Learnability of periodic activation functions: General results

Michal Rosen-Zvi,[1] Michael Biehl,[2] and Ido Kanter[1]
[1]*Minerva Center and the Department of Physics, Bar-Ilan University, Ramat Gan 52900, Israel*
[2]*Institut für Theoretische Physik, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany*

(Received 26 February 1998)

On-line learning in the presence of continuous periodic activation functions is studied analytically. The effect of the ambiguity (an infinite number of inputs with different local fields can produce the same output) on the learnability is examined. A universal interplay between the general features of the activation function (wave number, parity, etc.) and the critical learning rate is found. Analytical results are extended also to multilayer architectures with nonlinear output units. Results are compared with simulations. [S1063-651X(98)06409-5]

PACS number(s): 84.35.+i, 07.05.Mh, 89.70.+c

The theory of learning has benefited to a great extent from the application of statistical physics methods, see e.g., [1,2] for reviews. Statistical mechanics provides the tools to investigate, for instance, large neural networks [3], learning a rule from randomized example data. It allows one to calculate typical properties, such as the generalization error, which quantifies the average amount of disagreement between *student* and unknown rule.

One successful line of research concerns the physics of so-called on-line learning processes [3,4] and was initiated in [5,6]. From a practical point of view, on-line learning is particularly attractive because it uses only the latest from a sequence of examples for training. This obviously reduces the storage needs and computational effort in comparison with memory based off-line prescriptions. On the other hand, this very property makes it possible to investigate a variety of learning scenarios analytically. The learning dynamics is described exactly in terms of coupled differential equations for self-averaging order parameters in the thermodynamic limit.

A most remarkable outcome of this theory is that the performance of efficient on-line algorithms is, despite their simplicity, comparable with that of sophisticated off-line or batch prescriptions [6–9]. Among the different architectures that have been studied in this framework are such diverse systems as the simple perceptron (e.g., [5,6]), specific multilayer networks of threshold units (e.g., [10,11]), and two-layered structures constructed from continuous units [12–14].

Apart from perhaps one exception, the so-called reversed wedge perceptron [15], all networks investigated so far consist of units with monotonic activation functions, where the most prominent examples are continuous sigmoidal or linear units and discontinuous threshold neurons. For the latter the generalization error approaches its minimum value according to a power law with the number of examples, in general (e.g., [7,9]). In contrast, an exponential decay is typical for networks consisting of units with differentiable sigmoidal activations, provided the learning rate is smaller than some critical value η_c [12,13]. Furthermore, nontrivial transient behavior of these systems, like the occurrence of quasistationary plateau states in the learning dynamics, can be theoretically understood within this framework [11,13,14].

It remains an open question precisely which features of the activation functions determine the properties of the learning network. (a) Does a universal behavior exist in the sense that there is an interplay between general features of the activation function and, for instance, the critical learning rate? (b) Is the above mentioned exponential asymptotic decay a direct consequence of the continuous nature of the activation or is it crucial that the function is monotonic and invertible? (c) How does the critical learning rate depend on the above properties? (d) What is the relevance of symmetries in the transfer functions? (e) Will new classes of behavior emerge when the scope of possible characteristics is extended? (f) What is the effect, if any, on the nature of plateaus in the learning process of multilayer architectures with nonlinear output units?

In order to address and investigate these questions we study in this paper neural networks with continuous but periodic activation function. In such networks an infinite number of different local fields or internal representations can produce the same output. Due to this ambiguity an example $\{\boldsymbol{\xi},\tau(\boldsymbol{\xi})\}$ contains less information about the rule than in cases where the transfer functions are invertible. We will investigate here to what extent this property affects the dynamics of learning.

Throughout the following we restrict the analysis to teacher-student scenarios with perfectly matching network architectures. Training is guided by the quadratic deviation $\epsilon(\boldsymbol{\xi})=[\sigma(\boldsymbol{\xi})-\tau(\boldsymbol{\xi})]^2/2$, which compares the student's response σ with the rule or teacher output τ for a given high-dimensional input vector $\boldsymbol{\xi}\in\mathbb{R}^N$. Accordingly, the generalization error is defined as the average $\langle\epsilon(\boldsymbol{\xi})\rangle$ over the distribution of inputs. Throughout this paper we will consider random vectors $\boldsymbol{\xi}$ with independent, identically distributed components of zero mean and unit variance.

Upon the presentation of a single example input-output pair $\{\boldsymbol{\xi}^\mu,\tau^\mu=\tau(\boldsymbol{\xi}^\mu)\}$ the vector $\vec{W}$ of all adjustable parameters in the student network is updated according to the following stochastic gradient descent prescription:

$$\vec{W}(\mu)=\vec{W}(\mu-1)-\frac{\eta}{N}\vec{\nabla}_{\vec{W}}\epsilon(\boldsymbol{\xi}^\mu)|_{\vec{W}(\mu-1)}, \tag{1}$$

where the same learning rate η is used everywhere in the network. The vector $\vec{W}$ includes in the following only the weights of the student network (input-to-hidden as well as hidden-to-output). We plan to study the adjustment of additional parameters like the phases of the considered activation function in the formalism.

Note that so far only two-layered systems with linear output units have been treated analytically [13,14]. The consideration of periodic activations enables us to extend this formalism to networks of several layers with nonlinear units in each layer. The details of the solution and its mathematical insight will be presented after the discussion of the perceptron.

As a first example we consider a teacher-student scenario with matching single unit networks. Normalized teacher weights $\boldsymbol{B} \in \mathbb{R}^N, \boldsymbol{B}^2 = 1$ define the rule output

$$\tau(\boldsymbol{\xi}) = g(y) = \sin(ky + \theta) \tag{2}$$

with $y = \boldsymbol{B} \cdot \boldsymbol{\xi}$ for any N-dimensional input vector $\boldsymbol{\xi}$. The additional parameters k and θ fix the wave number and phase of the periodic activation function, respectively.

The above mentioned ambiguity is most clearly studied in the single node. For sigmoidal activations, y is uniquely determined by τ and, assuming the activation function is known, each example provides a linear equation of the form $\boldsymbol{B} \cdot \boldsymbol{\xi} = g^{-1}(\tau)$. Here, however, an infinite number of overlaps produces the same output:

$$\tau(y_n) = \tau(y_o) \quad \text{with} \quad y_n = y_o \pm 2\pi n/k, \quad n \in \mathbb{N}. \tag{3}$$

Due to this ambiguity, an example $\{\boldsymbol{\xi}, \tau\}$ contains less information about the unknown rule than in cases with sigmoidal, i.e., invertible monotonic transfer functions.

In general one would expect that the density of values y has a finite width for realistic data. The assumed specific input distribution results in a Gaussian density with $\langle y \rangle = 0$ and $\langle y^2 \rangle = 1$. The number M_{eff} of overlaps y in the range $-1 < y < 1$, which can be assigned to the same output, increases linearly with (large) k. Thus the wave number is a direct measure of how pronounced the effect of the ambiguity will be.

The complexity of learning P examples in the case of invertible monotonic transfer functions (like tanh) is the same as solving P linear equations with N variables (N weights of the student). In contrast, the case of learning P examples with periodic activation functions results in

$$M_{\text{eff}} \propto k^P \tag{4}$$

possible different sets of P linear equations, due to the ambiguity (3). Therefore, the learning process in the case of periodic functions has to overcome two difficulties: (I) to find the most appropriate set of equalities among exponentially many with the size of the training set P and (II) then to solve the set of equalities as for monotonic activation functions.

The single weight vector $\boldsymbol{J} \in \mathbb{R}^N$ parametrizes the student hypothesis $\sigma(\boldsymbol{\xi}) = \sin(k\boldsymbol{J} \cdot \boldsymbol{\xi} + \theta)$ about the unknown rule, whereas the correct values of θ and k are taken to be known in advance. Note that assuming knowledge of the wave number does not constitute a restriction of our model since the norm of the student weights will not be fixed in the learning process. Thus, any mismatched value $\hat{k} \neq k$ in the student could be compensated for by tuning $Q = \boldsymbol{J}^2$ such that $\sqrt{Q}\,\hat{k} = k$.

In order to calculate the generalization error we observe that the randomness of the input enters only through the quantities $x = \boldsymbol{J} \cdot \boldsymbol{\xi}$ and $y = \boldsymbol{B} \cdot \boldsymbol{\xi}$, which are distributed according to a two-dimensional Gaussian density with $\langle x \rangle = \langle y \rangle = 0$, $\langle x^2 \rangle = Q$, $\langle y^2 \rangle = 1$, and $\langle xy \rangle = R = \boldsymbol{J} \cdot \boldsymbol{B}$. We obtain

$$\epsilon_g = \tfrac{1}{2}[1 - A_- + [A_+ - \tfrac{1}{2}(e^{-2k^2 Q} + e^{-2k^2})]\cos(2\theta)] \tag{5}$$

where $A_\pm = e^{-k^2(1+Q\pm 2R)/2}$.

Learning proceeds according to the on-line gradient descent prescription (1), here the change of weights upon presentation of example $\{\boldsymbol{\xi}^\mu, \tau^\mu = g(y^\mu)\}$ is

$$\boldsymbol{J}(\mu) = \boldsymbol{J}(\mu-1) - \frac{\eta}{N}(\sigma^\mu - \tau^\mu)k\,\cos(kx^\mu + \theta)\boldsymbol{\xi}^\mu$$

$$\text{with} \quad x^\mu = \boldsymbol{J}(\mu-1) \cdot \boldsymbol{\xi}^\mu. \tag{6}$$

Following the method described at length in e.g., [12,13], recursion relations for the self-averaging $R(\mu) = \boldsymbol{J}(\mu) \cdot \boldsymbol{B}$ and $Q(\mu) = \boldsymbol{J}^2(\mu)$ are derived which become, in the limit $N \to \infty$, differential equations in continuous time $\alpha = \mu/N$. The average over the sequence of independent vectors $\boldsymbol{\xi}$ is performed as outlined above and one obtains the deterministic equations of motion

$$\frac{dR}{d\alpha} = \frac{\eta k^2}{2}\{[(R+1)A_+ - 2Re^{-2k^2 Q}]\cos(2\theta) - (R-1)A_-\}, \tag{7}$$

$$\begin{aligned}\frac{dQ}{d\alpha} = {}& \eta k^2\{[(R+Q)A_+ - 2Qe^{-2k^2 Q}]\cos(2\theta) - (Q-R)A_-\}\\ &+ \frac{\eta^2 k^4}{8}[2(e^{-2k^2 Q} - e^{-2k^2} - B_- + A_+)\cos(2\theta) + 3\\ &- A_-^4 - 2A_- + (2B_+ - e^{-8k^2 Q} - A_+^4)\cos(4\theta)]\end{aligned} \tag{8}$$

with $A_\pm$ from Eq. (5) and $B_\pm = e^{-k^2(1+9Q\pm 6R)/2}$.

First we observe that for all values of θ a configuration with $R = Q = 1$ that corresponds to perfect learning ($\epsilon_g = 0$) is a fixed point of Eqs. (7) and (8). The corresponding linearization of the system is of the form $d/d\alpha\,(R,Q)^\top = M(1,1)(1-R,1-Q)^\top$ where the matrix $M(1,1)$ has one eigenvalue that is linear in η and always negative. The second one has a quadratic term in η and becomes positive for learning rates larger than the critical value

$$\eta_c = \frac{8}{k^2}\,\frac{e^{-2k^2}\cos(2\theta) + 1}{3 + e^{-8k^2}\cos(4\theta) + 4e^{-2k^2}\cos(2\theta)},$$

which indicates that perfect learning is only possible for $\eta < \eta_c$. Provided this condition is satisfied, the generalization error decreases like $\epsilon_g \propto \exp[-\lambda\alpha]$ for large α, i.e., its asymptotic behavior is similar to the case of invertible transfer functions.

The generic behavior of the critical learning rate for small wave numbers $k \to 0$ is $\eta_c \propto k^{-2}$. For the special case θ

$=0$ [$g(x)=\sin(kx)$] we obtain, for instance, $\eta_c\approx 2/k^2$, which coincides with the well-known result for the linear perceptron [3]. Note that by expanding Eq. (6) for $k\rightarrow 0$ and $\theta=0$ one obtains a linear update of the form $\boldsymbol{J}(\mu)-\boldsymbol{J}(\mu-1)=-(\eta^2k^2)(x^\mu-y^\mu)\boldsymbol{\xi}^\mu$.

Only the exceptional, *even* activation function $g(x)=\cos(kx)$ ($\theta=\pi/2$) yields the much larger value $2/(3k^4)$ for small wave numbers. One can show that $\eta_c(\theta=\pi/2)>\eta_c(\theta=0)$ holds true for all values of k. This seems to contradict the intuition that learning should be harder for even activation functions since the ambiguity affects the most probable inputs (small local fields). Note, however, that the above consideration concerns the asymptotic properties, but the initial decrease of ϵ_g for $R\ll 1$ is faster for the activation $\sin(kx)$.

It can be shown that the interplay of the general expandable activation function $g(x)$ and the critical learning rate has the following form:

$$\lim_{k\rightarrow 0} g(x)=a_o+\sum_{m=m_0}^{\infty} a_m\,(kx)^m \Rightarrow \eta_c\propto k^{-2m_0} \tag{9}$$

in the limit $k\rightarrow 0$.

The limit $k\rightarrow\infty$ corresponds to a highly oscillatory behavior of the activation. Independent of the specific value of θ one finds that the critical rate decreases like $\eta_c\propto 1/k^2$ in this limit. This universal behavior reflects that successful learning is hindered drastically by the ambiguity discussed above.

The exceptional properties of the activation $g(x)=\cos(kx)$ are related to the fact that in this particular case the output (and thus also ϵ_g) is invariant under a sign change of all weights ($a_1=0$), which is reflected in the existence of fixed points with $R=0$ in the system (7), (8).

In the limit $\eta\rightarrow 0$ the quadratic term in Eq. (8) can be neglected and learning proceeds on a rescaled time scale $\eta\alpha$. The following features distinguish the behavior in comparison to invertible activation functions. (a) *The number of fixed points*: In addition to the attractive configuration ($R=1,Q=1$) we find, independent of k, for the cos activation, the attractive fixed point $(-1,1)$ with $\epsilon_g=0$ due to symmetry and repulsive stationary states (0,0) and (0,1/3). Learning requires initial knowledge $|R(0)|>0$ in order to break the $\pm\boldsymbol{J}$ symmetry. In finite systems fluctuations will guarantee $R(0)=O(1/\sqrt{N})$, a macroscopic overlap will be achieved after a characteristic time of order $\ln N$ [14]. For the sin case Fig. 1 displays the corresponding fixed points of the system for a specific value of k. For a nonzero $\eta<\eta_c$, the number of repulsive fixed points increases with k, which also reflects the ambiguity Eq. (4). (b) *Flow* $Q\rightarrow\infty$: The solid line in Fig. 1 marks a separatrix $\tilde{Q}(R)$: for $Q>\tilde{Q}(R)$ the student is unable to learn and the length of $\boldsymbol{J}$ diverges as $\alpha\rightarrow\infty$. This can be understood in the context of the ambiguity problem as large Q corresponds to an effective large wave number $\sqrt{Q}k$ in the student unit.

Next we extend the formalism to networks with one hidden layer and nonlinear hidden and output units. We consider a rule given by the nonoverlapping teacher network $NM{:}M{:}1$ where the NM-dimensional input $\boldsymbol{\xi}=(\boldsymbol{\xi}^{(1)},\ldots,\boldsymbol{\xi}^{(M)})$ consists of M disjoint subsets, each of which is available to only one of the hidden nodes

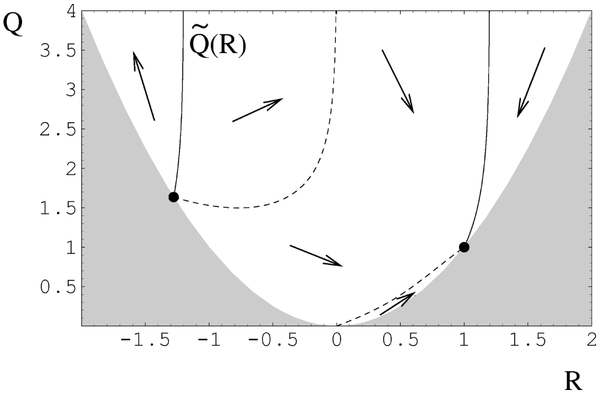

FIG. 1. Fixed points (●) for $g(x)=\sin(x)$ in the limit $\eta\rightarrow 0$, solid lines display the zeros of $dR/d\alpha$ (7), dashed lines correspond to $dQ/d\alpha=0$ in the allowed region $Q>R^2$. Arrows indicate the direction of the temporal evolution according to the signs of the right-hand side of Eqs. (7) and (8).

$$\tau=\tilde{g}\left(\sum_{i=1}^{M} v_i g(y_i)\right), \quad \text{where} \quad y_i=\boldsymbol{B}_i\cdot\boldsymbol{\xi}^{(i)}, \tag{10}$$

where M is the number of hidden units and $g(\tilde{g})$ are the activations of the hidden (output) units and v_i denote the hidden to output weights. The student is assumed to have the same architecture and internal fields $x_i=\boldsymbol{J}_i\cdot\boldsymbol{\xi}^{(i)}$. A learning process of the form (1) is taken to modify the adjustable weights $\{\boldsymbol{J}_i,w_i\}_{i=1,2,\ldots,M}$.

For simplicity we concentrate on the prototype multilayer net with $M=2$ hidden units with $\tilde{g}(x)=g(x)=\sin(kx)$. The explicit form of the student output is $\sigma=\sin\{k[w_1\sin(kx_1)+w_2\sin(kx_2)]\}$ and the quadratic deviation is $\epsilon=(\sigma-\tau)^2/2$. Using standard trigonometric identities, σ (similarly τ) can be rewritten:

$$\begin{aligned}\sigma=&\sin[w_1k\sin(kx_1)]\cos[w_2k\sin(kx_2)]\\&+\cos[w_1k\sin(kx_1)]\sin[w_2k\sin(kx_2)].\end{aligned} \tag{11}$$

Furthermore, relations of the form

$$\sin[w\sin(kx)]=2\sum_{m=0}^{\infty} J_{2m+1}(w)\sin[(2m+1)kx] \tag{12}$$

enable us to write the output as a linear combination of perceptrons with all wave numbers mk, integer m, weighted by appropriate Bessel functions. The trigonometric relation (11) and the expansion (12) are crucial for the extension of the analysis to multilayered networks. Such mathematical simplifications are not available in the case of sigmoidal activation functions.

We obtain the equations of motion

$$\begin{gathered}dR_i/d\alpha=\eta\langle\delta_{i1}y_i\rangle, \quad dQ_i/d\alpha=2\eta\langle\delta_{i1}x_i\rangle+\eta^2\langle\delta_{i1}^2\rangle,\\ dw_i/d\alpha=\eta\langle\delta_{i2}y_i\rangle,\end{gathered} \tag{13}$$

where $\delta_{1\mu}$ is given by

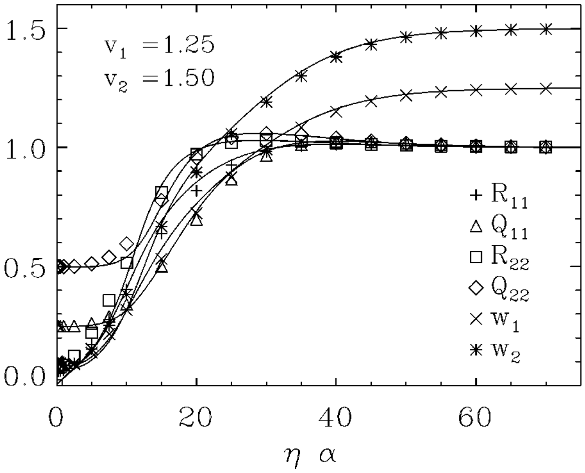

FIG. 2. Learning curves for $2N{:}2{:}1$ with $g(x)=\sin(x)$. Solid lines are the result of numerical integration, Eq. (13), with $\eta\to 0$, symbols correspond to simulations with $N=80$, $\eta=0.02$ averaged over 20 independent runs. Standard error bars would be approximately the size of the symbols.

$$C_{i\mu}\sum_{m_{1,2}=0}^{\infty} J_{2m_1}(w_1)J_{2m_2}(w_2)\gamma_{m_1}\gamma_{m_2}\cos(2m_1x_1)$$

$$\times\cos(2m_2x_2)-4J_{2m_1+1}(w_1)J_{2m_2+1}(w_2)$$

$$\times\sin[(2m_1+1)x_1]\sin[(2m_2+1)x_2]$$

with $C_{i1}=\Delta w_i\cos(x_i)$, $C_{i2}=\Delta\sin(x_i)$, $\gamma_m=2-\delta_{m,0}$, and $\Delta=\tau-\sigma$. The average over the input distribution is denoted by $\langle\cdots\rangle$ and can be performed analytically as an integration over M independent two-dimensional Gaussian densities with $\langle x_i\rangle=\langle y_i\rangle=0$, $\langle x_i^2\rangle=Q_i=\boldsymbol{J}_i\cdot\boldsymbol{J}_i$, $\langle y_i^2\rangle=1$, and $\langle x_iy_i\rangle=R_i=\boldsymbol{J}_i\cdot\boldsymbol{B}_i$ [16].

Figure 2 shows the result of a numerical integration of Eqs. (13) in the limit $\eta\to 0$ where the quadratic term is neglected. The analytical results for typical initial conditions are in good agreement with simulations of a system with $N=80$ and $\eta=0.02$.

Again, the configuration of perfect learning, $\epsilon_g=0$, is a fixed point of the dynamics (13), which is attractive for small enough learning rates. In the limit $k\to\infty$, the critical value is found to be $\eta_c\propto 1/k^4$ independent of the actual phases in the activation of different units.

In contrast, when $k\to 0$, one obtains, for instance, for $\tilde{g}(x)=g(x)=\sin(kx)\,[\cos(kx)]$ the critical rate $\eta_c\propto 1/k^4$ $(1/k^8)$, respectively. The general rule from which the scaling of η_c with small k can be obtained is

$$\eta_c\propto 1/(\nabla_J\epsilon)^2$$

for general $\boldsymbol{J}_i$. One derivative is due to the form of the learning algorithm, Eq. (1), and the second factor stems from the asymptotic expansion $\boldsymbol{J}_i\to\boldsymbol{B}_i$. This is consistent with the above result (9) for the single node.

The extension of the presented analytical approach to more general architectures, including overlapping receptive fields, different phases and wave numbers for each node, is currently studied and much more involved. Preliminary results show that, for instance, plateaus persist in overlapping machines with a nonlinear output unit.

Finally we would like to point out that in the case of periodic activations a mapping exists between different types of two-layer perceptrons. The architecture we discussed above, where the output depends on the *sum* of the hidden unit activities (soft committee machine), is equivalent to a network where the total output is a function of the *product* of its hidden units (soft parity machine [12]). For instance, for $N{:}2{:}1$ architectures and $\tilde{g}(x)=g(x)=\sin(kx)$ the output is given by $\sin\{k[\sin(kx_1)+\sin(kx_2)]\}=\sin\{2k\sin[k(x_1+x_2)/2]\cos[k(x_1-x_2)/2]\}$, which can be interpreted as a soft parity machine with weight vectors $\boldsymbol{J}_{\pm}=(\boldsymbol{J}_1\pm\boldsymbol{J}_2)/2$ and an output wave number $2k$.

I.K. acknowledges support from the Israel Academy of Science.

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

[2] M. Opper and W. Kinzel, in *Models of Neural Networks III*, edited by E. Domany, J. L. van Hemmen, and K. Schulten (Springer, Berlin, 1996).

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

[4] S. Amari, IEEE Trans. Electromagn. Compat. **16**, 299 (1967); Neurocomp. **5**, 185 (1993).

[5] W. Kinzel and P. Ruján, Europhys. Lett. **13**, 473 (1990).

[6] O. Kinouchi and N. Caticha, J. Phys. A **26**, 6243 (1992).

[7] M. Opper, Phys. Rev. Lett. **77**, 4671 (1996).

[8] C. van den Broeck and P. Reimann, Phys. Rev. Lett. **76**, 2188 (1996).

[9] J. Kim and H. Sompolinsky, Phys. Rev. Lett. **76**, 3021 (1996).

[10] M. Copelli, O. Kinouchi, and N. Caticha, Phys. Rev. E **53**, 6341 (1995).

[11] H. Sompolinsky, N. Barkai, and H. S. Seung, in *Neural Networks: The Statistical Mechanics Perspective*, edited by J. H. Oh, C. Kwon, and S. Cho (World Scientific, Singapore, 1995).

[12] M. Biehl and H. Schwarze, J. Phys. A **28**, 643 (1995).

[13] D. Saad and S. A. Solla, Phys. Rev. Lett. **74**, 4337 (1995); Phys. Rev. E **52**, 4225 (1995).

[14] M. Biehl, P. Riegler, and C. Wöhler, J. Phys. A **29**, 4769 (1996).

[15] J. Inoue, H. Nishimori, and Y. Kabashima (unpublished).

[16] P. Riegler and M. Biehl, J. Phys. A **28**, L507 (1995).