduces a statistical error that does *not* tend to zero with increasing number of samples. This implies that for the Ising perceptron the number of realizable output sequences is not the same for all pattern sets in general position.

The typical VC dimension d_{VC}^{typ} can principally be obtained from $\Delta^{typ}(p)$ as the number of patterns for which $\Delta^{typ}(p)$ starts to deviate from 2^p. Due to the statistical errors in $\Delta^{typ}(p)$, a separate evaluation of d_{VC}^{typ} is more appropriate. For this, we calculate $\Delta(\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p)$ for a random set of patterns. If equal to 2^p, the set is enlarged by another random pattern and $\Delta(\{\boldsymbol{\xi}^\mu\})$ is calculated again. This step is repeated until the set is no longer shattered. The number of patterns in the set (-1) gives a value for d_{VC}^{typ}. These values are averaged over many random samples. The results are shown in Fig. 5. The dependence of d_{VC}^{typ} on N is roughly given by

$$d_{\mathrm{VC}}^{typ}(N) \propto \begin{cases} 0.5N & \text{(Ising perceptron)} \\ 0.6N & \text{(committee tree)} \\ 0.88N & \text{(parity tree)}. \end{cases} \tag{15}$$

VI. BOUNDS FOR d_{VC}

The exact value of d_{VC} for the Ising perceptron with binary or zero threshold and binary patterns is not known, not even in the limit $N\to\infty$. Only bounds can be provided.

An arbitrary set of classifiers that are parametrized by N

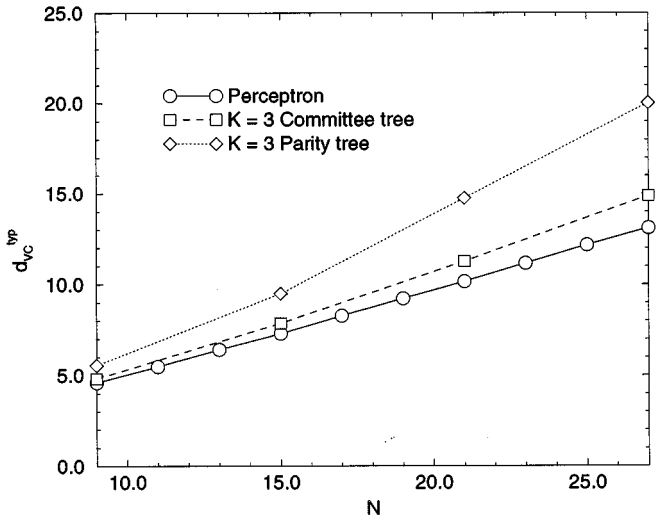

FIG. 5. Numerical values of $d_{\mathrm{VC}}^{typ}(N)$ for the Ising perceptron, the $K=3$ Ising committee tree, and the $K=3$ Ising parity tree. The straight lines between the points are guides to the eye.

bits, like the Ising perceptron, cannot produce more than 2^N distinct output vectors on any set of input patterns. So we have

$$d_{\mathrm{VC}}(N) \leqslant N \tag{16}$$

as a general upper bound for ''Ising-like'' classifiers.

Finding good lower bounds is a bit more tedious. It can be achieved by explicit construction of shattered sets. In those cases, however, where $d_{\mathrm{VC}}^{\mathrm{typ}} \ll d_{\mathrm{VC}}$, shattered sets with cardinality greater than $d_{\mathrm{VC}}^{\mathrm{typ}}$ are rare and consequently hard to find by random search.

A. Ising perceptron

For the Ising perceptron it is shown in Appendix A that d_{VC} is the same for N odd, zero threshold and $N-1$, binary threshold. Therefore we can safely restrict ourselves to the case N odd and no threshold.

In Ref. [27] a special pattern set is given that yields

$$d_{\mathrm{VC}}(N) \geqslant \tfrac{1}{2}(N+3). \tag{17}$$

Shattered sets with cardinality $\frac{1}{2}(N+3)$ are not too rare; they do show up in the statistical algorithm of Sec. V. To get an improved lower bound for general values of N, we consider a restricted variant of the Ising perceptron, the *balanced* Ising perceptron where the couplings have minimum ''magnetization'':

$$\sum_i J_i = \pm 1. \tag{18}$$

The balanced Ising perceptrons are a subset of the usual Ising perceptrons, hence any pattern set that is shattered by the former is as well shattered by the latter.

Now let $\{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p\}$ be a shattered set for the balanced Ising perceptron with N nodes and let $\{\sigma_1, \ldots, \sigma_p\}$ be an output vector that is realized by the balanced weight vector $\mathbf{J}$. Going from N to $N+2$, we define $p+1$ patterns

$$\tilde{\boldsymbol{\xi}}^{\nu} = (-, \boldsymbol{\xi}^{\nu}, -), \quad 1 \leqslant \nu \leqslant p \tag{19a}$$

$$\tilde{\boldsymbol{\xi}}^{p+1} = (-, \underbrace{-, \ldots, -}_{N \text{ times}}, +) \tag{19b}$$

and new couplings

$$\mathbf{J}^{\pm} = (+, \mathbf{J}, -), \quad \mathbf{J}^{\mp} = (-, \mathbf{J}, +). \tag{20}$$

These couplings preserve the output values of the ''old'' patterns

$$\mathrm{sgn}(\mathbf{J}^{\pm} \tilde{\boldsymbol{\xi}}^{\nu}) = \mathrm{sgn}(\mathbf{J}^{\mp} \tilde{\boldsymbol{\xi}}^{\nu}) = \sigma_{\nu}, \quad 1 \leqslant \nu \leqslant p, \tag{21}$$

while the balance property ensures that both classifications of the new pattern can be realized:

$$\mathrm{sgn}(\mathbf{J}^{\pm} \tilde{\boldsymbol{\xi}}^{p+1}) = \mathrm{sgn}\left(-2 - \sum_{i=1}^{N} \mathbf{J}_i \right) = -1, \tag{22a}$$

$$\mathrm{sgn}(\mathbf{J}^{\mp} \tilde{\boldsymbol{\xi}}^{p+1}) = \mathrm{sgn}\left(2 - \sum_{i=1}^{N} \mathbf{J}_i \right) = +1. \tag{22b}$$

Note that both $\mathbf{J}^{\pm}$ and $\mathbf{J}^{\mp}$ are balanced. This allows us to apply Eqs. (19) and (20) recursively to obtain the lower bound

$$d_{\mathrm{VC}}(N) \geqslant \tfrac{1}{2}(N + 2c - N_0), \quad N \geqslant N_0 \tag{23}$$

for the *general* Ising perceptron, where c is given by the cardinality of a shattered set for the *balanced* Ising perceptron with N_0 nodes.

Now we are left with the problem of finding large shattered sets for the balanced Ising perceptron. A partial enumeration (see below) yields shattered sets with cardinality $c = 7, 11, 13$ for $N = 9, 15, 17$. This gives

$$d_{\mathrm{VC}}(N) \geqslant \begin{cases} \frac{1}{2}(N+5) & \text{for } N \geqslant 9 \\ \frac{1}{2}(N+7) & \text{for } N \geqslant 15 \\ \frac{1}{2}(N+9) & \text{for } N \geqslant 17. \end{cases} \tag{24}$$

The corresponding shattered sets are listed in Appendix C. This sequence of increasing lower bounds indicates that probably $\lim_{N\to\infty} d_{\mathrm{VC}}(N)/N > \frac{1}{2}$.

There is a method that surely finds the largest possible shattered set, i.e., the exact value of d_{VC}: *exhaustive enumeration* of all shattered sets. The overwhelming complexity of $O(2^{N^2})$ limits this approach to small values of N, however. Nevertheless, the results obtained for $N \leqslant 9$ are already quite remarkable [27]:

$$d_{\mathrm{VC}}(3) = 3, \quad d_{\mathrm{VC}}(5) = 4, \quad d_{\mathrm{VC}}(7) = 7, \quad d_{\mathrm{VC}}(9) = 7. \tag{25}$$

Again the corresponding shattered sets are listed in Appendix C. They share a common feature: Using transformations that do not change $\Delta(p)$ (see Appendix C), they can be transformed into quasiorthogonal pattern sets, i.e., sets where the patterns have minimum pairwise overlap

$$\boldsymbol{\xi}^{(\mu)} \cdot \boldsymbol{\xi}^{(\nu)} = \begin{cases} \pm 1, & \mu \neq \nu \\ N, & \mu = \nu. \end{cases} \tag{26}$$

(Exact orthogonality cannot be achieved for N odd.)

This observation appears reasonable. Consider a shattered set of patterns. The corresponding cells in weight space have nonzero volume $V(\{\sigma_{\mu}\})$, i.e., each cell contains at least one weight vector $\mathbf{J}$. If we enlarge the shattered set by an additional pattern, each cell must divide in two cells of nonzero volume. This process can be repeated until the first nondivisible cell appears. If we assume that the divisibility of a cell decreases with its volume, we must look for cell structures where the volume of the smallest cell is maximized. This is the case for *equisized* cells, i.e., for orthogonal patterns (Fig. 6).

Quasiorthogonal pattern sets can easily be built from the rows of Hadamard matrices (see Appendix B). These are $4n \times 4n$ orthogonal matrices with ± 1 entries. To get quasiorthogonal patterns of odd length N, we either cut out one column $(N = 4n-1)$ or add an arbitrary column

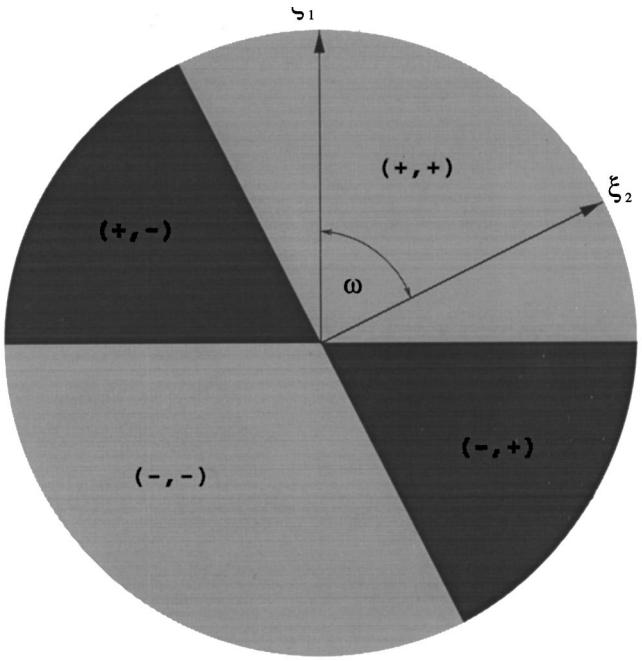

FIG. 6. Spherical perceptron with $N=2$. The four cells in weight space induced by patterns $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are equisized for $\omega=\pi/2$, i.e., for orthogonal patterns.

$(N=4n+1)$. It is clear that there are many quasiorthogonal pattern sets with p elements that can be constructed from a given Hadamard matrix. By *partial enumeration*, i.e., by evaluation of some of them, we were able to find shattered sets that exceed the lower bound given by Eq. (24) for certain values of N:

$$d_{\mathrm{VC}}(N)\geqslant\begin{cases}13, & N=15\\ 17, & N=23\\ 19, & N=27\\ 24, & N=24.\end{cases} \tag{27}$$

The corresponding pattern sets are listed in Appendix C. Systems with $N>31$ were not investigated. Note that the lower bound for $d_{\mathrm{VC}}(N=31)$ is larger than the value reported in Ref. [27].

Figure 7 summarizes our results for the $N\leqslant 31$. Both the exact values and the lower bounds provided by Eqs. (24) and (24) clearly exceed the maximum value $d_{\mathrm{VC}}=\frac{1}{2}(N+3)$ found by the statistical method in Sec. V. The somewhat irregular behavior of the lower bounds does not rule out a more regular sequel of the true $d_{\mathrm{VC}}(N)$, including well-defined asymptotics. However, if the limit $\lim_{N\to\infty} d_{\mathrm{VC}}/N$ exists, it will probably be larger than 0.5.

B. Committee tree

To get a lower bound for $d_{\mathrm{VC}}^{\mathrm{CT}}$, the VC dimension of the Ising committee tree with binary patterns, we explicitly construct a shattered set based on shattered sets for the Ising perceptron. Let $\{\boldsymbol{\tau}^{\nu}\}$ be a shattered set of an Ising perceptron with K nodes and $\{\boldsymbol{\xi}^{\mu}\}$ be a shattered set of the N/K-node subperceptron. Then we build patterns $\boldsymbol{\Xi}^{\mu,\nu}$ for the committee tree by concatenation

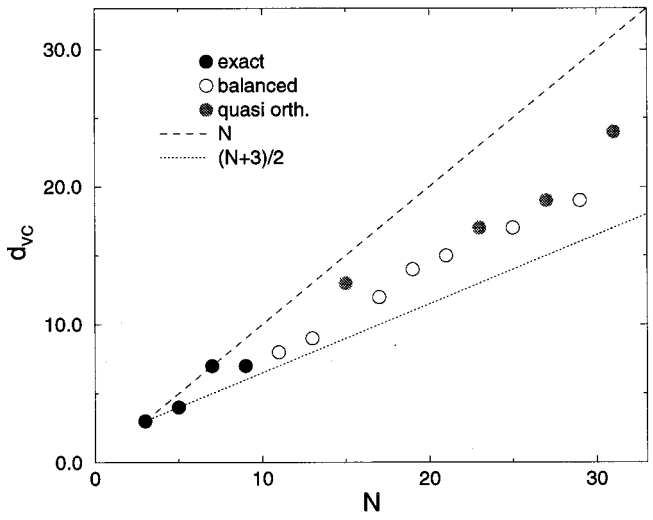

FIG. 7. VC dimension of the Ising perceptron with binary patterns vs N. The circles labeled quasiorthogonal and balanced are lower bounds for the true d_{VC}.

$$\boldsymbol{\Xi}^{\mu,\nu}=(\tau_1^{\nu}\boldsymbol{\xi}^{\mu},\tau_2^{\nu}\boldsymbol{\xi}^{\mu},\ldots,\tau_k^{\nu}\boldsymbol{\xi}^{\mu}). \tag{28}$$

We prove that the set $\{\boldsymbol{\Xi}^{\mu,\nu},1\leqslant\mu\leqslant d_{\mathrm{VC}}(N/K),1\leqslant\nu\leqslant d_{\mathrm{VC}}(K)\}$ is shattered.

Let $\{\sigma_{\mu,\nu}\}$ be a given output sequence of length $d_{\mathrm{VC}}(K)d_{\mathrm{VC}}(N/K)$. Since the $\{\boldsymbol{\tau}^{\nu}\}$ are shattered, we can always find $\{W_k^{\mu}=\pm 1\}_{k=1}^{K}$ such that

$$\sigma_{\mu,\nu}=\mathrm{sgn}\left(\sum_{k=1}^{K} W_k^{\mu}\tau_k^{\nu}\right) \tag{29}$$

for all ν and μ. Now we choose the couplings in the kth subperceptron such that

$$W_k^{\mu}=\mathrm{sgn}\left(\sum_{i=1}^{N/K} J_i^{(k)}\xi_i^{\mu}\right),\quad k=1,\ldots,K. \tag{30}$$

This is always possible because $\boldsymbol{\xi}^{\mu}$ is taken from a shattered set. Combining Eqs. (29) and (30), we get

$$\sigma_{\mu,\nu}=\mathrm{sgn}\left(\sum_{k=1}^{K}\mathrm{sgn}\sum_{i=1}^{N/K} J_i^{(k)}\tau_k^{\nu}\xi_i^{\mu}\right),\quad \nu=1,\ldots,\quad d_{\mathrm{VC}}(K). \tag{31}$$

i.e., the patterns (28) form a shattered set and we find

$$d_{\mathrm{VC}}^{\mathrm{CT}}(N)\geqslant d_{\mathrm{VC}}(K)d_{\mathrm{VC}}(N/K). \tag{32}$$

Note that this lower bound matches the upper bound N whenever $d_{\mathrm{VC}}(K)$ $[d_{\mathrm{VC}}(N/K)]$ meet their upper bounds K $[N/K]$. Examples include $K=3$ or 7 and $N=21$, $K=7$, and $N=49$.

This lower bound is much larger than the values for $d_{\mathrm{VC}}^{\mathrm{typ}}$ found in Sec. V. If we assume that $\alpha_{\mathrm{VC}} = \lim_{N\to\infty} d_{\mathrm{VC}}(N)/N$ is well defined for the Ising perceptron, Eq. (32) reads

$$d_{\mathrm{VC}}^{\mathrm{CT}}(N) \geqslant N\alpha_{\mathrm{VC}}^2, \quad N \gg K \gg 1. \tag{33}$$

C. Parity tree

We follow the same strategy and construct a shattered set from the patterns of a shattered set $\{\boldsymbol{\xi}^\nu\}$ of the subperceptrons. The first pattern is simply built from K consecutive patterns $\boldsymbol{\xi}^1$,

$$\boldsymbol{\Xi}^0 = (\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^1). \tag{34}$$

All other patterns differ from $\boldsymbol{\Xi}^0$ in only one subpattern

$$\boldsymbol{\Xi}^{\mu,k} = (\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^1,)\boldsymbol{\xi}^\mu, \boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^1)$$
$$\uparrow k\text{th position}, \tag{35}$$

where $k = 1, \ldots, K$ and $\mu = 2, \ldots, d_{\mathrm{VC}}(N/K)$. This set of $k[d_{\mathrm{VC}}(N/K) - 1] + 1$ patterns is shattered.

Proof. Let $\{\sigma_0, \sigma_{k,\mu}\}$ be a given output sequence for our patterns. We choose the weights $\mathbf{J}^{(k)}$ in the subperceptrons such that

$$\begin{aligned} \mathrm{sgn}(\mathbf{J}^{(1)} \cdot \boldsymbol{\xi}^1) &= \sigma_0, \\ \mathrm{sgn}(\mathbf{J}^{(k>1)} \cdot \boldsymbol{\xi}^1) &= 1, \\ \mathrm{sgn}(\mathbf{J}^{(1)} \cdot \boldsymbol{\xi}^\mu) &= \sigma_{1,\mu}, \\ \mathrm{sgn}(\mathbf{J}^{(k>1)} \cdot \boldsymbol{\xi}^\mu) &= \sigma_0 \sigma_{k,\mu} \end{aligned} \tag{36}$$

for $\mu = 2, \ldots, d_{\mathrm{VC}}(N/K)$. This is always possible because $\{\boldsymbol{\xi}^\nu\}$ is shattered. With this assignment of weights, the parity tree maps $\{\boldsymbol{\Xi}^0, \boldsymbol{\Xi}^{\mu,n}\}$ to the prescribed output sequence.

Our shattered set provides us with a lower bound for the VC dimension of the parity tree

$$d_{\mathrm{VC}}^{\mathrm{PT}}(N) \geqslant K[d_{\mathrm{VC}}(N/K) - 1] + 1. \tag{37}$$

For $K = 1$ the parity tree is equivalent to the simple perceptron and Eq. (37) reduces to $d_{\mathrm{VC}}^{\mathrm{PT}}(N) = d_{\mathrm{VC}}(N)$. If one inserts the lower bounds for d_{VC} into the right-hand side of Eq. (37), the resulting values are generally larger than $d_{\mathrm{VC}}^{\mathrm{PT}}$, but the differences are much smaller than for the committee tree, and for some values of N, $d_{\mathrm{VC}}^{\mathrm{typ}}$ even exceeds the right-hand side of Eq. (37). We do not know whether Eq. (37) is only a bad lower bound or whether the maximum shattered sets for the parity tree are not as atypical as for the Ising perceptron and the committee tree.

VII. CONCLUSION

The VC dimension is one of the central quantities to characterize the information processing abilities of feedforward neural networks. The determination of the VC dimension of a given network architecture is, however, in general, a nontrivial task.

In the present paper we have shown that even for the simplest feedforward neural networks this task requires rather sophisticated techniques if both the couplings of the network and the inputs are restricted to binary values ± 1. This is mainly due to the fact that the VC dimension defined by a supremum over all pattern sets of given size is determined by *atypical* pattern sets. Consequently, Monte Carlo methods as well as analytical estimates involving pattern averages do not yield reliable results and one has to resort to exact enumeration techniques. These methods are naturally restricted to small dimensions of the input space, but the results obtained can be used to get lower bounds for the VC dimension of larger systems. In some cases, even tighter bounds can be derived from number theoretic arguments.

Complementary one could argue that *typical* situations are of more interest than the worst case. Accordingly, a typical VC dimension $d_{\mathrm{VC}}^{\mathrm{typ}}$ has been defined in Sec. II. One always has $d_{\mathrm{VC}}^{\mathrm{typ}} \leqslant d_{\mathrm{VC}}$ since an average can never be larger than the supremum.

For the Ising perceptron ($J_i = \pm 1$) we found $d_{\mathrm{VC}} = N$ as long as the patterns are allowed to take on the value 0, regardless of whether we use real-valued or $\{0,1\}$ patterns. If, however, also the patterns are Ising-like, i.e., $\boldsymbol{\xi} \in \{\pm 1\}^N$ our numerical results suggest

$$\tfrac{1}{2}(N+3) < d_{\mathrm{VC}}(N) < N \tag{38}$$

for general N. For large N, the VC dimension is presumably *substantially* larger than the typical VC dimension $d_{\mathrm{VC}}^{\mathrm{typ}}(N) \propto N/2$.

Similar results are found for two simple examples of multilayer networks: the committee and the parity tree with Ising couplings. Here the results are

$$d_{\mathrm{VC}}(K) d_{\mathrm{VC}}(N/K) \leqslant d_{\mathrm{VC}}^{\mathrm{CT}}(N) \leqslant N \tag{39}$$

for the committee tree and

$$K[d_{\mathrm{VC}}(N/K) - 1] + 1 \leqslant d_{\mathrm{VC}}^{\mathrm{PT}}(N) \leqslant N \tag{40}$$

for the parity tree with K hidden nodes. For the committee tree we find again that $d_{\mathrm{VC}}^{\mathrm{typ}} < d_{\mathrm{VC}}$. For the parity tree our data do not allow us to draw the same conclusion, but this may be due to the low quality of the lower bound in Eq. (40).

We finally note that the growth function $\Delta(p)$ related to the VC dimension is used to derive the famous Vapnik-Chervonenkis bound for the asymptotic difference between learning and generalization error. This bounds results from the analysis of the worst case. It would be interesting to investigate whether a similar bound for the *typical* generalization behavior could be obtained from $\Delta^{\mathrm{typ}}(p)$, which, in general, is much easier to determine.

APPENDIX A: SYMMETRIES

Let $\{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p\}$ be a set of binary ± 1 patterns and $\Delta(\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p)$ the number of different output sequences $(\sigma_1, \ldots, \sigma_p)$ that can be realized by the Ising perceptron for this particular set of patterns. $\Delta(\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p)$ is invariant under the following transformations on $\{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p\}$: complement a whole pattern, $\boldsymbol{\xi}^\mu \mapsto -\boldsymbol{\xi}^\mu$; interchange two patterns, $\boldsymbol{\xi}^\mu \leftrightarrow \boldsymbol{\xi}^\nu$; complement one entry in all patterns

$(\mu=1,\ldots,\ p)$, $\xi_i^\mu \mapsto -\xi_i^\mu$ and interchange two entries in all patterns $(\mu=1,\ldots,p)$, $\xi_i^\mu \leftrightarrow \xi_j^\mu$. Applying these transformations, we can always achieve that all patterns have $\xi_N^\mu=-1$.

Now we assume that N, the number of couplings, is odd and that there is no threshold. Let $\mathbf{J}$ be a weight vector that realizes an output sequence $(\sigma_1,\ldots,\sigma_p)$ for a pattern with $\xi_N^\mu=-1$:

$$\sigma_\nu=\mathrm{sgn}\left(\sum_{i=1}^{N-1} J_i\xi_i^\nu - J_N\right), \quad 1\leqslant\nu\leqslant p. \tag{A1}$$

We use the left $N-1$ bits of $\boldsymbol{\xi}^\nu$ as a pattern set for the Ising perceptron with $N-1$ input units and a binary threshold Θ. Identifying Θ with J_N in Eq. (A1), it becomes obvious that $\Delta(\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^p)$ is the same in both cases. Hence we may restrict ourselves to the case N odd and no threshold to discuss the VC dimension of the Ising perceptron.

Now we consider a two-layer feedforward network with K perceptrons (spherical or Ising) operating between input and hidden layer (weight vectors $\mathbf{J}^{(k)}$ and an Ising perceptron as decoder function with weight vector $\mathbf{J}^{(0)}$). Suppose that a given output sequence is realized by a weight vector with some entries $J_k^0=-1$ in the decoder perceptron. The output sequence is left unchanged if we set $J_k^0=+1$ and at the same time complement all weights in the kth subperceptron $\mathbf{J}^{(k)}\mapsto -\mathbf{J}^{(k)}$. This transformation allows us to realize any realizable output sequence with all $J_k^0=+1$. Hence the VC dimension of the committee machine equals the VC dimension of the two-layer perceptron with Ising weights in the output layer.

APPENDIX B: HADAMARD MATRICES

A Hadamard matrix is an $m\times m$ matrix H with ±1 entries such that

$$HH^T=mI, \tag{B1}$$

where I is the $m\times m$ identity matrix. If H is an $m\times m$ Hadamard matrix, then $m=1$, $m=2$, or $m\equiv 0 \mod 4$. The reversal is a famous open question: Is there a Hadamard matrix of order $m=4n$ for every positive n? The first open case is $m=428$.

If H and H' are Hadamard matrices of order m and m', respectively, their Kronecker product $H\otimes H'$ is a Hadamard matrix of order mm'. Starting with the 2×2 Hadamard matrix

$$H_2=\begin{pmatrix} -1 & -1 \\ -1 & +1 \end{pmatrix}, \tag{B2}$$

this gives Hadamard matrices of order $4,8,16,\ldots,2^n$, the so-called *Sylvester-type* matrices. For example,

$$H_{2^3}=\begin{pmatrix} -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & +1 & -1 & +1 & -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 & -1 & -1 & +1 & +1 \\ -1 & +1 & +1 & -1 & -1 & +1 & +1 & -1 \\ -1 & -1 & -1 & -1 & +1 & +1 & +1 & +1 \\ -1 & +1 & -1 & +1 & +1 & -1 & +1 & -1 \\ -1 & -1 & +1 & +1 & +1 & +1 & -1 & -1 \\ -1 & +1 & +1 & -1 & +1 & -1 & -1 & +1 \end{pmatrix}. \tag{B3}$$

Let q be an odd prime power. Then Hadamard matrices of *Paley type* can be constructed for

$$m=\begin{cases} q+1 & \text{for } q\equiv 3 \mod 4 \\ 2(q+1) & \text{for } q\equiv 1 \mod 4. \end{cases} \tag{B4}$$

Paley's construction [28] relies on the properties of finite Galois fields GF(q) [29], where q is an odd prime power, especially on the *quadratic character* χ of GF(q), defined by

$$\chi(x)=\begin{cases} 0 & \text{if } x=0 \\ +1 & \text{if } x\neq 0 \text{ is a square} \\ -1 & \text{otherwise.} \end{cases} \tag{B5}$$

Then, for any $a\neq 0$

$$\sum_{x\in \mathrm{GF}(q)} \chi(x)\chi(x-a)=-1. \tag{B6}$$

To construct a Paley-type matrix for $q\equiv 3 \mod 4$, we start with the $q\times q$ matrix $M=(m_{ij})$ whose rows and columns are indexed by the elements of GF(q):

$$m_{ij}=\begin{cases} -1 & \text{if } i=j \\ \chi(i-j) & \text{if } i\neq j. \end{cases} \tag{B7}$$

Hence, by Eq. (B6)

$$\sum_{j\in \mathrm{GF}(q)} m_{hj}m_{ij}=\begin{cases} q, & h=i, \\ -1, & h\neq i. \end{cases} \tag{B8}$$

Now adjoin one row and one column with all entries $+1$ to get a Hadamard matrix of order $q+1$. This gives Hadamard matrices of order 4,8,12,20,24,28, ...

For example, $q=11$. The Galois field GF(11) is equivalent to the integers $\{0,\ldots,10\}$ together with their addition and multiplication modulo 11. The squares are 1,4,9,5,3 and we get

$$H_{11+1}=\begin{pmatrix} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 & +1 & +1 & +1 & -1 & -1 & -1 & +1 & -1 \\ +1 & -1 & -1 & +1 & -1 & +1 & +1 & +1 & -1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 & +1 & -1 & +1 & +1 & +1 & -1 & -1 & -1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 & +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 & +1 & +1 & -1 \\ +1 & -1 & -1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 & +1 & +1 \\ +1 & +1 & -1 & -1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 & +1 \\ +1 & +1 & +1 & -1 & -1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & +1 & +1 & +1 & -1 & -1 & -1 & +1 & -1 & -1 & +1 & -1 \\ +1 & -1 & +1 & +1 & +1 & -1 & -1 & -1 & +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & +1 & +1 & +1 & -1 & -1 & -1 & +1 & -1 & -1 \end{pmatrix}. \tag{B9}$$

For $q\equiv 1 \bmod 4$, the construction starts with the $(q+1)\times(q+1)$ matrix $M=(m_{ij})$, indexed by $\mathrm{GF}(q)\cup\{\infty\}$ as

$$m_{\infty j}=m_{j\infty}=1 \quad \text{for all } j\in \mathrm{GF}(q), \tag{B10}$$

$$m_{\infty\infty}=0, \tag{B11}$$

$$m_{ij}=\chi(j-i) \quad \text{for } i,j\in \mathrm{GF}(q). \tag{B12}$$

M is symmetric and orthogonal. To get from M to a Hadamard matrix of order $2(q+1)$, we define the auxiliary matrices A and B by

$$A=\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad B=\begin{pmatrix} 1 & -1 \\ -1 & -1 \end{pmatrix} \tag{B13}$$

and replace every 0 in M by B, every $+1$ by A, and every -1 by $-A$. This gives Hadamard matrices of order 12,20,28,36,52, ... For example, $q=5$. GF(5) is equivalent to the integers $\{0,\ldots,4\}$ and their addition and multiplication modulo 5. The squares are 1 and 4 and we get

$$H_{2(5+1)}=\begin{pmatrix} B & A & A & A & A & A \\ A & B & A & -A & -A & A \\ A & A & B & A & -A & -A \\ A & -A & A & B & A & -A \\ A & -A & -A & A & B & A \\ A & A & -A & -A & A & B \end{pmatrix}. \tag{B14}$$

The first value of $m=4n$ where neither the Sylvester nor the Paley construction applies is $m=92$.

APPENDIX C: GALLERY OF SHATTERED SETS

For $N\leqslant 9$ the exact values of d_{VC} have been obtained by exhaustive enumerations. Shattered sets of maximum cardinality are

$N=3$	$N=5$	$N=7$	$N=9$
+++	+++++	−−−−−−−	−−−−−−−−−
−−+	−+−−+	−+−+−+−	+−+−+−+−+
−+−	−+−+−	−−++−−+	−−−+−−++
	−++−−	−++−−++	+−++−−++−
		−−−−+++	−−−−−++++
		−+−++−+	+−+−++−+−
		−−+++−−	−−−+++−−

The sets for $N=3$ and $N=5$ are obtained from the rows of the Sylvester-type Hadamard matrix H_{2^2}. For $N=3$, the first column and the last row has been deleted. For $N=5$, a column $(+1,-1,-1,-1)$ has been adjoined. The sets for $N=7$ and $N=9$ are obtained from the rows of the Sylvester-type Hadamard matrix H_{2^3}; confer Eq. (B3). For $N=7$, the eighth column and row have been deleted, and for $N=9$, a column with alternating ± 1's has been adjoined.

The largest shattered sets we could find for the balanced Ising perceptron with binary patterns are

$N=9$	$N=15$	$N=17$
−−−−−−−−−	−+−+−+−+−+−+−	−−+−+−+−+−+−+−+
−−−−−−+++	−−++−−++−−++−−+	+−−++−−++−−++−−++
−−−−+−−+	−++−−++−−++−−++	−−++−−++−−++−−+−
−−+−−−−+	−−−−++++−−−−+++	+−−−−++++−−−−++++
−+−+−+−+−	−+−++−+−−+−++−+	−−+−++−+−−+−++−+−
−+−++−+−−	−−++++−−−−++++−	+−−++++−−−−++++−−
−++−−++−−	−−−−−−−−++++++	−−++−+−−+−++−+−−+
	−+−+−+−++−+−+−+	+−−−−−−−−++++++++
	−−++−−++++−−++−	−−+−+−+−++−+−+−+−
	−+−++−+−+−+−−+−	+−−++−−++++−−++−−
	−−++++−−+−−−−+	−−++−−++−+−−++−−+
		+−−−−++++++++−−−−
		−−+−++−+−+−+−−+−+

These pattern sets lead to the lower bounds in Eq. (24). The set for $N=9$ has been found by exhaustive enumeration and has no simple relation to a Hadamard matrix. The patterns for $N=15$ are rows 2–11, 14, and 15 of the Sylvester-type Hadamard matrix H_{2^4} with the last column deleted. The patterns for $N=17$ are rows 2–14 of H_{2^4}, extended by a column of alternating ± 1's.

Pattern sets that exceed the bounds given in Eq. (24) can be constructed for these values of N: $N=15$, delete the last column from the Sylvester-type Hadamard matrix H_{2^4} and then the first 13 rows form a shattered pattern set; $N=23$, delete the last column from the Hadamard matrix $H_2 \otimes H_{11+1}$ and then the first 17 rows form a shattered pattern set; $N=27$, delete the last column from the Paley-type Hadamard matrix $H_{2(13+1)}$ and then the first 19 rows form a shattered pattern set; and $N=31$, delete the last column from the Sylvester-type Hadamard matrix H_{2^5} and then the rows number 2 to number 25 form a shattered pattern set with 24 patterns.

[1] V. N. Vapnik and A. Y. Chervonenkis, Theory Probab. Its Appl. (USSR) **16**, 264 (1971).
[2] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data* (Springer, Berlin, 1982).
[3] D. Haussler, M. Kearns, and R. Schapire, in *The IVth Annual Workshop on Computational Learning Theory (COLT 91), Santa Cruz, 1991* (Morgan Kaufmann, San Mateo, CA, 1991).
[4] J. M. R. Parrondo and C. van den Broeck, J. Phys. A **26**, 2211 (1993).
[5] A. Engel, Mod. Phys. Lett. B **8**, 1683 (1994).
[6] E. Baum and D. Haussler, Neural Comput. **1**, 151 (1989).
[7] A. Engel and C. van den Broeck, Phys. Rev. Lett. **71**, 1772 (1993).
[8] J. A. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).
[9] N. Sauer, J. Combinatorial Theory A **13**, 145 (1972).
[10] M. Opper, Phys. Rev. E **51**, 3613 (1995).
[11] T. M. Cover, IEEE Trans. Electron. Comput. **EC-14**, 326 (1965).
[12] H. Gutfreund and D. Stein, J. Phys. A **23**, 2613 (1990).
[13] W. Krauth and M. Mézard, J. Phys. (Paris) **50**, 3057 (1989).
[14] W. Krauth and M. Opper, J. Phys. A **22**, L519 (1989).
[15] B. Derrida, R. B. Griffith, and A. Prügel-Bennett, J. Phys. A **24**, 4907 (1991).
[16] R. Monasson and D. O'Kane, Europhys. Lett. **27**, 85 (1994).
[17] A. Engel and M. Weigt, Phys. Rev. E **53**, R2064 (1996).
[18] M. Weigt and A. Engel, Phys. Rev. E (to be published).
[19] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[20] E. Gardner, J. Phys. A **21**, 257 (1988).
[21] R. Monasson and R. Zecchina, Phys. Rev. Lett. **75**, 2432 (1995).
[22] R. Monasson and R. Zecchina, Mod. Phys. Lett. B **9**, 1887 (1996).
[23] S. Cocco, R. Monasson, and R. Zecchina, Phys. Rev. E **54**, 717 (1996).
[24] J. Stambke, Diploma thesis, University of Giessen, 1992 (unpublished).
[25] E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992).
[26] E. Barkai and I. Kanter, Europhys. Lett. **14**, 107 (1990).
[27] S. Mertens, J. Phys. A **29**, L199 (1996).
[28] T. Beth, D. Jungnickel, and H. Lenz, *Design Theory* (Bibliographisches Institut, Mannheim, 1985), Chap. I.9.
[29] Rudolf Lidl and Harald Niederreiter, *Introduction to Finite Fields and their Applications* (Cambridge University Press, Cambridge, 1994).