Training a perceptron in a discrete weight space

Michal Rosen-Zvi and Ido Kanter
Minerva Center and the Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel

(Received 5 February 2001; revised manuscript received 15 May 2001; published 20 September 2001)

Learning in a perceptron having a discrete weight space, where each weight can take $2L+1$ different values, is examined analytically and numerically. The learning algorithm is based on the training of the continuous perceptron and prediction following the clipped weights. The learning is described by a new set of order parameters, composed of the overlaps between the teacher and the continuous/clipped students. Different scenarios are examined, among them on-line learning with discrete and continuous transfer functions. The generalization error of the clipped weights decays asymptotically as $\exp(-K\alpha^2)$ in the case of on-line learning with binary activation functions and $\exp(-e^{|\lambda|\alpha})$ in the case of on-line learning with continuous one, where α is the number of examples divided by N, the size of the input vector and K is a positive constant. For finite N and L, perfect agreement between the discrete student and the teacher is obtained for $\alpha \propto L\sqrt{\ln(NL)}$. A crossover to the generalization error $\propto 1/\alpha$, characterizing continuous weights with binary output, is obtained for synaptic depth $L > O(\sqrt{N})$.

DOI: 10.1103/PhysRevE.64.046109 PACS number(s): 84.35.+i, 02.50.Cw, 05.20.−y

I. INTRODUCTION

The study of neural networks as a tool for understanding learning processes has benefited various applications [1,2]. We are interested in the perceptron learning ability as an archetype of feed networks that are able to learn. Most of the perceptrons that have been studied until now are under two totally different constraints, two extremes. Either the teacher weight vector is restricted to a binary space (the Ising teacher), or it is continuous, confined to a hypersphere. Only a few aspects of the learning ability of weights, which are confined to have a finite number of values, have been studied. These systems are the intermediate case, in which the weights are confined to finite space $(2L+1)^N$ when L is an integer and N stands for the input size [3–5].

The generalization ability of such networks, in which the synapse has a finite depth, has been analyzed by using replica calculations and has been found to have interesting nontrivial behavior of phase transition. The learning procedure composed of two phases: one in which the learning ability is very limited, the generalization error is finite. Another phase is when the generalization error is exactly zero, perfect learning is gained, and it occurs in a finite α, where α is the number of patterns divided by the size of the input N [5]. Nevertheless, replica calculations do not involve practical algorithms that one may use in order to obtain that learning behavior. In the Ising case, for instance, although a phase transition was predicted, no practical algorithm reproduces this discontinuous behavior [6,7].

In contrast to batch learning, when all the examples are used together to achieve perfect learning, on-line learning is a procedure in which an update rule is used and learning in each step utilizes only the last of a sequence of examples. Such an algorithm drastically reduces the computational effort compared with batch learning and no explicit storage of a training set is required [8]. It was shown that there is no updating rule that uses only the discrete vector for updating and results in perfect learning [9].

In this paper we address the issue of learning from a finite depth teacher. The method we introduce is based on the clipping of a continuous perceptron. Having an artificial continuous weight vector enables smooth learning; clipping it results in a discrete student $\vec{W}^S$, whose components are close to those of the teacher. This method has been used successfully in the Ising perceptron [6,7,10,11]. The questions that arise from the procedure above are; whether learning is possible at all and if it is possible, does it give better results than the learning in a continuous space. It seems very natural that if the weights' depth is very large, i.e., there are many possible values to each weight, the learning behavior of the discrete weights will be exactly the same as those of a continuous weight. However, in the following we examine if and what are the scaling relations between both properties L and N.

Our main results are as follows. (a) Learning in the case of finite depth is possible by using a continuous precursor. This result was confirmed both analytically and numerically. (b) On-line learning scenario: In the case of a binary output the generalization error decays superexponentially with α, $\epsilon_g \propto \exp(-K_1\alpha^2)$. whereas in the case of continuous output the generalization error decays much faster, $\exp[-K_2\exp(K_3\alpha)]$, where all the constants K_i are positive constants. (c) Perfect learning is obtained when N is very large but finite, unlike the continuous perceptrons performance. Quantitatively, for a given N and L perfect learning is achieved for $\alpha_f \propto O(L\sqrt{\ln(LN)})$. (d) A crossover to the behavior of the generalization error in the presence of continuous weights occurs for $L > o(\sqrt{N})$.

The paper is organized as follows. In Sec. II the architectures and the dynamical rules are defined as well as the continuous and discrete students. In Sec. III the order parameters are defined and the relations between the overlaps of the continuous teacher with the discrete/continuous students are derived analytically. In Sec. IV, the dynamical evolution of the order parameters in the case of binary output is derived analytically and confirmed by simulations. In Sec. V the case of large synaptic depth and the crossover to the continuous weights is studied. In Sec. VI, the perfect learning in finite N systems is examined both analytically and numerically. Section VII is devoted to analyzing results in the case of continuous output. Finally, in Sec. VIII results are concluded and open questions are addressed.

II. THE MODEL

A. The architecture

We investigate a teacher-student scenario where both nets are single-layer feed forward. The examples are generated by the so-called teacher, which is known to be restricted to a well-defined discrete set of values. We define a synaptic depth L and a set of discrete values as follows [3,5]:

$$W_i^T = \pm\frac{1}{L}, \pm\frac{2}{L} \cdots \pm 1. \tag{1}$$

When the zero value is part of the game, the possible values of the weights are

$$W_i^T = 0, \pm\frac{1}{L}, \pm\frac{2}{L} \cdots \pm 1. \tag{2}$$

For the sake of simplicity in this paper we present results only for including zero case [Eq. (2)]. It is easy to generalize our results to any other set including the one presented in Eq. (1), which converges to the Ising case when $L=1$.

The components of the input patterns $\vec{\xi}_i^{\mu}$ are independent random variables. In the following they are drawn from a Gaussian distribution with zero mean and unit variance. The size of the teacher, the student, and the input is N. For any input $\vec{\xi}$ the so-called teacher generates an output S according to a rule

$$S = F\left(\frac{\vec{W}^T \cdot \vec{\xi}}{\sqrt{N}}\right). \tag{3}$$

In the following we discuss both binary and continuous rules. The student has in mind the rule F and the discrete set of values that the teacher is confined to. In addition, in an on-line learning scenario, the student is given in each time step μ, the input $\vec{\xi}^{\mu}$, and the teacher's output S^{μ}, whereas in batch learning the set $(\vec{\xi}^{\mu}, S^{\mu})$ $\mu = 1 \dots \alpha N$ is given altogether.

B. Dynamics of the weights

A continuous precursor for the student $\vec{J}$ is needed for learning from a discrete teacher. The learning procedure, having a continuous student, is well known. In an on-line scenario, at each step the continuous student updates its weight vector according to some learning algorithm f. The generic form of the learning algorithm is

$$\vec{J}^{\mu+1} = \vec{J}^{\mu} + \frac{\eta}{\sqrt{N}} f(S^{\mu}, x_J^{\mu}) \vec{\xi}^{\mu} S^{\mu}, \tag{4}$$

where η is the learning rate and x_J is the student's local field, $x_J \equiv 1/\sqrt{N}\vec{J} \cdot \vec{\xi}$. Such a learning algorithm means that at each learning step μ, the current weight vector $\vec{J}^{\mu}$ is updated according to the new example $\vec{\xi}^{\mu}$ and each example is presented only once.

In an off-line scenario, there is a set of examples $\vec{\xi}^{\mu}$ $\mu = 1 \dots \alpha N$ and they are used altogether to gain perfect learning. There are methods in which the off-line learning is made according to a rule that defines an additive quantity of all the examples. Such procedures were shown to end up in perfect learning [12,13]. Since having a discrete teacher is merely a special case, not using the knowledge that the teacher is confined to a discrete set of values gives the well-known results; an exponential decay in the case of continuous rule (on-line learning [14,15]) and a power law decay in the case of binary rule (on-line and off-line learning [12,13,16–18]).

The way to gain from the knowledge of the discrete nature of the weights is at the center of our work, and is based on having in addition a discrete student $\vec{W}^S$ derived from the continuous one using the following clipping procedure. A continuous weight is clipped to the nearest discrete value, among the $2L+1$ possibilities. Such a clipping procedure is the optimal one with the lack of any prior knowledge about the weights except that each value appears with the same probability. We define limit values λ_l, which are arranged in an increasing order. The limit values divide the continuous region of the precursor weight vector components into $2L+1$ intervals, according to the number of the available values as in Eq. (2). The clipping process is such that J_i is mapped onto l/L for $J_i \in (\lambda_l, \lambda_l + 1)$. The set of limits includes $\{\lambda_{-l}, \lambda_{-l-1}, \dots \lambda_{-1}, \lambda_0, \lambda_1 \dots \lambda_{l+1}\}$. It is given by the following mathematical rule:

$$W_i^S = \sum_{l=-L}^{L} \frac{l}{L} [\theta(\lambda_{l+1} - J_i) - \theta(\lambda_l - J_i)], \tag{5}$$

where θ is the Heaviside function.

Since the value of those limits λ_l is somewhat unclear, we would like to exemplify it with some specific cases. In the case of $L=1$, Eq. (1), for instance, due to symmetry it is obvious that the limit between -1 and 1 should be 0. Hence, one introduces the following limits: $\lambda_{-1} = -\infty$, $\lambda_0 = 0$, $\lambda_1 = \infty$. Evaluating the mapping equation results in the well-known clipping rule, $W_i^S = \mathrm{sgn}(J_i)$, [6,10]. Finding the appropriate value for all other cases but the Ising perceptron becomes more complicated, the continuous space is no longer divided into two clear regions and hence one has to consider carefully the value of the limits.

In this paper we chose to nail down the general results by focusing on the including zero case, $L=1$, i.e., $W_i = 0, \pm 1$. This case is known as the diluted Ising case and some other aspects of it have been studied in Refs. [19,3,20]. It contains the simplicity of the Ising case on the one hand and introduces more generality concerning discrete values on the other hand. In this case, there is only one unknown parameter λ_1 since $\lambda_2 = -\lambda_{-1} = \infty$ and $\lambda_0 = -\lambda_1$.

While choosing the value of the limits, (in the last case it means choosing only the value of λ_1) one should take into consideration the *a priori* knowledge about the weights of teacher. It is clear that the limits should scale with the student norm, since the exact set of values that the continuous student ends up with is irrelevant. The mapping rule ensures that the discrete student ends up with the same values as those of the teacher. This will be shown only after analyzing the new order parameters and their dependence on the former one, as presented in the following section.

III. THE ORDER PARAMETERS

Evaluating the agreement between teacher and student is done by calculating either the generalization error or the order parameters. The generalization error ϵ_g is calculated by taking the average of the student/teacher disagreement over the distribution of input vectors. The generalization error is given, in principle, by the overlaps between the vectors, (the so-called order parameters). However, in order to go into details one has to first define the rule, [F in Eq. (3)]. This will be done in the following sections. In the following we concentrate on introducing the complete set of order parameters and their inter-relations.

In our case there are three vectors and hence two interdependent sets of order parameters. One set concerns the continuous overlaps,

$$R_J \equiv \frac{1}{N}\vec{J}\cdot\vec{W}^T,$$

$$Q_J \equiv \frac{1}{N}\vec{J}\cdot\vec{J}, \tag{6}$$

and the other set concerns the discrete vector's overlaps,

$$R_W \equiv \frac{1}{N}\vec{W}^S\cdot\vec{W}^T,$$

$$Q_W \equiv \frac{1}{N}\vec{W}^S\cdot\vec{W}^S. \tag{7}$$

We note that the dynamical evolution of the continuous set of order parameters, Eq. (6), is independent of the clipped order parameters, since the *training* is done only following the continuous weights. Contrary to the training process, the *prediction* of the generalization properties is made following the clipped student. Hence, finding the quantitative interplay between the continuous set of order parameters, Eq. (6), and the discrete set of order parameters, Eq. (7), is the cornerstone for the analytical description of the generalization ability of the student.

In this section we examine the relationship between the clipped set and the continuous one. The development of R_J and Q_J is not influenced by the clipping method. Hence, examination of the above relationship enables us to determine the development of the clipped order parameters and results in a description that provides the whole picture of the learning process.

The teacher's norm is determined according to the *a priori* probabilities for each discrete value. Having equal probability and taking the thermodynamic limit results in the norm

$$T \equiv \frac{1}{N}\vec{W}^T\cdot\vec{W}^T = \frac{1}{L^2 n_L}\sum_{l=1}^{L} l^2 = \frac{1}{3}+\frac{1}{3L}, \tag{8}$$

where n_L is defined as the number of optional values $n_L = 2L+1$. The order parameters in the clipped machines R_W and Q_W as a function of those of the continuous machine R_J and Q_J are evaluated as follows:

$$R_W = \left\langle \frac{1}{N}\sum_i W_i^T \frac{l}{L}[\theta(\lambda_{l+1}-J_i)-\theta(\lambda_l - J_i)]\right\rangle,$$

$$Q_W = \left\langle \frac{1}{N}\sum_i \frac{l^2}{L^2}[\theta(\lambda_{l+1}-J_i)-\theta(\lambda_l - J_i)]\right\rangle, \tag{9}$$

where $\langle A\rangle$ is an average over the known constraints and the known overlaps

$$\langle A\rangle \equiv \frac{\mathrm{Tr}_{W^T}\int dJ_i\,\delta(J_i^2-NQ_J)\,\delta(J_iW_i^T-NR_J)A}{\mathrm{Tr}_{W^T}\int dJ_i\,\delta(J_i^2-NQ_J)\,\delta(J_iW_i^T-NR_J)}, \tag{10}$$

and the summations are over all the possible values of l, starting from $l=-L$, $-L+1,\dots,L$, and over $i = 1\dots N$. The validity of this average is based on the assumption that all vectors $\vec{J}$ that are consistent with the constraints are taken with equal probability. This assumption is violated when the updating of the continuous vector itself is made according to the clipped one, (see [6,11]).

The results are

$$R_W = \frac{1}{2L^2 n_L}\sum_l l'[\mathrm{erf}(\Phi_{l+1,l'})-\mathrm{erf}(\Phi_{l,l'})],$$

$$Q_W = \frac{1}{2L^2 n_L}\sum l^2[\mathrm{erf}(\Phi_{l+1,l'})-\mathrm{erf}(\Phi_{l,l'})], \tag{11}$$

where the summation is over all the possible values of l,l', and we define

$$\Phi_{l,l'} \equiv \frac{\dfrac{\lambda_l}{\sqrt{Q_J}}-\dfrac{\rho_J}{\sqrt{T}}\dfrac{l'}{L}}{\sqrt{2(1-\rho_J^2)}}, \tag{12}$$

where $\rho_J \equiv R_J/\sqrt{T}\sqrt{Q_J}$, $\rho_W \equiv R_W/\sqrt{T}\sqrt{Q_W}$ are the geometrical order parameters.

In the limit $L\to\infty$ the summation in Eq. (11) can be replaced by an integral. Calculating the integrals in this limit results in the obvious identities $R_W = R_J$ and $Q_W = Q_J$. Note that taking integrals instead of summation imposes an inequality. The difference $\Phi_{l,l'}-\Phi_{l+1,l'}$ tends to zero as long as $L\gg 1/\sqrt{1-\rho_J^2}$, [see Eq. (12)]. Hence, in the event that L is very large, learning with the continuous student or learning with the clipped version produces the same result as long as ρ_J is smaller than $2/L$. This limit is discussed in Sec. VI.

We exemplify the general results in the case of the diluted Ising perceptron. In this case we use the following limits:

$$\lambda_2=-\lambda_{-1}=\infty,$$

$$\lambda_1=-\lambda_0, \tag{13}$$

and the teacher's norm is $T=2/3$. The mapping above gives

$$R_W=\frac{1}{3}[\operatorname{erf}(A_+)+\operatorname{erf}(A_-)],$$

$$Q_W=1-\frac{1}{3}\operatorname{erf}(A_0)+\frac{1}{3}\operatorname{erf}(A_-)-\frac{1}{3}\operatorname{erf}(A_+), \tag{14}$$

where $A_\pm=(\rho_J/\sqrt{T}\pm\lambda_1/\sqrt{Q_J})/\sqrt{2(1-\rho_J^2)}$ and $A_0=\lambda_1/\sqrt{2Q_J(1-\rho_J^2)}$.

From Eq. (14) one can verify that at the limit $\alpha\to\infty$ when the continuous order parameters achieve perfect learning, $\rho_J\to 1$, the discrete order parameters achieve perfect learning as well, $R_W\to 2/3$, $Q_W\to 2/3$, and $\rho_W\to 1$ given that the positive quantity λ_1 is smaller than $\lambda_1<\sqrt{Q_J/T}$.

In general, in order that the discrete student will gain perfect learning it is necessary that the relation $\sqrt{Q_J/T}(l-1)<\lambda_l<\sqrt{Q_J/T}l$ holds for any positive l. Note that the interpretation of the above constraint is that in the vicinity of perfect learning the precursor might be focused around any set of discrete symmetric values, but not necessarily the ones that the clipped student has.

One of the conclusions concerning λ_l is that the law according to which ϵ_g decays is independent of the exact value of the limit value λ_l. It depends only on the ruler (binary/continuous), the specific strategy of learning (on-line/off-line), and the learning algorithm one uses. In the following we analyze all these variations.

IV. BINARY OUTPUT

In an on-line learning scenario one can write equations of motion that determine the development of the order parameters as a function of α. The rate of convergence depends on the rule, F [Eq. (3)] and the learning algorithm that one uses f [Eq. (4)]. Fine tuning is achieved by choosing the learning rate η.

We analyze learning procedure in the case of binary rule,

$$S=\operatorname{sgn}(x), \tag{15}$$

where x is the local field and the generalization error as a function of ρ is known to be

$$\epsilon_g=\frac{1}{\pi}\cos^{-1}(\rho). \tag{16}$$

Although it was shown that using the ''expected stability'' algorithm that maximizes the generalization gain per example leads to an upper bound for the generalization ability [17], we concentrate on the so-called AdaTron or relaxation learning algorithm. The reason is that this latter algorithm in a specific case (for zero stability, $\kappa=0$) performs comparably well. Moreover, unlike the ''expected stability'' algorithm it does not require additional computations in the student network besides the updating of its weights, and the analysis is simpler [8].

The convergence to perfect learning depends on the learning rate. If it is too large, perfect generalization becomes impossible. The transition from a learnable situation to unlearnable occurs at η_c. In the following, in order to simplify the analysis, we choose a fixed learning rate $\eta=1$, which is below η_c in all scenarios.

We update the artificial continuous weight vector $\vec{J}$. The updating is made as in Eq. (4) according to the following learning rule:

$$J_i^{\mu+1}=J_i^{\mu}-\frac{\eta}{\sqrt{N}}\left(\frac{\vec{J}^{\mu}\cdot\vec{\xi}^{\mu}}{\sqrt{N}}\right)\xi_i^{\mu}\,\theta\left(-\frac{\vec{J}^{\mu}\cdot\vec{\xi}^{\mu}}{\sqrt{N}}S^{\mu}\right). \tag{17}$$

The equations for the order parameters with $\eta=1$ are

$$\frac{d\rho_J}{d\alpha}=-\frac{\rho_J}{2\pi}\cos^{-1}(\rho_J)+\frac{1}{\pi}\left(1-\frac{\rho_J^2}{2}\right)\sqrt{1-\rho_J^2},$$

$$\frac{dQ_J}{d\alpha}=\frac{Q_J}{\pi}[\rho_J\sqrt{1-\rho_J^2}-\cos^{-1}(\rho_J)]. \tag{18}$$

In the limit $\alpha\to\infty$, one can expand the right-hand side of the first equation around $\rho_J=1$. The next step is to plug the result of $\rho_J(\alpha)$ up to the first order corrections in α in the second equation. One can find the following power law:

$$\rho_J\sim 1-2\left(\frac{3\pi}{4}\right)^2\frac{1}{\alpha^2},$$

$$Q_J\sim Q_0\left(1-\pi^2\left(\frac{3}{4}\right)^3\frac{1}{\alpha^2}\right). \tag{19}$$

Note that in the case of a binary output unit, perfect learning is achieved as soon as the angle between the vectors goes to zero, independent of the student's norm.

The solution of Eq. (18) describes only the development of the continuous perceptron's overlaps. The next step is to map the continuous precursor to the clipped one as defined by Eq. (11). Since in the case of binary ruler the student's norm converges to some unknown value, it seems only natural to choose a limit set λ_l that scales with $\sqrt{Q_J}$. As a result ρ_W, $[R_W/\sqrt{Q_W}T$, see Eq. (11)] is only a function of ρ_J and does not depend on Q_J. Hence, substituting the asymptotic behavior of ρ_J [Eq. (19)] into ρ_W, one can find the typical asymptotic behavior of ρ_W. In general, the clipped order parameter ρ_W is composed of a devision between two different sums of error functions. The argument of each error function consists of $1/(1-\rho_J^2)$. Asymptotically $1-\rho_J \rightarrow 1/\alpha$ and the first correction to the error function scales superexponential with α, $\exp[-K(\lambda,L)\alpha^2]$ where $K(\lambda,L)$ is independent of α. The leading correction of ρ_W is determined by $K-\min(\lambda,L)$ over all error functions. Finally, the generalization error in the limit $\alpha\rightarrow\infty$ is given by $\epsilon_g \propto \sqrt{(1-\rho_W)/2}/\pi$ [see Eq. (16)], and hence

$$\epsilon_g \propto \frac{\exp[-K(\lambda,L)\alpha^2]}{\alpha^{\frac{1}{2}}}, \qquad (20)$$

where $K(\lambda,L)$ is determined by the minimal value upon all $|\lambda_l - l/L\sqrt{Q_J/T}|$, [for a specific example see Eq. (24)].

One way of choosing λ_l is simply "half the way" between the constrained values, i.e., $\lambda_{-L}=\lambda_{L+1}=\infty$ and otherwise

$$\lambda_l = \frac{1}{L}\left(l-\frac{1}{2}\right)\sqrt{\frac{Q_J}{T}}. \qquad (21)$$

In the case where the limits are defined as in Eq. (21) it is possible to calculate the asymptotic decrease of the generalization error for *any* given depth L,

$$K(\lambda,L) = \frac{1}{12\pi^2(L^2+L)}. \qquad (22)$$

We exemplify the aforementioned discussion in the diluted Ising perceptron. We use the limits as in Eq. (13) and assume $\lambda_1 = c\sqrt{Q_J/T}$. In that case

$$\rho_W = \frac{\mathrm{erf}(a_+)+\mathrm{erf}(a_-)}{\sqrt{T}\sqrt{9-3\,\mathrm{erf}(a_0)-3\,\mathrm{erf}(a_+)+3\,\mathrm{erf}(a_-)}}, \qquad (23)$$

where $a_\pm = \rho_J \pm c/\sqrt{2T(1-\rho_J^2)}$ and $a_0 = c/\sqrt{2T(1-\rho_J^2)}$. In the limit of large α one finds

$$\epsilon_g \propto \frac{\exp(-b_c\alpha^2)}{\alpha^{1/2}}, \qquad (24)$$

where for $c \geqslant 1/2 b_c = c^2/6\pi^2$ and otherwise $b_c = (1-c)^2/6\pi^2$. One can see that choosing $c=1/2$ results in the fastest decay of the generalization error.

The analytical results are compared with simulations in the case of a teacher of the type of the diluted Ising perceptron with the following parameters; $\lambda = 0.5\sqrt{Q_J/T}$ and $\lambda = 0.3\sqrt{Q_J/T}$, see Fig. 1. The initial conditions for the continuous student weight vector are $Q_J(\alpha=0)=T=2/3$ and $R_J(\alpha=0)=0$. The weight components were drawn out of a Gaussian distribution. We used $\eta=1$, $N=3000$ and each point was averaged over 50 samples. One can see in Fig. 1 that the analytical results given by Eq. (24) are in agreement with simulations.

One can see that the superexponential decay is independent of the accurate value of λ. However, two important parameters do depend on the exact choice of λ. One is the decay rate, the factor $K(\lambda)$ in the large α limit. One can see, for instance, that the optimal limit, $\lambda=0.5$, results in a faster decay than the limit $\lambda=0.3$. The second is the exact α or the exact value of ρ_J at which the clipped version gives a better result than the continuous one. We named this value as ρ_T. For $\rho_J<\rho_T$ the clipping lowers the overlap ρ_J since the learning solution does not contain enough information about the real direction of the teacher $\vec{W}^T$ so that clipping only leads the solution to "forget" a little about the learned pattern without bringing it closer to the exact solution. In the other region, when $\rho_J>\rho_T$, clipping becomes efficient because the learning solution is near the exact one. The numerical results of ρ_T according to the mapping, [Eq. (23)], are $\rho_T \sim 0.92$ for $\lambda_1 = 0.5\sqrt{Q_J/T}$ and $\rho_T\sim 0.97$ for $\lambda_1 = 0.3\sqrt{Q_J/T}$, see Fig. 1.

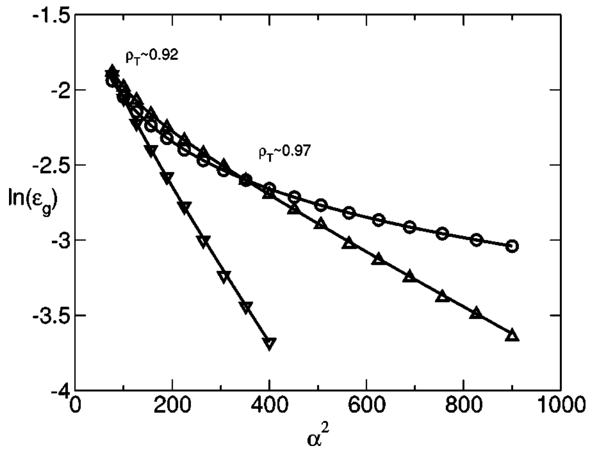

FIG. 1. Simulation results of $\ln(\epsilon_g)$ of the continuous precursor (○) and of the clipped vector vs α^2. The clipping is made according to the mapping in Eq. (13), where the results are for $\lambda_1 = 0.5\sqrt{Q_J/T}$ (∇) and $\lambda_1 = 0.3\sqrt{Q_J/T}$ (△). Error bars are smaller than symbols. Solid lines are the numerical integrals [Eq. (18)]. ρ_T refers to the point at which a transition occurs between a superior performance by the continuous/clipped perceptron (see text).

V. LARGE SYNAPTIC DEPTH

In this section we examine the crossover of the generalization error in the presence of continuous weights as we increase the synaptic depth. As long as the synaptic depth $L<O(\sqrt{N})$, the generalization error still vanishes superexponentially, Eq. (20), where the prefactor decreases with L. For $L\geqslant O(\sqrt{N})$ the learning is characterized by the features of spherical constrained learning.

The first step towards the continuous case limit is to find out the change of the decay of the generalization error as a function of L. We focus on the binary unit in the on-line scenario. The analytic tractability of this model enables a profound study of the influence of the synaptic depth over the learning features.

In the last model the generalization decays superexponentially, $\epsilon_g \sim \exp(-K\alpha^2)$, Eq. (20). The factor K depends on the limits one chooses λ_l. Hence, in order to maintain consistency, we use the abovementioned limits, Eq. (21), and K is given by Eq. (22). We should emphasize that only one dominated term out of many superexponential terms arising from

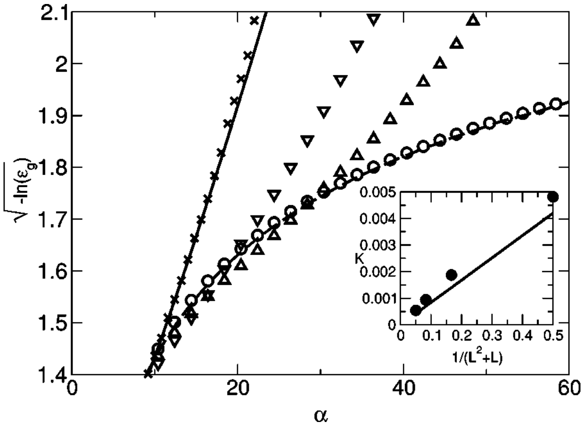

FIG. 2. Simulation results of $\sqrt{-\ln(\epsilon_g)}$ in the case of $L=1$ (diluted Ising) ($\times$), $L=2$ (∇), $L=3$ ($\triangle$), and $L=157$ ($\bigcirc$) vs α. The analytical result obtained by the numerical integration of Eq. (18) and Eq. (23) is presented for the diluted Ising case (solid line). The dashed line is the analytical curve for $\sqrt{-\ln(\epsilon_g^J)}$, where ϵ_g^J is the generalization error of the *continuous* student. Inset: the dependence of the prefactor $K(L)$ vs $1/(L^2+L)$. Simulation results (circles) and analytical results (solid line) are following Eq. (22).

the asymptotic expansion of all the error functions [Eq. (14)], was kept in Eq. (20). As soon as the deviations between different factors in the exponent are too small, one has to integrate all the terms together instead of neglecting all but one. Such a procedure results in a crossover from a superexponential decay to power law behavior.

Analytical and simulation results of the generalization error in varieties of synaptic depths are presented in Fig. 2. Simulations were carried out with $N=630$ and each point is averaged over 100 samples. The inset shows the estimated slope K as a function of the depth L. The solid line is the analytical results, Eq. (22), and the circles are the slopes estimated by simulations for $L=1,2,3,4$. The deviation from the analytical curve is probably due to higher order corrections in α. Note that only at the very end of the learning procedure, the linearity of $\ln(\epsilon_g)$ in α^2 can be achieved. In addition, at this stage of learning ($\alpha \gg 1$), one has to bear in mind deviations due to finite size corrections in N.

We now present an argument supporting the statement that the generalization performance of finite depth machines coincides with the performance of continuous machines as soon as $L \sim \sqrt{N}$. This scaling is found by taking into account that: (a) the difference between two available values is of order of $1/L$; (b) the distribution of the continuous student values around the teacher's value is a Gaussian with a variance of $\sqrt{1-\rho_J^2}=\epsilon_g^J$, where ϵ_g^J is the generalization error of the continuous student. Having a learning procedure (in the continuous space) in a finite dimension results in a generalization error ϵ_g^J, which is different from the analytical predictions. The variance is of order $\sqrt{1/N}$ [21]. Hence, an estimation to the order of the lower value that ϵ_g^J gets in a specific run will be $\sqrt{1/N}$. As a consequence, having a discrete machine of depth L when

$$\frac{1}{L} \ll \sqrt{1-\rho_J^2} \sim \epsilon_g^J \sim \sqrt{\frac{1}{N}} \tag{25}$$

or $L \gg \sqrt{N}$, gives the same results as those of continuous learning. Note that Eq. (25) is consistent with the mathematical constraint that was pointed out in Sec. III when we discussed the continuous limit [after Eq. (11)]. Indeed, the simulations show indeed that in the case of $L=157 \gg \sqrt{N}$, where $N=630$, the discrete vector's performance coincides with the analytical learning curve of the *continuous* student.

It is worth pointing out that a similar result was found when analyzing the possibility of learning from a discrete teacher by a discrete student using a general updating rule [9]. The last analysis uses a totally different argument, resulting in the conclusion that only when the teacher's depth is of order $\sqrt{N}$, it is possible to learn the rule using an updating rule that depends on the discrete weights, i.e., only then it behaves as if we have a continuous machine.

VI. FINITE SYSTEMS—PERFECT LEARNING

The theoretical results presented in the previous sections exhibit the typical behavior of the generalization error and the order parameters. The main result is the fast decay of the generalization error of the clipped perceptron to zero, Eq. (20). In the case of teacher and student with continuous weights and finite N, the generalization error is always finite distance from zero, even in the asymptotic stage of the learning process. In contrast to the continuous case, the learning of a perceptron with discrete weights and finite N is characterized by a transition to perfect learning, as was found for the Ising perceptron [11]. Performing simulations in that case results in a perfect learning at some stage, since in the clipping version the student knows exactly the teacher's optional values. Hence, the overlap becomes exactly one, $\rho_W=1$, and the generalization error becomes exactly zero as well, $\epsilon_g=0$.

In order to estimate the number of steps needed for perfect learning α_f we use the analytical approximation valid in the $\alpha \to \infty$ regime. At that regime we have an analytical approximation of the interdependence of ϵ_g and α [Eq. (20) and Eq. (31)]. In addition, the minimal step before perfect learning is well defined: $\rho_W=1-2/(LN)$ or $\epsilon_g \sim \sqrt{1/(LN)}$. Hence, we can find the interplay between α and N.

It was shown that except for some special cases such as trapping in symmetric phases in MLN, [22,23] the analytical equations for the development of the order parameters are accurate in the leading order. Finite dimension N affects the deterministic equations for the mean values of the order parameters by having broader distributions for the order parameters, and the covariances scale as $1/N$. Moreover, extensive numerical simulations show that the finite size corrections to ρ_J scale with $1/N$ [24,25,21] and hence they are negligible.

In the binary output perceptron the generalization error drops superexponentially, Eq. (20). Hence, perfect learning is determined by

$$\exp[-K(\lambda,L)\alpha^2] \sim \sqrt{1/(LN)}. \tag{26}$$

If one uses the set of limits as in Eq. (21) then the dependence of K on L is given by Eq. (22). Deriving α_f from the last equation results in, $\alpha_f \sim L\sqrt{\ln LN}$. This result indicates

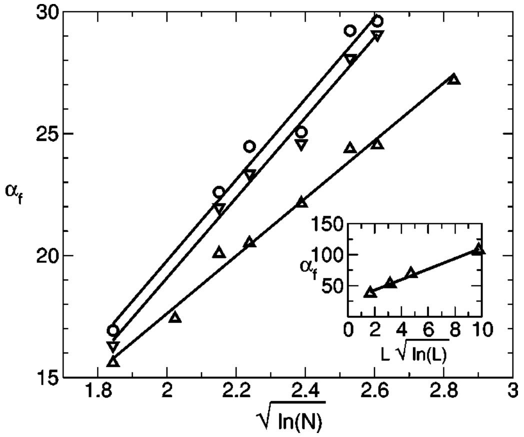

FIG. 3. Simulation results of α_f, the number of rescaled steps necessary to achieve perfect learning vs $\sqrt{\ln N}$. Simulations for the diluted Ising perceptron, in the case of a binary output unit, with $\lambda_1 = 0.4\sqrt{Q_J/T}$ (∇), $\lambda_1 = 0.5\sqrt{Q_J/T}$ ($\triangle$), and $\lambda_1 = 0.6\sqrt{Q_J/T}$ ($\bigcirc$). Solid lines correspond to the linear fit of least squares error. Inset: Simulation results of α_f vs $\sqrt{L \ln L}$ for $N = 630$, $L = 2,3,4,7$ and the limit values are chosen according to Eq. (21). The solid line is the least squares fit.

quantitatively that for any chosen limit λ_l, the number of learning steps necessary to achieve perfect learning is finite as long as N and L are finite.

Figure 3 presents results of α_f obtained in simulations for the diluted Ising perceptron with $c = 0.4$, $c = 0.5$, and $c = 0.6$, [Eqs. (23) and (24)]. Results where averaged over $M(N)$ training sets, were values of $M(N)$ ranging from 5000 to 20 in accordance to N that is varied between 30 and 9000. To get results in a lower dimension N, we averaged over a larger number of simulations M.

One can see from the obtained values of $\alpha_f(N,c)$ in Fig. 3 that the last quantity is indeed linear in $\sqrt{\ln N}$. Note that the obtained slope in Fig. 3 for $c = 0.4$ and $c = 0.6$ is the same as we expect, since b_c is symmetric around $c = 1/2$. Indeed, one can see in the inset that $\alpha_f(L)$ in the case of $N = 630$, increases linearly with $L\sqrt{\ln L}$. As $L \to \infty$ an infinite number of examples are needed for perfect learning, there is a crossover to the spherical case as discussed in the previous chapter.

Small deviations from a straight line in Fig. 3 are expected to be a consequence of the following approximations:

(a) We took as an analytical curve [Eq. (26)] only the asymptotic function that is an expansion valid in infinite α.

(b) We neglected the polynomial corrections in Eq. (26) such as $1/\sqrt{\alpha}$.

(c) We derived Eq. (26) from the analytical calculation of $\rho_J(\alpha)$. The latter quantity itself is influenced by finite size effects as explained above.

As was shown in previous sections, $c = 0.5$ gives the best performance in the asymptotic learning procedure, lower α_f for all N, and is confirmed in our simulations, Fig. 3. In the thermodynamic limit $N \to \infty$, $\alpha_f \to \infty$ as expected.

VII. CONTINUOUS UNIT

We now study the case of continuous output perceptrons with finite depth. As long as one uses a continuous activation function, the generalization error decreases exponentially, (see for instance [14,15,18]). In order to learn a rule that is defined by a finite depth vector, we used a spherical vector for the student weight vector $\vec{J}$ and clipped it in order to have a discrete student weight vector $\vec{W}^S$. The updating of the spherical student weight vector is done according to the gradient descent method,

$$\vec{J}^{\mu+1} = \vec{J}^{\mu} - \frac{\eta}{\sqrt{N}} \nabla_{\vec{J}} \;\; \epsilon(\vec{J}^{\mu}, \vec{\xi}^{\mu}). \tag{27}$$

The error $\epsilon(\vec{J}^{\mu}, \vec{\xi}^{\mu})$ measures the deviation of the student from the teacher's output for a particular input $\vec{\xi}$. The generalization error of a student is defined as the averaged error

$$\epsilon_g = \left\langle \frac{1}{2} [S(\vec{J}, \vec{\xi}) - S(\vec{W}^T, \vec{\xi})]^2 \right\rangle_{\vec{\xi}}. \tag{28}$$

Since the learning features of all kinds of continuous transfer functions are more or less the same, we chose to concentrate on the ‘‘sin’’ activation function

$$S = \sin(kx). \tag{29}$$

The periodic activation function sin was found to be learnable given that the period k is small enough [15]. In the following we will simplify our analysis by taking $k = 1$ and the learning rate $\eta = 1$. Since the learning curves of the continuous version are the same as if there was a rule defined by a continuous teacher (having the finite depth limitation is merely a special case of the spherical constraint) and the learning rate we chose is small enough, we find that perfect learning is an attractive fixed point in both scenarios.

Linearizing the equations of motion around these fixed points results in the following form (which holds for all continuous transfer functions):

$$R_J = 1 - \frac{c_1}{\det V} V_{22} \exp(\gamma_1 \alpha) + \frac{c_2}{\det V} V_{12} \exp(\gamma_2 \alpha),$$

$$Q_J = 1 + \frac{c_1}{\det V} V_{21} \exp(\gamma_1 \alpha) - \frac{c_2}{\det V} V_{11} \exp(\gamma_2 \alpha). \tag{30}$$

The matrix $V(1,1)$ arises from the linearization, $d/d\alpha \;\; (R,Q)^{\top} = V(1,1)(1-R, 1-Q)^{\top}$ where γ_1, γ_2 are its eigenvalues and both are negative. The constants c_1, c_2 are determined from the numerical solution of the equations of motion.

In order to get a description of the discrete learning one has to use the mapping relations as in Eq. (5). The generalization error of the finite depth student depends directly upon the order parameters as can be found by taking the average over the local field distributions Eq. (28). The general result of this calculation at the $\alpha \to \infty$ regime is

$$\epsilon_g \sim \exp(-C_0 e^{|K|\alpha}), \tag{31}$$

where K and C_0 depend only on the learning rate η, the limits one chooses λ_l and the specific activation function.

The explicit expression of C_0 and K for the sin activation function and $\eta=1$ is give below. The equations of motion are

$$\frac{dR_J}{d\alpha}=\frac{1}{2}[(R_J+1)D_+-2R_Je^{-2Q_J}-(R_J-1)D_-],$$

$$\frac{dQ_J}{d\alpha}=[(R_J+Q_J)D_+-2Q_Je^{-2Q_J}-(Q_J-R_J)D_-]+\frac{1}{8}[2(e^{-2Q_J}-e^{-2}-E_-+D_+)+3-D_-^4-2D_-+(2E_+-e^{-8Q_J}-D_+^4)], \tag{32}$$

with $D_\pm=\exp[-(1+Q_J\pm 2R_J)/2]$ and $E_\pm=\exp[-(1+9Q_J\pm 6R_J)/2]$. As $\alpha\to\infty$, one gets two eigenvalues, $\gamma_1\sim-0.30$ and $\gamma_2\sim-0.69$. Using Eq. (14), rescaling R_W and Q_W by the teacher's norm 2/3, and taking the limit values λ_l to be as defined in Eq. (21). Collecting everything we have the leading order correction in the limit $\alpha\to\infty$,

$$R_W\sim 1-\frac{\exp(-0.15\alpha)}{2\sqrt{\pi}K_1}\exp(-K_1^2e^{0.30\alpha}),$$

$$Q_W\sim 1+\frac{\exp(-0.15\alpha)}{\sqrt{\pi}K_1}\exp(-K_1^2e^{0.30\alpha}), \tag{33}$$

where

$$K_1^2=\frac{c^2}{L^2+L}, \tag{34}$$

and c is a rescaled constant determined by the initial conditions only. The generalization error as a function of the discrete parameters is

$$\epsilon_g=\frac{1}{2}\left[1-d_-+d_+-\frac{1}{2}(e^{-2Q_W}+e^{-2})\right], \tag{35}$$

with $d_\pm=\exp[-(1+Q_W\pm 2R_W)/2]$. Expanding the last equation around $R_W\to 1$ and $Q_W\to 1$, we obtain that the generalization error decays very fast,

$$\epsilon_g\sim\exp(-K_1^2e^{0.30\alpha}). \tag{36}$$

We performed simulations in the diluted Ising case $L=1$ and in the case of $L=2$. Results are averaged over 10 samples and $N=3000$. In Fig. 4 the development of the discrete as well as the continuous order parameters as a function of α in the case of $L=1$ are presented. The solid lines are the analytical numerical integrals of Eq. (32). Note that the transition in this scenario is from a poor generalization of the clipped version compared to that of the continuous one, to a situation in which the clipped version has a better performance and it occurs in the same $\rho_T\sim 0.92$ as in the binary unit. This quantity is related to the clipping rule and is independent of the specific transfer function one tries to learn.

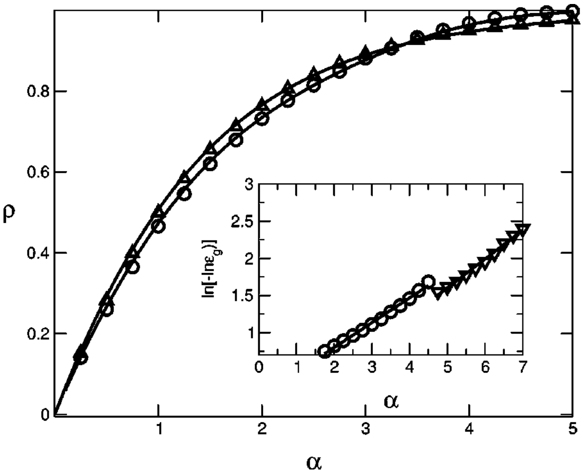

FIG. 4. Simulation results of $\rho_J(\triangle)$ and ρ_W $(\bigcirc)$ vs α in the diluted Ising case. Solid lines are the numerical integrals [Eqs. (14) and (32)]. Inset: $\ln[-\ln(\epsilon_g)]$ vs α obtained in simulations for $L=1$ (circles) and $L=2$ (triangles) with $N=3000$. Solid lines are the least squares fit. The slope was found to be 0.33 in the case of $L=1$ and 0.38 in the case of $L=2$.

The inset of Fig. 4 shows the decay of the generalization error for $L=1$ (circles) and $L=2$ (triangles). We plotted $\ln(-\ln\epsilon_g)$ as a function of α and according to the above analysis, Eqs. (34) and (36), the slope of the linear curve should be independent of L and equal to 0.30, whereas the constant in the linear formula depends on L. We obtained in simulations for $L=1$, 0.33 ± 0.01, and for $L=2$, 0.38 ± 0.01. Considering the fact that we are dealing with an approximation that is valid only in the $\alpha\to\infty$ and simulations obtained are at finite α, the results are comparable with analytical predictions. The generalization error of the clipped version for larger α ($\alpha>7$ in our case) gives better results than those predicted by the analysis. Its values are exactly zero due to the finite size effects discussed in Sec. V.

Following the same arguments used in order to find an estimation of the number of examples needed for gaining perfect learning, one finds that in the case of continuous output $\alpha_f\sim\ln(\ln N)$. It is obvious from the analytical calculations and the above simulations that clipping a continuous vector in order to learn a finite depth teacher results in extremely fast learning. The learning in finite dimension is characterized by α_f, above which one gets perfect learning of the discrete vector. All of these unique characteristics of discrete learning disappear as soon as the weight depth is of the order of $\sqrt{N}$, as found in Sec. VI.

VIII. CONCLUSIONS

In this paper, we presented an analysis of the simplest neural network, the perceptron, that learns from examples given by another perceptron, the teacher, which is confined to a discrete space. In fact, we used two students; a continuous precursor and its clipped version.

We analyzed the new set of order parameters arising from the clipping method. We discussed the issue of how to clip and what set of limits λ_l is the best choice. We found that it depends specifically on the kind of optimization one imposes. We showed that during the very first step after reaching some overlap ρ_T, a transition occurs and the clipped version results in a better performance then the nonclipped one, i.e., the benefit from the clipping is evident only after the learning is nearly accomplished, after gaining large ρ_J. For optimizing the learning time, by means of minimizing the generalization error for a given finite α, the best value is given by minimization of ρ_W with respect to λ. In the diluted Ising perceptron, for instance, the optimal value for better performance around $\rho_J \sim 0.9$ was found to be $\lambda_1 \sim 0.425\sqrt{Q_J/T}$. These results suggest that it is possible to optimize the generalization error of the clipped perceptron by the choice of a dynamical $\lambda_1 = \lambda_1(\alpha)$. In this paper we introduce the limits that result in the fastest decrease in the limit $\alpha \rightarrow \infty$. To conclude, choosing the limits that give the fastest decrease is given in Eq. (21) as explained in Secs. IV and VII.

As one can see from the definitions in Eq. (2), it is natural to choose the continuous weight vector as the one that is not constrained to a hypersphere, than to choose a vector constrained to a hypercube space. It was shown that in the case of storing random patterns, pretraining a continuous student whose weight vectors are constrained to the volume of a hypercube results in a better performance [7]. Open questions remain: what is the quantitative benefit that one can gain in a learning procedure by using the cubical constraint?; and can a learning strategy be designed that fulfills this constraint?

We studied the case of a very large L and show a scaling relation between L and N arising from the analysis. For $L \sim O(\sqrt{N})$ the learning curve is the one typical of the continuous case. However, it should remain clear that learning is the same as having a continuous student unless $\alpha \rightarrow \infty$ and $\rho_J \rightarrow 1$. In that regime the fast decay that characterizes the clipped learning appears.

All discrete computers actually correspond to a similar situation, where all available properties have a finite representation. The machine uses some kind of clipping by rounding the numbers. In fact, the process carried out by computers updates the clipped version by adding a continuous quantity to each weight component that depends on the mismatch between the *discrete* student and the teacher. The next step is rounding the student's weights. In such a scenario, a precursor is *not* used. In analysis of the latter, one has to make use of different method than the one presented here and it is beyond the scope of our analysis. An intermediate case where there is a precursor but its updating is done according to the clipped version has been analyzed in Refs. [26,27]. However, even if computers use larger memory space for the calculations during the learning (a kind of ''continuous'' precursor) and give final results by limited parameters (rounded ones) and hence the learning procedure is a kind of finite space, there should be difference between the expected results in the continuous machines and the actual results in the finite machines. The difference between the learning in the continuous student and the learning in the clipped one, as predicted here, can be significant only in the $\alpha \rightarrow \infty$ regime or small depth. Visualizing them is usually impossible since they are smaller than the measurement scale.

ACKNOWLEDGMENTS

Discussions with W. Kinzel and M. Biehl are acknowledged. We would like to thank M. Biehl for a careful reading of the manuscript. We are grateful to A.C.C. Coolen for bringing to our attention Ref. [27] and for fruitful discussions. The research was partially supported by the Israel Academy of Science.

[1] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).
[2] T. L. H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).
[3] H. Gutfreund and Y. Stein, J. Phys. A **23**, 2613 (1990).
[4] I. Kanter, Europhys. Lett. **17**, 181 (1992).
[5] R. Meir and J. F. Fontanary, J. Phys. A **25**, 1149 (1992).
[6] J. Schietse, M. Bouten, and C. Van den Broeck, Europhys. Lett. **32**, 279 (1995).
[7] M. Bouten, L. Reimers, and B. Van Rompaey, Phys. Rev. E **58**, 2378 (1998).
[8] M. Biehl and P. Riegler, Europhys. Lett. **28**, 525 (1994).
[9] W. Kinzel and R. Urbanczik, J. Phys. A **31**, L27 (1998).
[10] C. Van den Broeck and M. Bouten, Europhys. Lett. **22**, 223 (1993).
[11] M. Rosen-Zvi, J. Phys. A **33**, 7277 (2000).
[12] F. Vallet, Europhys. Lett. **8**, 747 (1989).
[13] F. Vallet and J. G. Gailton, Phys. Rev. A **41**, 3059 (1990).
[14] D. Saad and S. A. Solla, Phys. Rev. Lett. **74**, 4337 (1995); D. Saad and S. A. Solla, Phys. Rev. E **52**, 4225 (1995).
[15] M. Rosen-Zvi, M. Biehl, and I. Kanter, Phys. Rev. E **58**, 3606 (1998).
[16] M. Opper, Phys. Rev. Lett. **77**, 4671 (1996).
[17] O. Kinouchi and N. Caticha, J. Phys. A **25**, 6243 (1992).
[18] M. Biehl and H. Schwarze, J. Phys. A **28**, 643 (1995).
[19] M. Bouten, A. Komoda, and R. Serneels, J. Phys. A **23**, 2605 (1990).
[20] D. Malzahn, Phys. Rev. E **61**, 6261 (2000).
[21] l P. Sollich and D. Barber, Europhys. Lett. **38**, 477 (1997).
[22] W. Kinzel, Philos. Mag. B **77**, 1455 (1998).
[23] D. Barber, D. Saad, and P. Sollich, Europhys. Lett. **34**, 151 (1996).
[24] B. Derrida, R. B. Griffiths, and A. Prügel-Bennett, J. Phys. A **24**, 4907 (1991).
[25] A. Buhot, J. M. Torres Moreno, and M. B. Gordon, Phys. Rev. E **55**, 7434 (1997).
[26] E. Gardner and B. Derrida, J. Phys. A **22**, 1983 (1989).
[27] H. Chakravorty (unpublished).