RAPID COMMUNICATIONS

PHYSICAL REVIEW E VOLUME 53, NUMBER 3 MARCH 1996

Multifractal analysis of the coupling space of feedforward neural networks

A. Engel and M. Weigt

Institut für Theoretische Physik, Otto-von-Guericke-Universität Magdeburg, PSF 4120, 39016 Magdeburg, Germany

(Received 14 July 1995)

Random input patterns induce a partition of the coupling space of feedforward neural networks into different cells according to the generated output sequence. For the perceptron this partition forms a random multifractal for which the spectrum $f(\alpha)$ can be calculated analytically using the replica trick. A phase transition in the multifractal spectrum corresponds to the crossover from percolating to nonpercolating cell sizes. Instabilities of negative moments are related to the Vapnik-Chervonenkis (VC) dimension [Theor. Prob. Appl. **16**, 264 (1971)].

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

Multifractal concepts were originally introduced in the context of developed turbulence [1] and chaotic dynamical systems [2] and have become since then a standard tool to analyze physical systems with richer structure than that induced by dilation symmetry alone (for reviews see [3]). Contrary to simple scale invariant situations as provided, e.g., by systems at second order phase transitions that can be classified with the help of a few critical exponents only, the description of multifractals requires a full range of scaling exponents specified by a continuous function $f(\alpha)$. The reason for this multitude of exponents is the fact that different moments of the underlying probability distribution are dominated by different fractal subsets of the system.

The simplest examples of multifractal measures are provided by deterministic recursive constructions such as the two-scale Cantor set [3]. However, many multifractals observed experimentally as, e.g., in diffusion limited aggregation, are generated by random processes. It is therefore of general interest to analyze simple models of random systems that exhibit multifractality [4]. A particularly simple case is provided by fractals on which a measure of constant density is distributed. The multifractal properties are then of purely geometrical origin and characterize the fractal support itself [5].

In the present paper we show that the coupling space of simple feedforward neural networks storing random input-output mappings displays multifractality. The corresponding spectrum $f(\alpha)$ can be determined explicitly using the replica trick. On the one hand, these systems may hence serve as examples to test the properties of random multifractals. On the other hand, the multifractal analysis of the cell structure imposed on the coupling space by the random input-output mappings refines and extends the standard statistical mechanics analysis of the storage [6] and generalization properties [7] of these systems.

We consider a perceptron with N input bits $\xi_i = \pm 1$, $i = 1, \ldots, N$, and one output $\sigma = \pm 1$. The output is given as the sign of the scalar product between the input and the coupling vector $\boldsymbol{J}$ of the perceptron, i.e.,

$$\sigma = \mathrm{sgn}(\boldsymbol{J}\boldsymbol{\xi}) = \mathrm{sgn}\left(\sum_i J_i \xi_i \right). \tag{1}$$

We will mainly consider the case in which the $\boldsymbol{J}$-vector is binary, $J_i = \pm 1$ (Ising perceptron) and defer the discussion of continuous couplings (spherical perceptron) to the end of this paper.

Given $p = \gamma N$ different input patterns $\boldsymbol{\xi}^\mu$ we can associate with every pattern a hyperplane perpendicular to it that cuts the coupling space of the perceptron into two halves according to the possible output $\sigma^\mu = \pm 1$. If the inputs are generated independently at random we will hence find a *random partition of the coupling space* into at most 2^p cells. These cells can be labeled by their output sequences $\boldsymbol{\sigma} = \{\sigma^\mu\}$ and their size gives the probability $P(\boldsymbol{\sigma})$ that, for given input sequence $\{\boldsymbol{\xi}^\mu\}$, the outputs $\boldsymbol{\sigma}$ are generated by a randomly chosen coupling vector $\boldsymbol{J}$. It is well known that the storage and generalization properties of the neural network are closely related to this probability distribution [8–11].

The natural scale for the cell sizes in the thermodynamic limit $N \to \infty$ is $\epsilon = 2^{-N}$. Due to the random orientation of the hyperplanes, however, the cells will differ significantly in size from each other. To describe these fluctuations quantitatively we introduce the crowding index $\alpha(\boldsymbol{\sigma})$ by

$$P(\boldsymbol{\sigma}) = \epsilon^{\alpha(\boldsymbol{\sigma})} \tag{2}$$

and characterize $P(\boldsymbol{\sigma})$ by its moments

$$\langle P^q \rangle = \sum_{\boldsymbol{\sigma}} P^q(\boldsymbol{\sigma}) = \epsilon^{\tau(q)} \tag{3}$$

with the mass exponent $\tau(q)$. As usual [1,2] the multifractal spectrum $f(\alpha)$ is given by the Legendre transform of $\tau(q)$:

$$f(\alpha) = \min_q \left[\alpha q - \tau(q) \right] \tag{4}$$

and $\mathcal{N}(\alpha) = \epsilon^{f(\alpha)}$ gives the number of cells of size ϵ^α.

For large N we expect that τ and f become self-averaging, i.e., they will no longer depend on the choice of the random inputs $\boldsymbol{\xi}^\mu$. We can therefore calculate $\tau(q)$ from

$$\tau(q) = -\lim_{N \to \infty} \frac{1}{N \ln 2} \left\langle\!\!\left\langle \ln \sum_{\boldsymbol{\sigma}} P^q(\boldsymbol{\sigma}) \right\rangle\!\!\right\rangle \tag{5}$$

where $\langle\langle\rangle\rangle$ denotes the average over the distribution of the input patterns which we take as

$$\mathscr{P}(\xi_i^\mu)=\prod_{i,\mu}\left[\tfrac{1}{2}\delta(\xi_i^\mu+1)+\tfrac{1}{2}\delta(\xi_i^\mu-1)\right]. \tag{6}$$

The calculation of $\tau(q)$ uses a variant of the replica trick introduced recently by Monasson and O'Kane [12] and will only be sketched. Starting with the definition of $P(\boldsymbol{\sigma})$

$$P(\boldsymbol{\sigma})=\sum_{\boldsymbol{J}}\prod_{\mu=1}^{p}\theta(\sigma^\mu \boldsymbol{J}\boldsymbol{\xi}^\mu) \tag{7}$$

with the theta function $\theta(x)=1$ if $x\geq 0$ and $\theta(x)=0$ else we introduce one replica index α running from 1 to q in order to represent the qth power of P and another replica index a running from 1 to n to represent the ln in Eq. (5) in the usual way [13]. We then find

$$\tau(q)=-\lim_{N\to\infty}\frac{1}{N\ln 2}\times\lim_{n\to 0}\frac{1}{n}\left[\left\langle\left\langle\sum_{\sigma_\mu^a}\sum_{J^{a\alpha}}\prod_{\mu,a,\alpha}\theta(\sigma_\mu^a J^{a\alpha}\xi^\mu)\right\rangle\right\rangle-1\right]. \tag{8}$$

Next the average is performed and the resulting expression is written as a saddle-point integral over the elements of the overlap matrix

$$Q_{ab}^{\alpha\beta}=\frac{1}{N}\sum_i J_i^{a\alpha}J_i^{b\beta} \tag{9}$$

and its conjugate $\hat{Q}_{ab}^{\alpha\beta}$. To solve the remaining extremalization problem we assume replica symmetry (RS) which in the present problem is given by the ansatz [12]

$$\begin{aligned} Q_{ab}^{\alpha\beta}&=Q_1 \quad \text{if } a=b,\alpha\neq\beta,\\ Q_{ab}^{\alpha\beta}&=Q_0 \quad \text{if } a\neq b, \end{aligned} \tag{10}$$

and similarily for $\hat{Q}_{ab}^{\alpha\beta}$. Hence Q_1 denotes the typical overlap between couplings in the same cell (same output vector $\boldsymbol{\sigma}$) whereas Q_0 characterizes the overlap between couplings in different cells. The reliability of the RS ansatz will be discussed below.

In this way we get

$$\begin{aligned}\tau(q)=\frac{1}{\ln 2}\,\mathrm{extr}_{Q_1,Q_0,\hat{Q}_1,\hat{Q}_0}\Bigg[&\frac{q\hat{Q}_1}{2}[1-(1-q)Q_1]-\frac{q^2}{2}\hat{Q}_0Q_0\\&-\int Dz_0\ln\int Dz_1\cosh^q\left(\sqrt{\hat{Q}_1-\hat{Q}_0}z_1+\sqrt{\hat{Q}_0}z_0\right)\\&-\gamma\int Dt_0\ln 2\int Dt_1 H^q\left(\frac{\sqrt{Q_1-Q_0}t_1+\sqrt{Q_0}t_0}{\sqrt{1-Q_1}}\right)\Bigg]\end{aligned} \tag{11}$$

where $Dx=\exp(-x^2/2)/\sqrt{2\pi}$ and $H(y)=\int_y^\infty Dx$. Inspection of the saddle-point equations resulting from (11) reveals that these are always fulfilled for $Q_0=\hat{Q}_0=0$. Physically this is due to the symmetry $(\boldsymbol{J},\boldsymbol{\sigma})\leftrightarrow(-\boldsymbol{J},-\boldsymbol{\sigma})$ in Eq. (1) which ensures that to every cell there is a "mirror" cell of equal size. As a consequence Eq. (11) simplifies to

$$\begin{aligned}\tau(q)=\frac{1}{\ln 2}\,\mathrm{extr}_{Q_1,\hat{Q}_1}\Bigg[&\frac{q\hat{Q}_1}{2}[1-(1-q)Q_1]\\&-\ln\int Dz\,\cosh^q(\sqrt{\hat{Q}_1}z)\\&-\gamma\ln 2\int Dt\,H^q\left(\frac{\sqrt{Q_1}t}{\sqrt{1-Q_1}}\right)\Bigg].\end{aligned} \tag{12}$$

Extremizing this expression with respect to Q_1 and $\hat{Q}_1$ numerically and using Eq. (4) we find $f(\alpha)$ as shown in Fig. 1 for different values of the loading parameter γ. Also shown are results from exact enumerations up to $N=30$. The inset gives a finite size analysis demonstrating very good agreement between the analytical results and the extrapolation from the numerical data.

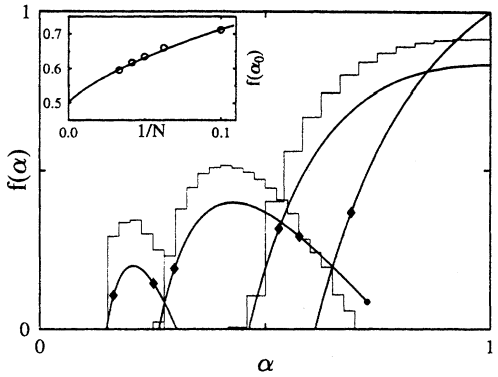

FIG. 1. Multifractal spectrum $f(\alpha)$ characterizing the cell structure of the coupling space of an Ising perceptron classifying γN random input patterns for $\gamma=0.2$, 0.4, 0.833, and 1.245 (from left to right). The diamonds mark the region of validity of the replica symmetric ansatz. The dot denotes the location of the instability of negative moments for $\gamma=0.4$ (see text). The histograms are exact enumeration results for $N=30$, $p=6$; $N=30$, $p=12$; and $N=24$, $p=20$ respectively. The inset shows a finite size analysis of $f(\alpha_0)$ for $\gamma=0.5$. The correct value for $N\to\infty$ is 0.5, the line describes the first correction to the saddle point for $N<\infty$, and the statistical error of the numerical results is smaller than the symbol size.

For small values of γ the $f(\alpha)$ curves have the typical bell-shaped form. The two zeros $\alpha_{min}(\gamma)$ and $\alpha_{max}(\gamma)$ specify the largest [14] and the smallest cell occurring with nonzero probability respectively. Moreover $\alpha_0(\gamma)=\mathrm{argmax} f(\alpha)$ defines the size of the typical cell and $f(\alpha_0)=\gamma$ indicates that for small γ all possible cells do indeed occur. Finally, from the normalization of $P(\boldsymbol{\sigma})$ we find for all γ that $\tau(q=1)=0$. This implies that the $f(\alpha)$ curves are all tangent to the line $f=\alpha$. We denote the abscissa of the tangential point by $\alpha_1(\gamma)$. Cells of size ϵ^{α_1} contribute most to the coupling space.

For larger values of γ it becomes important that due to $J_i=\pm 1$ the coupling space is discrete and the cell sizes must always be multiples of ϵ. Values of α larger than 1 thus correspond to empty cells. For $\alpha_0=1$ the *typical* cell is empty and hence $\alpha_0(\gamma_c)=1$ determines the storage capacity

γ_c [11]. From the figure we infer $\gamma_c \cong 0.833$. Similarly $\alpha_1 = 1$ indicates that the coupling space is dominated from cells containing a single $\boldsymbol{J}$-vector only. Therefore $\alpha_1(\gamma_g) = 1$ defines the threshold to perfect generalization [11] and the results shown in the figure yield $\gamma_g = 1.245$.

Both the values for γ_c and γ_g have been derived previously [15] and in fact $\alpha(\gamma) = 0$ is similar to the zero-entropy condition used to derive them. However, it should be emphasized that the justification of the zero-entropy condition within the traditional Gardner approach requires one-step replica symmetry breaking (RSB). On the contrary, in the present approach not only does RS already give the correct results but anticipating that $Q_0 = \hat{Q}_0 = 0$ for $\alpha_1 < \alpha < \alpha_0$ (see below) we can even go without the replica index a and calculate $\tau(q)$ as an *annealed* average over the input distribution (6). This is technically much simpler than a one-step RSB calculation and may open the way to a more detailed analysis of multilayer networks also [16].

We now turn to the reliability of the RS results with $Q_0 = 0$. First of all we note that from (4) and (3) $f(\alpha)$ is the entropy of the discrete spin system $\boldsymbol{\sigma}$ [with Hamiltonian $\alpha(\boldsymbol{\sigma})$] and must therefore be non-negative. However, before f becomes negative we find for both positive and negative q a transition to a saddle point with $Q_0, \hat{Q}_0 > 0$ at values of $q = q_\pm$ given by

$$\sqrt{\alpha} = |q_\pm - 1| \hat{Q}_1 (1 - Q_1)[1 - (1 - q_\pm) Q_1]. \qquad (13)$$

Since Q_0 gives the typical overlap between $\boldsymbol{J}$-vectors belonging to different cells the analogy with the spin glass problem suggests that $Q_0 > 0$ signals broken ergodicity. This means in the present context that moments of $P(\boldsymbol{\sigma})$ with $q > q_+$ or $q < q_-$ are dominated by cells that no longer percolate in coupling space. Whereas for $q_+ > q > q_-$ all dominating cells can be reached from one another without entering cells of a different size this is no longer true for the remaining values of q. Note that the transition occurs always outside the interval (α_1, α_0).

We have also investigated the transverse stability of the RS saddle point using standard techniques [17]. We find that at $q = q_\pm$ the RS saddle point becomes unstable also with respect to RSB and this instability is not removed by $Q_0 > 0$ [18]. On the other hand, we believe that our qualitative picture of a percolation transition at q_c remains valid also in a RSB solution.

We have obtained similar results for the spherical perceptron (see also [14]). For small values of γ the $f(\alpha)$ curves are almost identical to those of the Ising perceptron. Since there is now no smallest possible size of a cell the storage capacity $\gamma_c = 2$ [9,6] has to be determined from $\alpha_0(\gamma_c) \to \infty$. For $\gamma \geq \gamma_c$ the spectrum $f(\alpha)$ is hence monotonously increasing and the asymptotic value of f for $\alpha \to \infty$ remains smaller than γ since a growing fraction of classifications cannot be realized. The generalization properties of the spherical perceptron are characterized by the information dimension $f(\alpha_1)$ of the multifractal cell structure which is related to the volume $\mathscr{V}$ of the version space by $\mathscr{V} = \epsilon^{f(\alpha_1)}$ [11]. The longitudinal and transverse instability of the RS solution with $Q_0 = 0$ occurs again at the same values $q_\pm$ of q given now by $\sqrt{\alpha} = |q_\pm - 1| Q_1$.

Finally we find from the RS results for both the Ising and the spherical perceptron an instability $\langle P^q \rangle \to \infty$ of negative moments if $q < q_c$ where $q_c = 1 - 1/\alpha$ for $\alpha \leq 1$ and $q_c = 0$ for $\alpha > 1$. The corresponding end point of the $f(\alpha)$ curve for $\gamma = 0.4$ is marked by a dot in the figure; for $\gamma = 0.2$ it occurs for $f < 0$. These divergencies are due to the possibility of empty cells with $P(\boldsymbol{\sigma}) = 0$. For small values of γ we find $q_c < q_-$ and the instability lies outside the region of validity of RS. In this case it is therefore necessary to include RSB to elucidate the nature of this divergence and we speculate that this is due to the fact that empty cells can only occur due to very rare realizations of the input patterns $\boldsymbol{\xi}^\mu$ [19]. For γ larger than a threshold value γ^{VC}, however, the instability occurs within the region of validity of RS. Then the probability of empty cells can no longer be exponentially small in N. The instability of negative moments is hence related to the Vapnik-Chervonenkis dimension of the neural network, more precisely γ^{VC} determined from $q_c(\gamma^{VC}) = q_-(\gamma^{VC})$ gives an upper bound on the VC dimension d_{VC} [20]. For the spherical perceptron we find in this way the well known exact result $\gamma^{VC} = 1$ [9]. For the Ising perceptron we get similarly $\gamma^{VC} \cong 0.557$. The VC dimension of the Ising perceptron is not known up to now. One can show that it is at least 0.5 [21] and there is numerical evidence that it is indeed 0.5 for large N [22]. Our RS upper bound is somewhat larger which may indicate that there is a *discontinuous* transition to RSB for the Ising perceptron. This question is currently under study.

In conclusion we have shown that a multifractal analysis of the phase space of neural networks allows one to refine the standard statistical mechanics approach to learning and generalization substantially. The instabilities found in the spectrum $f(\alpha)$ can be related to physical properties of these systems. We hope that these results may serve as a guide to understanding similar transitions in other examples of random multifractals. Moreover, it would be interesting to see whether there are other examples for percolation transitions in high-dimensional spaces that can be detected from a bifurcation of an overlap parameter Q_0.

We would like to thank Chris van den Broeck for bringing the connection between the cell structure of the perceptron and multifractals to our attention. Stimulating discussions with Günther Radons and Manfred Opper are gratefully acknowledged.

[1] B. B. Mandelbrot, J. Fluid. Mech. **62**, 331 (1974); U. Frisch and G. Parisi, in *Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics*, edited by M. Ghil, R. Benzi, G. Parisi (North-Holland, Amsterdam, 1985); R. Benzi, G. Paladin, G. Parisi, and A. Vulpiani, J. Phys. A **17**, 3521 (1984).

[2] P. Grassberger, Phys. Lett. **97A**, 227 (1983); T. C. Halsey, M. H. Jensen, L. P. Kadanoff, I. Procaccia, and B. I. Shraiman, Phys. Rev. A **33**, 1141 (1986).

[3] G. Paladin and A. Vulpiani, Phys. Rep. **156**, 147 (1987); J.

Feder, *Fractals* (Plenum, New York, 1988); *Fractals and Disordered Systems*, edited by A. Bunde and S. Havlin (Springer, Berlin, 1991).

[4] M. H. Jensen, G. Paladin, and A. Vulpiani, Phys. Rev. E **50**, 4352 (1994); Y. Y. Goldschmidt and T. Blum, *ibid.* **48**, 161 (1993).

[5] T. Tél and T. Vicsek, J. Phys. A **20**, L835 (1987).

[6] E. Gardner, J. Phys. A **21**, 257 (1988); E. Gardner and B. Derrida, *ibid.* **21**, 271 (1988).

[7] G. Györgyi and N. Tishby, in *Workshop on Neural Networks and Spin Glasses*, edited by K. Theumann and W. K. Koeberle (World Scientific, Singapore, 1990); M. S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).

[8] J. A. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).

[9] T. M. Cover, IEEE Trans. Electron. Comput. **EC-14**, 326 (1965).

[10] D. Haussler, M. Kearns, and R. Schapire, *Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension*, Proceedings *COLT '91* (Morgan Kaufmann, San Mateo, CA, 1991); M. Opper and D. Haussler, Phys. Rev. Lett. **66**, 2677 (1991).

[11] B. Derrida, R. B Griffith, and A. Prügel-Bennett, J. Phys. A **24**, 4907 (1991).

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

[13] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[14] M. Biehl and M. Opper, in *Neural Networks: The Statistical Mechanics Perspective*, edited by Jong-Houn Oh, Chulan Kwon, and Sungzoon Cho (World Scientific, Singapore, 1995).

[15] W. Krauth and M. Mezard, J. Phys. (Paris) **50**, 3057 (1989); G. Györgyi, Phys. Rev. Lett. **64**, 2957 (1990).

[16] R. Monasson and R. Zecchina, Phys. Rev. Lett. **75**, 2432 (1995).

[17] J. R. L. de Almeida, and D. Thouless, J. Phys. A **11**, 983 (1978); T. Temesvari, C. De Dominicis, and I. Kondor, *ibid.* **27**, 7569 (1994).

[18] This is reminiscent of the SK model in zero field. The analog of a nonzero field situation can be studied by introducing a threshold of the perceptron.

[19] For large N the probability for patterns distributed according to (6) not to be in general position scales as 2^{-N}.

[20] V. N. Vapnik and A. Y. Chervonenkis, Theor. Prob. Appl. **16**, 264 (1971); A. Engel, Mod. Phys. Lett. B **8**, 1683 (1994).

[21] D. Helmbold (private communication).

[22] G. Stambke, Diploma-Thesis, University of Giessen, 1992 (unpublished).