Multifractal *a priori* probability distribution for the perceptron

C. Van den Broeck and G. J. Bex
Limburgs Universitair Centrum, B-3590 Diepenbeek, Belgium

(Received 1 October 1997)

We calculate the multifractal spectrum of the *a priori* probability distribution for a perceptron. [S1063-651X(98)14902-4]

PACS number(s): 87.10.+e, 05.50.+q, 02.50.Cw, 64.60.Ak

In an effort to understand the ability of artificial networks to learn by example, a lot of research has been devoted recently to the study of very simple learning scenarios, in which all the details of the learning process can be calculated analytically [1–10]. In classification problems, the aim is to learn the rule that lies at the basis of a set of observed training examples. This, however, is only possible if one has some prior knowledge about the problem. In a Bayesian framework, this knowledge is quantified by the so-called *a priori* probability distribution, defined as the probability for any of the possible classification rules to be true, *a priori*. The role of the *a priori* probability distribution is particularly clear in one of the simpler theories of learning a rule by example [11,12], since it allows for the direct evaluation of the generalization error. In machines that learn from examples, such as Boolean or neural networks, the *a priori* probability distribution is dictated by the architecture of the network. Although the *a priori* probability distribution has been calculated numerically for a number of interesting cases [11,12], we are not aware of any analytic results. Furthermore, the numerical results suggest that this distribution may typically have fractal properties. In this paper we confirm this suspicion by an analytic calculation of the *a priori* probability distribution for a simple perceptron. It is found to be a monofractal for N, the number of input channels of the perceptron, going to ∞, but it becomes a multifractal if finite size corrections are included.

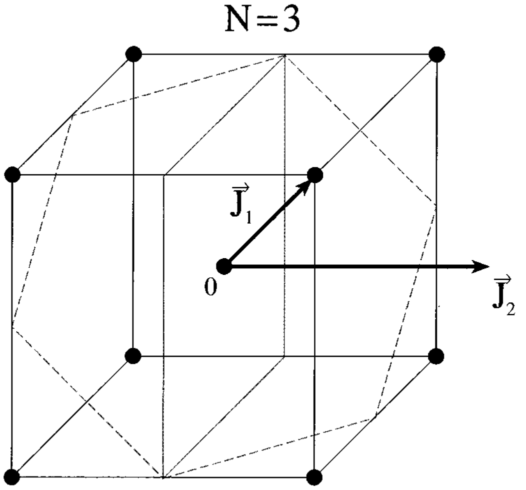

FIG. 1. Illustration of the classifications induced on the corners of the cube ($N=3$).

The perceptron [13] is a classifier characterized by an N-dimensional synaptic vector $\vec{J}$, that returns the following binary output ξ_o when presented with an input pattern $\vec{\xi}$:

$$\xi_o=\text{sgn}\left(\frac{\vec{J}\cdot\vec{\xi}}{\sqrt{N}}\right). \tag{1}$$

We start by evaluating the *a priori* probability distribution of this perceptron for the case of input patterns with binary

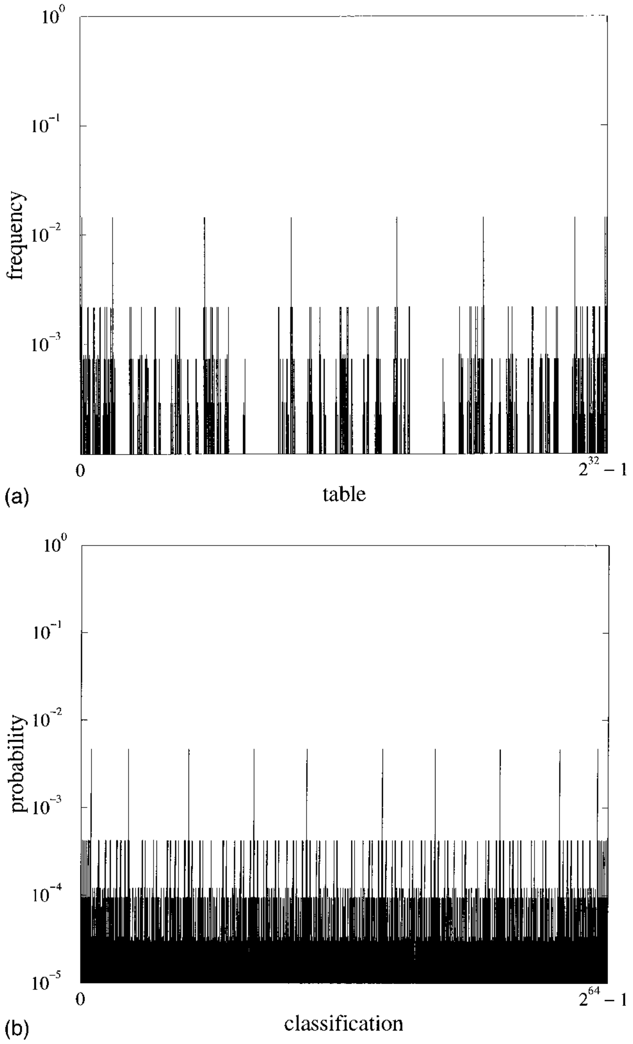

FIG. 2. *A priori* probability distribution for patterns at the corners of the hypercube in dimension N: Monte Carlo results for $N=5$ (a) and $N=6$ (b).

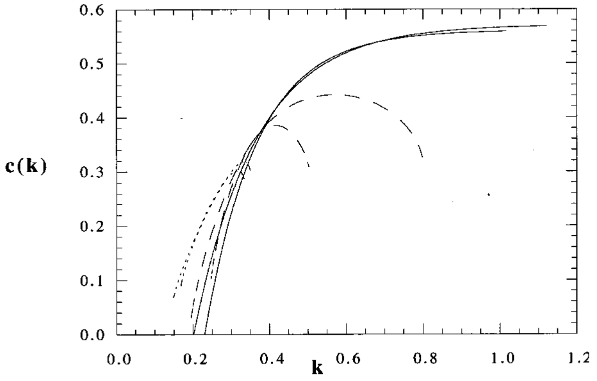

FIG. 3. Multifractal spectrum $c(k)$ as obtained from Eq. (4) combined with the numerical samplings presented in Figs. 2 and 4 [full lines: theory, cf. Eq. (9); dotted lines: patterns at the corners of the hypercube; dashed lines: random patterns; the curve with the higher maximum in each case corresponds to $N=6$, the lower maximum to $N=5$].

components $\xi_i=+1$ or -1. In total, there are 2^N such input patterns, namely, all the corners of the hypercube in dimension N. Every choice of $\vec{J}$, which are taken at random on the N sphere, $\vec{J}^2=N$, will induce a specific classification of these patterns. Apart from the well known fact that only linearly separable classifications can be implemented, we observe that the latter are, in fact, typically implemented with widely different *a priori* probabilities. This is illustrated in Fig. 1 for the case $N=3$. In this low-dimensional situation, only two types of classifications are possible, namely, those pointing more or less in the direction of the corners of the cube (there are eight of these, corresponding to the eight corners of the cube, see, e.g., $\vec{J}_1$), and those with a $\vec{J}$ vector more or less parallel to one of the axes of the cube (there are six such classifications, corresponding to the three axes, each with two directions, see, e.g., $\vec{J}_2$). However, it is clear from Fig. 1 that $\vec{J}_2$ can move around in a larger solid angle than $\vec{J}_1$ without modifying the classification of the corners of the hypercube. By an explicit calculation, one finds that the respective solid angles are equal to $\pi/2-3\arcsin(1/3)$ and $4\arcsin(1/3)$. The *a priori* probability thus consists of eight peaks of size $1/8-3\arcsin(1/3)/(4\pi)=0.0439$ and six peaks of size $\arcsin(1/3)/\pi=0.1082$. All the other $256-14=242$ classifications cannot be implemented and have probability zero. Results for higher dimension are more difficult to obtain analytically, but can be obtained numerically by generating vectors $\vec{J}$ with random orientation, and verifying which input-output table is reproduced. In Fig. 2, we have collected results for N = 5 and 6. Note that only a very small fraction of tables out of the total of 2^{2^N} can be generated by the perceptron. We find 4, 14, 104, 1882, and 94 572 different tables for N = 2, 3, 4, 5, and 6, respectively. These numbers should be compared with the result of Cover in Ref. [4], which gives the number W of linearly separable classifications for p patterns chosen at random in dimension N:

$$W=\begin{cases} 2^p, & p\leqslant N \\ 2\sum_{k=0}^{N-1}\binom{p-1}{k}, & p\geqslant N. \end{cases} \tag{2}$$

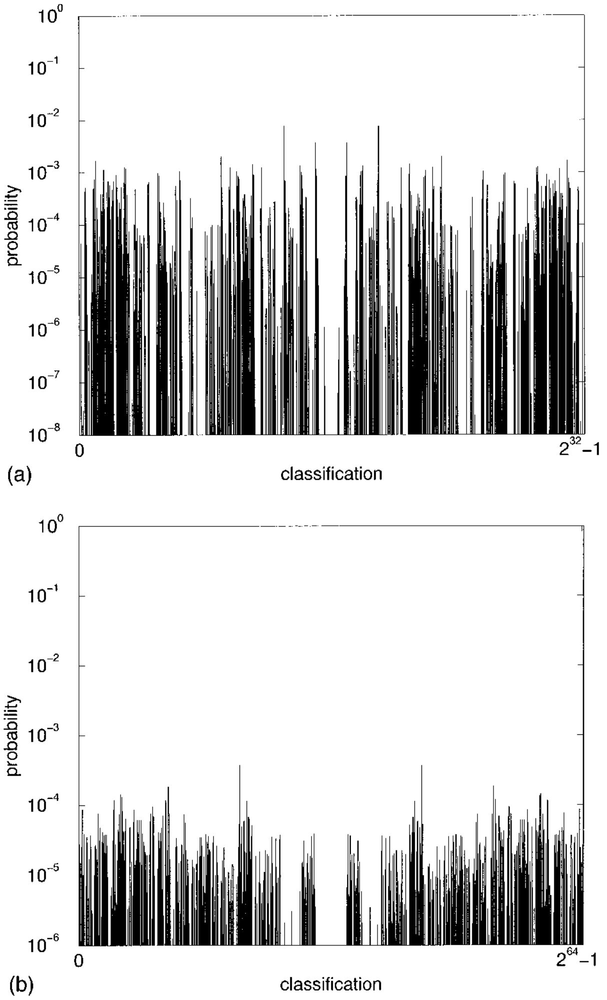

FIG. 4. *A priori* probability distribution for 2^N random patterns in dimension N: Monte Carlo results for $N=5$ (a) and $N=6$ (b).

The difference between both results is explained by the fact that the $p=2^N$ patterns that sit at the corners of the hypercube do not lie in a general configuration, so that the number of dichotomies that can be induced is much smaller.

In order to extract fractal or multifractal properties of a probability distribution such as the one represented in Fig. 2, one first needs to understand the correct scaling behavior with N. Indeed, one expects that as $N\rightarrow\infty$ the number of peaks in the probability distribution will diverge while the size of the peaks will go to zero. For patterns in a random configuration, we obtain from Eq. (2) (or directly from Sauer's bound [14], which gives the same result), that for $p=2^N$ the number of classifications that can be implemented, i.e., the number of nonzero peaks in the probability distribution, increases as

$$W\sim e^{N^2\ln 2}. \tag{3}$$

Even though we have not been able to prove that this scaling remains valid for the corners of the hypercube (which do not lie in a general configuration), we have applied it to the numerical results from Fig. 2. To find the multifractal spectrum $c(k)$ [15], we first evaluate its Legendre transform:

$$\phi(R)=\lim_{N\to\infty}\frac{\ln\sum_i P_i^R}{N^2}=\lim_{N\to\infty}\frac{\ln\int dk\ e^{N^2[c(k)-Rk]}}{N^2}$$
$$=\operatorname*{extr}_k[c(k)-Rk], \tag{4}$$

where the sum over i runs over all the classifications that have a nonzero *a priori* probability P_i. k is the singularity exponent of the probability $P\sim\exp(-N^2k)$, while $c(k)$ corresponds to the divergence as $\exp[N^2c(k)]$ of the number of such singularities. By inverse Legendre transform, one finds

$$c(k)=\operatorname*{extr}_R[\phi(R)+Rk]. \tag{5}$$

The results are represented in Fig. 3. One would hope to see convergence to a limiting form for N large. Unfortunately our results are restricted to rather small values of N, so that we could not get conclusive evidence. Turning to an analytic approach, one notes that $\phi(R)$, defined in Eq. (4), can be written explicitly as follows:

$$\phi(R)=\lim_{N\to\infty}\frac{1}{N^2}\ln\sum_{\{\xi_o^\mu\}}$$
$$\times\left\{\frac{\int d\vec{J}\ \ \delta(\vec{J}^2-N)\prod_{\mu=1}^{2^N}\Theta(\vec{J}\cdot\vec{\xi^\mu}\xi_o^\mu)}{\int d\vec{J}\ \ \delta(\vec{J}^2-N)}\right\}^R, \tag{6}$$

where the product over μ runs over all possible binary patterns (2^N factors) and the sum over ξ_o^μ over all their possible classifications (2^{2^N} terms). Unfortunately, we have not been able to evaluate the above expression, when the patterns $\{\vec{\xi}^\mu\}$ are the corners of the hypercube. The main problem is that these patterns are not random, so that one cannot perform an average that renders the calculation feasible.

In order to make progress, we turn to the *a priori* probability for a somewhat artificial but nevertheless interesting case, namely, that of $p=2^N$ patterns $\{\vec{\xi}^\mu\}$ randomly chosen on the sphere. One expects that $\phi(R)$ is self-averaging, and the average can be evaluated through a technique introduced some time ago by Monasson and O'Kane [16], and applied recently in the context of the perceptron [18]. In fact, for the present problem, an important simplification takes place, since it turns out that the annealed approximation, in which the average can be moved inside the logarithm, is correct. Since all the 2^p terms in $\Sigma_{\{\xi_o^\mu\}}$ are equal when the average is performed, the remaining calculation is in fact identical to the standard Gardner calculation, but with R playing the role of the usual replica index n:

$$\phi(R)=\langle\phi(R)\rangle=\lim_{N\to\infty}\frac{1}{N^2}\left\langle\ln\sum_{\{\xi_o^\mu\}}\left\{\frac{\int d\vec{J}\ \ \delta(\vec{J}^2-N)\prod_{\mu=1}^{p}\Theta(\vec{J}\cdot\vec{\xi^\mu}\xi_o^\mu)}{\int d\vec{J}\ \ \delta(\vec{J}^2-N)}\right\}^R\right\rangle$$
$$=\lim_{N\to\infty}\frac{1}{N^2}\ln 2^p\left\langle\left\{\frac{\int d\vec{J}\ \ \delta(\vec{J}^2-N)\prod_{\mu=1}^{p}\Theta(\vec{J}\cdot\vec{\xi^\mu}\xi_o^\mu)}{\int d\vec{J}\ \ \delta(\vec{J}^2-N)}\right\}^R\right\rangle.$$

Copying the well known replica symmetric result from the Gardner calculation [see, e.g., [4], but using the result before the limit $n=R\to 0$ is taken, see also Eq. (2.22) in [17] and Eq. (16) in [18]] we find

$$\phi(R)=\lim_{N\to\infty}\frac{1}{N^2}\operatorname*{extr}_q\left\{\frac{N(R-1)}{2}\ln(1-q)+\frac{N}{2}\ln[1\right.$$
$$\left.+(R-1)q]+p\ln\left[2\int Dt\ \ H^R\left(\frac{t\sqrt{q}}{\sqrt{1-q}}\right)\right]\right\}, \tag{7}$$

where $Dt=\exp(-t^2/2)/\sqrt{2\pi}$ and $H(x)=\int_x^\infty Dt$. The delicate point now is to find the correct scaling for q, which has the usual meaning of the typical overlap for two vectors $\vec{J}$ inside the regions that give the dominant contribution to $\phi(R)$. Since the number of these regions diverges as $\exp(N^2)$, their typical size will be of the order of $\exp(-N^2)$, so that q will go to 1 in the limit $N\to\infty$. The correct large N behavior of q can be inferred from more or less intuitive arguments. Here, we just note that the scaling

$$q=1-\frac{\mu^2N^2}{2^{2N}}, \tag{8}$$

where μ is of order one, yields the required result. The prefactor $p=2^N$ is neutralized and a remaining factor in N^2 cancels the one in front of the brackets in Eq. (7). Furthermore, by including lower order corrections [19], one obtains a bona-fide saddle point equation for the new variable μ, which can moreover be solved explicitly. By combining Eqs. (7) and (8), one thus finds

$$\phi(R)=(1-R)\left(\ln 2-\frac{\ln N}{N}+\frac{1}{N}\right)+\frac{1}{N}\left\{\frac{\ln R}{2}+(1-R)\times\ln\frac{2\int_0^{\infty}du\ \left[H^R(u)+H^R(-u)-1\right]}{\sqrt{2\pi}(1-R)}\right\}. \qquad (9)$$

Since the Legendre transform of $1-R$ is a constant, we conclude that to dominant order in N the spectrum is a monofractal with all the regions equally large. Their total number is of course exactly given by 2^{N^2}, in agreement with the Cover-Sauer result. The monofractal behavior persists if the logarithmic correction in N is included. Only at the level of $1/N$ corrections does a genuine multifractal spectrum arise. To compare these large N analytic results with numerical finite size results, we have performed Monte Carlo simulations for the case of $p=2^N$ random patterns in the same way as we did for the binary patterns. The profiles that are obtained for $N=5$ and $N=6$ are shown in Fig. 4, while the numerical results for the multifractal spectrum and comparison with the theory, cf. Eq. (9), are included in Fig. 3. The agreement is reasonable taking into account that we are working with very small values of N. One also notes that the spectrum is clearly different from the one for the hypercube, although the overall shape is the the same.

We close with a discussion of the above problem from a more geometric perspective. Consider the surface of the N sphere as it is being cut into pieces by p large circles with random orientation. With increasing p, the number of these pieces increases and their size decreases. The geometric properties of this randomly broken object can, in the limit $N\rightarrow\infty$ and with an appropriate corresponding scaling of p, be characterized by a multifractal spectrum. The latter was calculated in [18] for $p=\alpha N$, with α finite. Here, we considered the more unusual scaling $p=2^N$. Surprisingly, and in contrast to the multifractal behavior for $p\sim N$, we find that the spectrum is monofractal, at least to dominant order in N. Hence the size k that appears most often is also the size that covers almost all the surface of the sphere. As a result, picking at random one linearly separable classification from all the linearly separable classifications on a set of $p=2^N$ random examples is tantamount to picking a random perceptron teacher on the sphere. This, however, is no longer true for $p\sim N$, where the multifractal nature is in fact responsible for the distinction between the storage and generalization problem [18]. The difference between the $p\sim N$ and $p=2^N$ is, however, not entirely unexpected since a monofractal behavior is approached for $p=\alpha N$ in the limit $\alpha\rightarrow\infty$. A monofractal behavior to dominant order of N is therefore expected for any scaling in which p increases faster than N.

C.V.d.B. acknowledges support from the Program on Inter-University Attraction Poles, Prime Minister's Office, Belgian Government, and the Flemish FWO.

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

[2] E. Gardner and B. Derrida, J. Phys. A **22**, 1983 (1989).

[3] G. Györgyi and N. Tishby, in *Neural Networks and Spin Glasses*, edited by W.K. Theumann and R. Koberle (World Scientific, Singapore, 1990), p. 3.

[4] J. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computing* (Addison-Wesley, Reading, MA, 1991).

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

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

[7] M. Opper and W. Kinzel, in *Physics of Neural Networks III*, edited by E. Domany, J.L. Van Hemmen, and K. Schulten (Springer, Berlin, 1994).

[8] C. Van den Broeck, Acta Phys. Pol. B **25**, 903 (1994).

[9] A. Engel, Mod. Phys. Lett. B **8**, 1683 (1994).

[10] M. Bouten, J. Schietse, and C. Van den Broeck, Phys. Rev. E **52**, 1958 (1995).

[11] C. Van den Broeck and R. Kawai, Phys. Rev. A **42**, 6210 (1990).

[12] D.B. Schwartz, V.K. Samalam, S.A. Solla, and J.S. Denker, Neural Comput. **2**, 374 (1990).

[13] F. Rosenblatt, *Principles of Neurodynamics* (Spartan, New York, 1962).

[14] N. Sauer, J. Comb. Theory, Ser. A **13**, 145 (1972).

[15] Note that we have used the notation $c(k)$ rather than the more usual $\phi(\alpha)$ because the symbols α and ϕ are reserved to denote other quantities in the neural network community.

[16] R. Monasson and D. O'Kane, Europhys. Lett. **27**, 85 (1994).

[17] B. Derrida, R.B. Griffiths, and A. Prugel-Bennett, J. Phys. A **24**, 4907 (1991).

[18] M. Weigt and A. Engel, Phys. Rev. E **55**, 4552 (1996).

[19] By blindly applying this change of variable on Eq. (7), and solving the saddle point equation for μ, one gets the result given in Eq. (9). One has, however, to remember that q is in fact an integration variable so that the change from q to μ leads to an extra constant factor $\exp(-2N\ln 2)$, i.e., a term $-2\ln 2$ of order N. But such a term is clearly wrong since one has to find an exponent equal to 0 for $R=1$. The point is that to correctly include all the finite order corrections, one has already to do this at the level of the saddle point calculation that leads to Eq. (7). It turns out that this extra correction just yields another constant factor $\exp(+2N\ln 2)$, which exactly cancels the one coming from the change of integration variable.