Cavity approach to noisy learning in nonlinear perceptrons

Peixun Luo and K. Y. Michael Wong
Department of Physics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

(Received 19 July 2001; published 27 November 2001)

We analyze the learning of noisy teacher-generated examples by nonlinear and differentiable student perceptrons using the cavity method. The generic activation of an example is a function of the cavity activation of the example, which is its activation in the perceptron that learns without the example. Mean-field equations for the macroscopic parameters and the stability condition yield results consistent with the replica method. When a single value of the cavity activation maps to multiple values of the generic activation, there is a competition in learning strategy between preferentially learning an example and sacrificing it in favor of the background adjustment. We find parameter regimes in which examples are learned preferentially or sacrificially, leading to a gap in the activation distribution. Full phase diagrams of this complex system are presented, and the theory predicts the existence of a phase transition from poor to good generalization states in the system. Simulation results confirm the theoretical predictions.

DOI: 10.1103/PhysRevE.64.061912 PACS number(s): 87.18.Sn, 75.10.Nr, 02.50.−r, 07.05.Mh

I. INTRODUCTION

Since Hopfield's pioneer work on neural networks [1], statistical mechanics has been proved to be a powerful tool in the study of information processing. Mean-field theories such as the replica method [2–4] and the cavity method [2,5–7] are successfully developed to study these problems. In particular, it provides valuable insights to the learning of examples in neural networks by considering it as an energy minimization process. Early work used the replica method to study the learning problem in various situations [8–11]. It has the advantage of a readily-used mathematical formalism applicable to general cases, and has been applied to linear networks [12–15] and networks with binary outputs [11,16–21], dealing with learning tasks that are either realizable or unrealizable, random or teacher-generated data, and clean or noise corrupted data. These studies mainly focused on the global properties of the learning system, with less emphasis on the microscopic description of the examples and the weights in the system. Furthermore, most of these models were still remote from the differentiable nonlinear perceptron that is most commonly used today. Other work used the Green's function approach that is particularly convenient for linear networks [22], but these systems may not have the competitive effects among examples in nonlinear networks, which will be investigated in this paper. The annealed approximation is suitable for analyzing high-temperature learning [3], but the results cannot be directly extended to the more common case of low temperature.

A common phenomenon observed in the studies of learning from examples is the existence of phase transitions with abrupt improvement in the generalization ability of the networks once the training examples are sufficiently numerous, or the global parameters (e.g., the weight decay) are suitably tuned [23–27]. These transitions are often discontinuous. They arise when metastable states are present in the system, leading to discontinuous jumps in the network states, hystereses, and the disappearance of metastability at spinodal points. Multilayer perceptrons will exhibit a transition from permutation symmetric to specialized states [4]. In the present paper, we will see these effects in nonlinear perceptrons learning noisy examples. Here the competition between the locally stable states comes from the different learning strategies used to attain the systemwide energy minimum.

The cavity method is a suitable tool to study information competition effects in rule extraction from noisy examples. Large scale neural networks with many nodes can be considered as mean-field systems since, as far as the learning of one example is concerned, the influence of other examples can be regarded as a background satisfying some average properties. The success of the mean-field approach is illustrated by the capability of the replica method in describing the macroscopic properties of neural network learning [4]. However, the replica method provides much less interpretation on the processing of individual examples since its starting point is the quenched average of the free energy over the example distribution. The cavity method is an alternative version of mean-field theory. It is a generalization of the Thouless-Anderson-Palmer approach to spin glasses and starts from a microscopic description of the system elements [28,2]. In this method, mean-field equations are derived from self-consistent considerations. The method was subsequently generalized to learning problems [5,6,29] and yields macroscopic properties identical to the replica method while at the same time provides physical insights to the learning of individual examples. Recently, the cavity method was also applied to a number of problems in information processing [30].

In this paper, we study the learning of noisy examples in nonlinear perceptrons using the cavity method. Nonlinear networks have the following advantages: (i) compared with networks with binary output, gradient descent learning is possible, (ii) nonlinearity is representative of more complex networks, (iii) they have more resemblance with biological neurons [31]. Compared with previous studies, we will focus on the effects of information competition in the system, and their consequences on the energy landscape, the appearance of band gaps in the activation distribution, the choice between preferential and even-handed learning strategies as well as their possible relationship with phase transitions in this complex system. We analyze the parameter regimes with band gaps in the activation distribution, as well as the stability condition of the perturbative cavity approach. Simulation results show that the assumption of a smooth energy landscape usually works well when no gaps are present, but tends to fail when gaps appear. The phase diagram of this complex system is shown and the occurrence of phase transitions is investigated and compared with simulations.

The rest of this paper is organized as follows. After describing the model in the next section, we describe in Sec. III the cavity approach and introduce the cavity activation, which is the core microscopic variable in the cavity method. Three self-consistent equations are derived when a smooth energy landscape is assumed. In Sec. IV, we discuss the case when band gaps appear in the activation distribution. Phase transitions in nonlinear perceptrons and phase diagrams are the themes of Sec. V. In Sec. VI we summarize the results and their implications. Mathmetical details are appended at the end of the paper.

II. THE MODEL

Consider a student perceptron with N weights $J_j, j=1,\ldots,N$, connecting the N input nodes and the output node. It is trained to extract the rule of a teacher perceptron with the same architecture with N weights $B_j, j=1,\ldots,N$, where $\langle B_j\rangle=0$ and $\langle B_j^2\rangle=1$. A training set of p examples generated by the teacher and corrupted by noise is what the student can explore. Each example, labeled μ with $\mu=1,\ldots,p$, consists of the input vector $\boldsymbol{\xi}^\mu$ and the noisy output O_μ of the teacher. The input components ξ_j^μ are Gaussian random variables, with $\langle\xi_j^\mu\rangle=0$ and $\langle\xi_j^\mu\xi_k^\nu\rangle=\delta_{jk}\delta_{\mu\nu}$. The activation functions $f(x)$ of both perceptrons are differentiable and nonlinear, such as $\mathrm{sig}(x)\equiv(1-\tanh x)/2$, i.e., the teacher and student outputs are, respectively,

$$O_\mu\equiv f(\tilde{y}_\mu)\equiv f(y_\mu+T\eta_\mu) \quad \text{and} \quad f_\mu\equiv f(x_\mu), \tag{1}$$

where $y_\mu\equiv\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}$ is the teacher activation, η_μ is Gaussian noise with $\langle\eta_\mu\rangle=0$ and $\langle\eta_\mu^2\rangle=1,T$ is the noise temperature, and $x_\mu\equiv\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}$ is the student activation.

During the training procedure, one adapts the student network to minimize an energy function that measures the difference between student outputs f_μ and teacher outputs O_μ for all training examples. A natural energy function is the total quadratic error of examples in training set, $\Sigma_{\mu=1}^p(O_\mu-f_\mu)^2\equiv p\epsilon_t^2$, where we call ϵ_t training error. However, the final target of learning is to get a student perceptron that can generalize well to novel examples, i.e., to minimize the generalization error $\epsilon_g\equiv\langle[O(\boldsymbol{\xi})-f(\boldsymbol{\xi})]^2\rangle^{1/2}$, where $\langle\,\rangle$ is the average performed over the distribution of all inputs and the noise. We add a weight decay term to penalize excessively long weight vectors and speed up learning, and use the energy function

$$E=\frac{1}{2}\sum_\mu (O_\mu-f_\mu)^2+\frac{\lambda}{2}\sum_j J_j^2, \tag{2}$$

where λ is the weight decay strength. Minimizing the above energy function by gradient descent, one obtains the equilibrium state of the student perceptron given by

$$J_j=\frac{1}{\lambda\sqrt{N}}\sum_\mu (O_\mu-f_\mu)f'_\mu\xi_j^\mu, \tag{3}$$

where the prime in f'_μ represents the derivative of $f(x_\mu)$. Here we are interested in the dependence of the generalization error ϵ_g of the student perceptron in its equilibrium state (3) on the macroscopic parameters, such as the weight decay strength λ, the noise temperature T and the size of training set $\alpha\equiv p/N$. As in perceptrons with linear or discrete activation functions, the generalization error is essentially determined by the overlap of the student weight vector with the teacher weight vector R and the magnitude of the student weight vector q, which are defined as

$$q=\langle J_j^2\rangle_j \quad \text{and} \quad R=\langle J_jB_j\rangle_j, \tag{4}$$

where $\langle\,\rangle_j$ represents averaging over the N weights.

III. THE CAVITY METHOD

The cavity method developed in Refs. [6,29] is used to tackle the current problem in order to get more microscopic understanding of the mechanism in the learning of neural networks. After the student perceptron is trained with p examples, it reaches its energy ground state $\boldsymbol{J}$ given by Eq. (3). Suppose a new example with input vector $\boldsymbol{\xi}^0$ is fed to the student perceptron. The activation of example 0 is now given by

$$t_0\equiv\frac{1}{\sqrt{N}}\boldsymbol{J}\cdot\boldsymbol{\xi}^0, \tag{5}$$

which is called the cavity activation. Since the student $\boldsymbol{J}$ has no information about an example it has never learned, the cavity activation t_0 is a Gaussian variable for random inputs ξ_j^0 when $N\gg1$. It has a mean $\langle\langle t_0\rangle\rangle=0$ and covariances $\langle\langle t_0^2\rangle\rangle=q$ and $\langle\langle t_0y_0\rangle\rangle=R$, where $\langle\langle\,\rangle\rangle$ denotes the ensemble average. Hence the distribution of the cavity field is

$$P(t_0|y_0)=\frac{\exp\left[-\dfrac{(t_0-Ry_0)^2}{2(q-R^2)}\right]}{\sqrt{2\pi(q-R^2)}}. \tag{6}$$

Trained with all the $p+1$ examples $\{(\boldsymbol{\xi}_\mu,O_\mu)|\mu=0,1,\ldots,p\}$, the student perceptron reaches its equilibrium state $\boldsymbol{J}^0$, with

$$J_j^0=\frac{1}{\lambda\sqrt{N}}(O_0-f_0^0)(f_0^0)'\xi_j^0+\frac{1}{\lambda\sqrt{N}}\sum_\mu (O_\mu-f_\mu^0)(f_\mu^0)'\xi_j^\mu. \tag{7}$$

Here and below, variables with superscript 0 refer to those associated with the perceptron $\boldsymbol{J}^0$, which includes example 0 in its training set. We see that the generic student activation of example 0, $x_0 \equiv \boldsymbol{J}^0 \cdot \boldsymbol{\xi}^0/\sqrt{N}$, is not a Gaussian variable. (Although the correct notation of x_0 should be x_0^0, here we omit the superscript since it is sufficiently distinct from its cavity counterpart t_0.) However, it is reasonable to assume that the difference between $\boldsymbol{J}$ and $\boldsymbol{J}^0$ is small; the validity of this assumption will be discussed later. Following the perturbative analysis in Ref. [6], we show in Appendix A that, for a given corrupted teacher output $\tilde{y}_0$, there is a well defined relation between t_0 and $x_0, t_0 = t(x_0, \tilde{y}_0)$, where

$$t(x,\tilde{y}) = x - \gamma[f(\tilde{y}) - f(x)]f'(x). \tag{8}$$

Here the parameter γ is the local susceptibility and satisfies

$$1 - \gamma\lambda = \alpha\left\langle 1 - \frac{\partial x_\mu}{\partial t_\mu}\right\rangle_\mu, \tag{9}$$

where x_μ is a single-valued function of t_μ, and $\langle\ \rangle_\mu$ represents averaging over the p examples. In this section we will focus on the case that Eq. (8) presents a one-to-one mapping between x_μ and t_μ for a given $\tilde{y}_\mu$. As we shall see, this corresponds to a continuous activation distribution with no band gaps. In the next section we will discuss the case when t_μ has a one-to-many relation with x_μ, which will lead to the emergence of band gaps.

Combining Eqs. (6) and (8), we can derive the student activation distribution $P(x|\tilde{y},y)$,

$$P(x|\tilde{y},y) = P(t(x,\tilde{y})|y)\frac{\partial t(x,\tilde{y})}{\partial x}. \tag{10}$$

In turn, the distributions $P(\tilde{y}|y)$ and $P(y)$ are given by

$$P(\tilde{y}|y) = \frac{1}{\sqrt{2\pi T^2}}\exp\left[-\frac{(\tilde{y}-y)^2}{2T^2}\right], \tag{11}$$

$$P(y) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right). \tag{12}$$

Equation (9) for γ can now be transformed to an integral expression when N is very large,

$$1 - \gamma\lambda = \alpha\int dy P(y)\int d\tilde{y}P(\tilde{y}|y)\int dx P(x|\tilde{y},y)\left(1 - \frac{\partial x}{\partial t}\right), \tag{13}$$

where $\partial x/\partial t = (1 + \gamma\{[f'(x)]^2 - [f(\tilde{y}) - f(x)]f''(x)\})^{-1}$ from Eq. (8). Equation (13) can be simplified into an equation involving only double integrals,

$$1 - \gamma\lambda = \alpha\int Du\int Dv\left(1 - \frac{\partial x}{\partial t}\right), \tag{14}$$

where $Du \equiv du\exp(-u^2/2)/\sqrt{2\pi}$ and $Dv \equiv dv \times\exp(-v^2/2)/\sqrt{2\pi}$ are Gaussian measures, $\tilde{y} = \sqrt{1+T^2}u$, and x and t depend on u and v via

$$\frac{R}{\sqrt{1+T^2}}u + \sqrt{q - \frac{R^2}{1+T^2}}v = t(x,\tilde{y}). \tag{15}$$

The mean-field equation for R can be obtained by multiplying both sides of Eq. (3) with B_j and summing over j, yielding

$$R = \frac{\alpha}{\lambda}\int dy P(y)\int d\tilde{y}P(\tilde{y}|y)\int dx \times P(x|\tilde{y},y)[f(\tilde{y}) - f(x)]f'(x)y. \tag{16}$$

Using Eqs. (10)–(12), (14) and (15) and after elaborate integrating by part, we arrive at

$$R = \alpha\gamma\int Du\int Dv f'(\tilde{y})f'(x)\frac{\partial x}{\partial t}. \tag{17}$$

Similarly, multiplying both sides of Eq. (3) with J_j, and summing over j, we have another mean-field equation for q,

$$q - R^2 = \alpha\gamma^2\int Du\int Dv[f(\tilde{y}) - f(x)]^2[f'(x)]^2. \tag{18}$$

The three macroscopic parameters γ, R, and q can now be obtained by solving the three mean-field equations (14), (17), and (18) numerically for given values of α, λ, and T. Therefore, we can directly obtain the training error ε_t and generalization error ε_g, which depend on the generic activation x and cavity activation t, respectively,

$$\varepsilon_t^2 = \int Du\int Dv[f(\tilde{y}) - f(x)]^2, \tag{19}$$

$$\varepsilon_g^2 = \int Du\int Dv[f(\tilde{y}) - f(t)]^2. \tag{20}$$

The validity of the perturbative calculation can be checked by considering the stability condition of the equilibrium state. As derived in Appendix B, when the new example 0 is added, the magnitude of the change in the student weight vector is given by

$$\Delta_J \equiv \sum_j (J_j^0 - J_j)^2 = \frac{(x_0 - t_0)^2}{1 - \alpha\left\langle\left(1 - \frac{\partial x_\mu}{\partial t_\mu}\right)^2\right\rangle_\mu}. \tag{21}$$

Hence Δ diverges when the denominator approaches 0. This yields the stability condition

$$\alpha\left\langle\left(1 - \frac{\partial x_\mu}{\partial t_\mu}\right)^2\right\rangle_\mu < 1. \tag{22}$$

It is identical to the stability condition of the replica-symmetric (RS) ansatz in the replica approach [11,6], the so-called Almeida-Thouless (AT) condition [32].

In the region where the stability condition (22) is violated, the perturbative version of cavity method breaks down. It becomes possible that when a new example is added to the system, the ground state relocates to another metastable state. This corresponds to the picture of a rough energy landscape with many metastable states and the perturbative cavity method has to be modified [29]. In the formalism of the replica method, it was shown by Parisi that breaking of the replica symmetry is the thermodynamical transcription of the existence of many pure thermodynamic states [33]. Hence the RS and replica symmetry-breaking (RSB) approximations in the replica method describe the situations of smooth and rough energy landscapes, respectively.

IV. ACTIVATION DISTRIBUTIONS WITH BAND GAPS

When the activation function $f(x)$ is nonlinear, the behavior of the system may be very complex. This can be seen by considering Eq. (8) for a sufficiently large γ, when the generic activation x may become a multivalued function of the cavity activation t.

To compare the energy of the possible states, we consider the energy difference between the perceptron states $\boldsymbol{J}^0$ and $\boldsymbol{J}$. According to Eq. (3),

$$\Delta E \equiv E^0 - E = \frac{1}{2}(O_0 - f_0^0)^2 + \frac{1}{2}\sum_\mu [(O_\mu - f_\mu^0)^2 - (O_\mu - f_\mu)^2] + \frac{\lambda}{2}\sum_j [(J_j^0)^2 - J_j^2]. \tag{23}$$

Expanding the first summation to the second order $(x_\mu^0 - x_\mu)^2$ and substituting Eq. (3) and Eq. (7) to the second summation, we can simplify the above equation to

$$\Delta E = \frac{1}{2}(O_0 - f_0^0)^2 + \frac{1}{2\gamma}(x_0 - t_0)^2, \tag{24}$$

using the relation between the cavity activation t_0 and generic activation x_0 in Eq. (8). The first term is the primary change due to the newly added example, and the second term results from the adjustment of the background examples. In the multivalued region, when the energy minimum favors x_0 to take a value closer to t_0 (therefore, favorable to small background adjustment), the example is *sacrificed*. Otherwise, when the output f_0^0 is closer to the teacher's output O_0 (therefore favorable to small primary cost), the example is *preferentially* learned. The competition between the two possible responses to a new example leads to a discontinuity in the range of the generic activation x_0 when the cavity activation t_0 varies, accompanied by the appearance of gaps in the activation distribution for a given teacher output O_0.

To study this competition, we suppose that Eq. (8) has multiple solutions of x in a range of t for a given $\tilde{y}$. We are interested in the point $t_g(\tilde{y})$ where two solutions yield the same energy change ΔE. That is, there are two distinct values of x,$x_<$ and $x_>$, such that $t_g = t(x_<,\tilde{y}) = t(x_>,\tilde{y})$ and $\Delta E(x_<) = \Delta E(x_>)$. Then using Eq. (24), we arrive at the condition

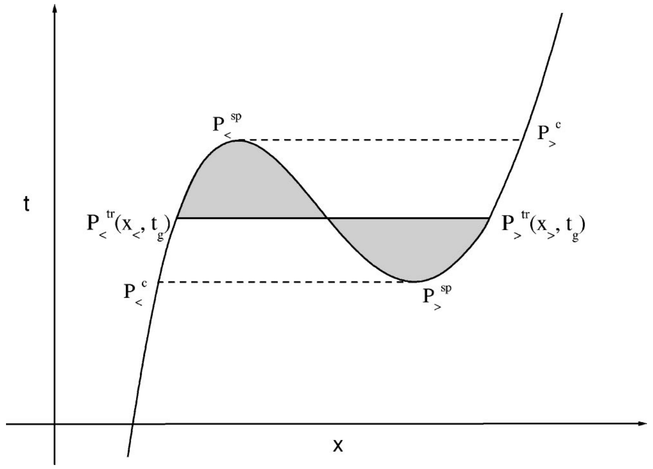

FIG. 1. The Maxwell's construction to determine the position of band gap t_g. In the figure, the areas of the two shaded regions equal to each other. Between $P_<^c$ and $P_<^{tr}$, the left state competes with a metastable right state between $P_>^{sp}$ and $P_>^{tr}$, but the left state remains the ground state. Similar competitions exist between $P_>^{tr}$ and $P_>^c$.

$$\int_{x_<(t_g)}^{x_>(t_g)} t(x)dx = t_g[x_>(t_g) - x_<(t_g)], \tag{25}$$

which is the Maxwell's construction as shown in Fig. 1. As the result of energy minimization, one of the two solutions of x is preferred on the left neighborhood of t_g, while the other is preferred on the right. Hence x is a function of t with a discontinuity at t_g. Consequently, a band gap appears in the student activation distribution for a given teacher output,

$$P(x|\tilde{y}) = 0 \qquad \text{when} \qquad x \in [x_<(t_g), x_>(t_g)]. \tag{26}$$

Extra terms should then be added to the mean-field equations Eqs. (14) and (17) for γ and R, as derived in Appendix C, namely,

$$1 - \gamma\lambda = \alpha\int Du \sum_i \int_{R_i} Dv \left(1 - \frac{\partial x}{\partial t}\right) - \alpha\int Du \sum_j G(t_g^j) \times [x_>(t_g^j) - x_<(t_g^j)], \tag{27}$$

$$R = \alpha\gamma\int Du \sum_i \int_{R_i} Dv f'(\tilde{y}) f'(x) \frac{\partial x}{\partial t} + \alpha\gamma\int Du \sum_j G(t_g^j) f'(\tilde{y})[f(x_>(t_g^j)) - f(x_<(t_g^j))], \tag{28}$$

where each term in the summations over i corresponds to an integration over a region R_i separated from each other by band gaps, and each term in the summations over j corresponds to a band gap. The Gaussian factor $G(t_g^j)$ is given by

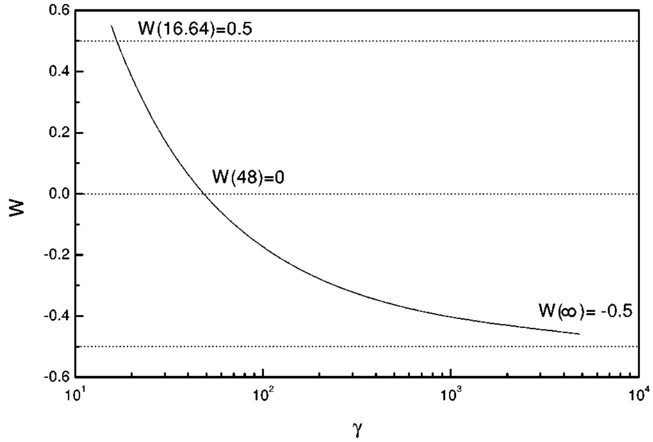

FIG. 2. The function W of γ defined by Eq. (D2).

$$G(t_g^j)=\frac{1}{\sqrt{2\pi\left(q-\frac{R^2}{1+T^2}\right)}}\exp\left[-\frac{\left(t_g^j-\frac{R}{\sqrt{1+T^2}}u\right)^2}{2\left(q-\frac{R^2}{1+T^2}\right)}\right]. \tag{29}$$

We note that the extra terms due to gaps are consistent with adding the delta function component $(x_>-x_<)\delta(t-t_g)$ to $\partial x/\partial t$ in Eq. (14) and $[f(x_>)-f(x_<)]\delta(t-t_g)$ to $f'(x)\partial x/\partial t$ in Eq. (17).

For the sigmoid function $f(x)=(1+e^{-x})^{-1}$, the necessary and sufficient condition for Maxwell's construction, as derived on Appendix D, is

$$f(\tilde{y})-\frac{1}{2}>W(\gamma)\quad \text{for}\quad x<0,$$

$$\frac{1}{2}-f(\tilde{y})>W(\gamma)\quad \text{for}\quad x>0, \tag{30}$$

where the function $W(\gamma)$ is monotonic, as shown in Fig. 2. The behavior of the activation distribution depends on the value of γ in the following three cases.

Case 1: $\gamma<(117+165\sqrt{33})/64\approx 16.64$. As $W(\gamma)>1/2$ and $0\leqslant f(\tilde{y})\leqslant 1$, the condition (30) cannot be satisfied for all teacher output $f(\tilde{y})$. Hence there is no gap in the activation distribution.

Case 2: $16.64<\gamma<48$. Here $0<W(\gamma)<1/2$. The activation distribution starts to develop a band gap that extends from $f(\tilde{y})=1$ to $f(\tilde{y})=1/2+W(\gamma)$ in the region $x<0$. Similarly, another band gap extends from $f(\tilde{y})=0$ to $f(\tilde{y})=1/2-W(\gamma)$ in the region $x>0$. The two band gaps are symmetric with respect to the point $(f(x),f(\tilde{y}))=(1/2,1/2)$. For intermediate teacher output between $1/2\pm W(\gamma)$, the distribution remains continuous.

Case 3: $\gamma>48$. Here $W(\gamma)<0$. The band gap in the region $x<0$ now extends from $f(\tilde{y})=1$ to $f(\tilde{y})=1/2+W(\gamma)<1/2$. Together with its symmetric counterpart in the region $x>0$, the activation distribution is three banded for $1/2+W(\gamma)<f(\tilde{y})<1/2-W(\gamma)$, beyond which the activation distribution remains two banded.

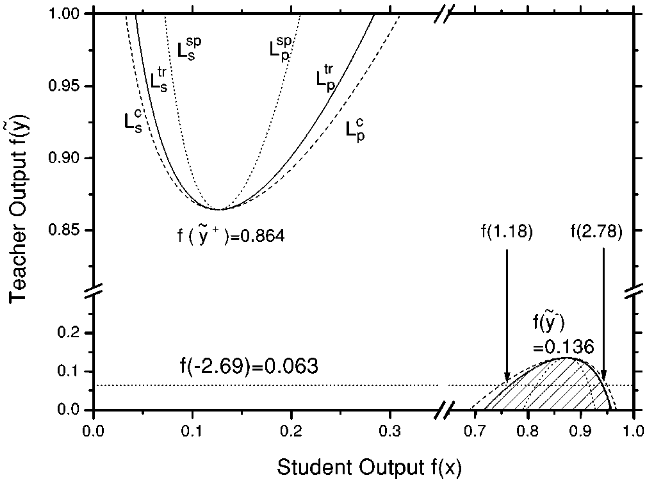

FIG. 3. The occurrence of band gap and preferential learning when $\gamma=20.55$. The picture is symmetric with respect to the point (1/2,1/2). For intermediate teacher output $0.136<f(\tilde{y})<0.864$, no band gaps exist. For related discussions in Figs. 4, 5, and Table I when $\alpha=3$ and $\lambda=0.002$, the present value of γ corresponds to $T=2.5$, and the choice of $\tilde{y}=-\sqrt{1+T^2}$ corresponds to the line of $f(\tilde{y})=0.063$, which cuts the boundaries of the shaded region at $x=1.18$ and $x=2.78$, indicating a band gap of $P(x|\tilde{y})$ at [1.18,2.78].

Case 2 is illustrated in Fig. 3, where a band gap exists in the regions that are shaded or enclosed by the transition lines L_s^{tr} and L_p^{tr} (subscripts s and p represent sacrificed and preferred states, respectively). Near the sacrificed band edge, the line L_s^c indicates the onset of competition. Between the lines L_s^c and L_s^{tr} (or the region close right to the shaded one), the sacrificed state is competing with a metastable preferred state, which appears between the spinodal line L_p^{sp} and the line L_p^{tr}, but the sacrificed state remain the ground state. Between line L_s^{tr} and the spinodal line L_s^{sp}, the sacrificed state becomes metastable and disappears at L_s^{sp}. Similar lines exist in the neighborhood of the preferred band edge (or the region close left to the shaded one).

Figure 3 shows that preferential learning occurs at extreme values of $f(\tilde{y})>f(\tilde{y}^+)$ or $f(\tilde{y})<f(\tilde{y}^-)$. The energy advantage of this learning strategy can be easily understood. In nonlinear perceptrons, changes in the student activation around these extreme values of $f(\tilde{y})$ do not result in significant changes in the training error of an example due to the saturation in this region, and if the cavity activation is very different from the teacher's activation, it is more economical to keep the student activation close to the cavity activation, so that the background adjustment remains small. In contrast, for intermediate values of $f(\tilde{y})$, the competitive effects are less, and no band gaps develop.

The width of the band gap can be narrowed when the existence of metastable states is taken into account. As shown in Fig. 3, metastable states exist inside the band gap as far as the spinodal lines L_s^{sp} and L_p^{sp}. Hence in finite-time

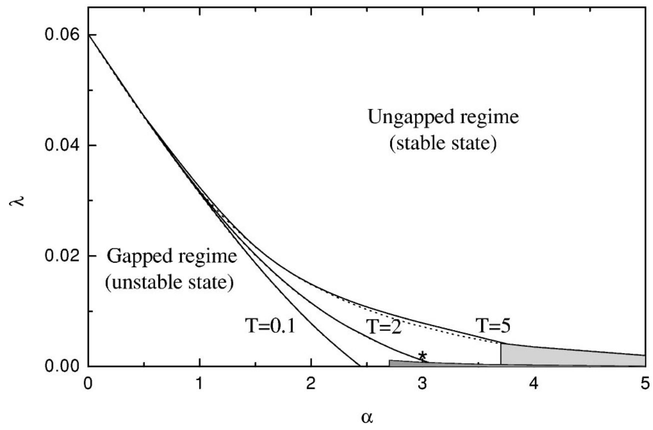

FIG. 4. Regimes of the existence of gapped activation distribution and regimes of unstable states for different noise temperatures. Below the solid lines, the perturbative cavity solutions are unstable, and below the dotted line, band gaps will appear in the activation distribution. The point $(\alpha=3,\lambda=0.002)$ is denoted by a star. The shaded regions indicate the existence of discontinuous phase transitions to be discussed in Sec. V. (For $T=0.1$, the shaded region is too small to be shown.)

simulations, the system may be trapped in metastable states. Conventionally, the narrowing of band gaps in simulations is interpreted by RSB effects in the replica method [19,20]. Here the narrowing can be explained by metastability in the perturbative cavity method, even without invoking the formalism of RSB.

This kind of preferential learning is clearly not present in linear perceptrons, even when perfect learning is impossible, since the activation x is a linear function of the cavity activation t, by virtue of Eq. (8). Hence preferential learning is a unique consequence of the nonlinearity of the perceptron activation.

Figure 4 shows the parametric regimes for the existence of gapped activation distributions as well as the unstable regimes of the perturbative cavity method (the boundary line being equivalent to the AT line in the replica method) for different noise temperatures. Since the development of a gap is already sufficient to cause an uncontrollable change in Eq. (21), the gapped regions lie inside the unstable regions. Furthermore, provided that α and T are not too large, the boundaries of the gapped and unstable regions are very close to each other. The region of small weight decay and large noise will be discussed in the next section, where the phase lines are modified when discontinuous transitions take place.

Trends for the existence of band gaps in the activation distribution can be observed from Fig. 4. Gaps exist only when the size α of training set is small, leaving ambiguities about the underlying rule. Furthermore, increasing the data noise broadens the gapped region, since it introduces conflicting information to be learned by the student. Finally, since weight decay restricts the flexibility in the weight space, and hence reduces the tendency for multiple minima, gaps are found for small λ.

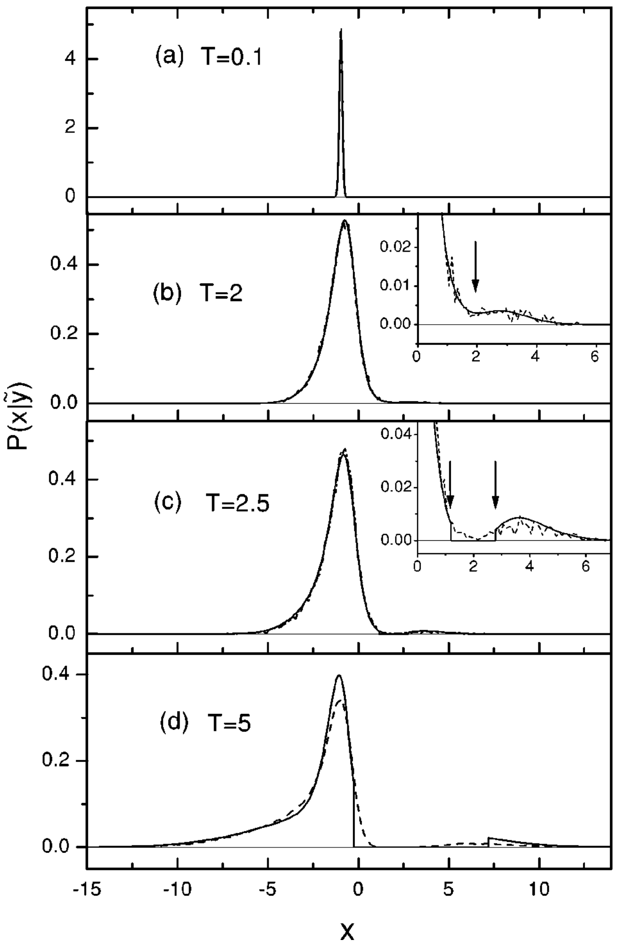

FIG. 5. Theoretical and simulation results of student activation distributions, indicated by solid and dashed lines, respectively, when $\alpha=3$ and $\lambda=0.002$ [denoted by a star in Fig. (4)]. We choose $\tilde{y}=-\sqrt{1+T^2}$ for different noise temperatures, so that $P(\tilde{y})$ are the same. The arrow in (b) shows the position of a pseudo gap and the arrows in (c) show the band gap [1.18, 2.78] from the theoretical prediction in Fig. 3.

We check the appearance of band gaps predicted in our theory with simulations in Fig. 5. In Fig. 5(a), $\gamma=11.1$ and the stability condition (22) is fulfilled. The student activation distribution in this case has a single band and is a sharp peak at $x=\tilde{y}$. When noise temperature T increases to 2 where $\gamma=14.9$, the location of the point $(\alpha=3,\lambda=0.002)$ in Fig. 4 is slightly above the boundary between gapped and continuous regimes. Correspondingly, there is a pseudogap developing in the activation distribution, as shown in Fig. 5(b). Comparing with simulation results, we see that the assumption of a smooth energy landscape used in the present work is valid in this regime. As shown in Table I , the theoretical and simulation results of macroscopic properties also agree well.

In Fig. 5(c), $T=2.5$ and the stability condition is violated. There is now a gap in both the theory and simulation. However, at a higher T in Fig. 5(d), the theoretical prediction of the band gap is broader and has sharper edges than the simulation one. At the same time, there are prominent differences of the corresponding ε_t,R, and, especially, q in Table I. Two arguments are relevant. First, the narrowing of the band gap can be explained by the presence of metastable states in the band gap as discussed in Fig. 3. These metastable states probably prevent the learning process to converge to the

TABLE I. The comparison of macroscopic parameters and errors obtained from theory (roman) and simulation (italics in brackets) for different T when $\alpha=3$ and $\lambda=0.002$.

T	γ	R	q	ϵ_t	ϵ_g
0.1	11.1	0.963 (*0.961*)	0.933 (*0.932*)	0.018 (*0.017*)	0.027 (*0.027*)
2.0	14.9	0.796 (*0.795*)	2.211 (*2.196*)	0.236 (*0.236*)	0.376 (*0.375*)
2.5	20.6	0.837 (*0.822*)	3.816 (*3.574*)	0.260 (*0.260*)	0.437 (*0.431*)
5.0	80.1	0.920 (*0.848*)	23.05 (*16.34*)	0.275 (*0.301*)	0.577 (*0.570*)

ground state, which, therefore, yields a value of q different from the theory. Second, due to the violation of the stability condition (22) when the band gap develops, a rough energy landscape as discussed previously [29] must be introduced to improve the agreement.

In Fig. 6, the variation of the activation distributions for different α at a given T and λ shows another trend of band gap evolution. One finds that while Fig. 4 shows that insufficient examples cause the appearance of band gaps, here it is possible that the fraction of examples located in the sacrificed band decreases with the size of the example set. Therefore, the competitive effects of learning strategies are serious only when both the noise and the size of training set are large, as one may expect intuitively.

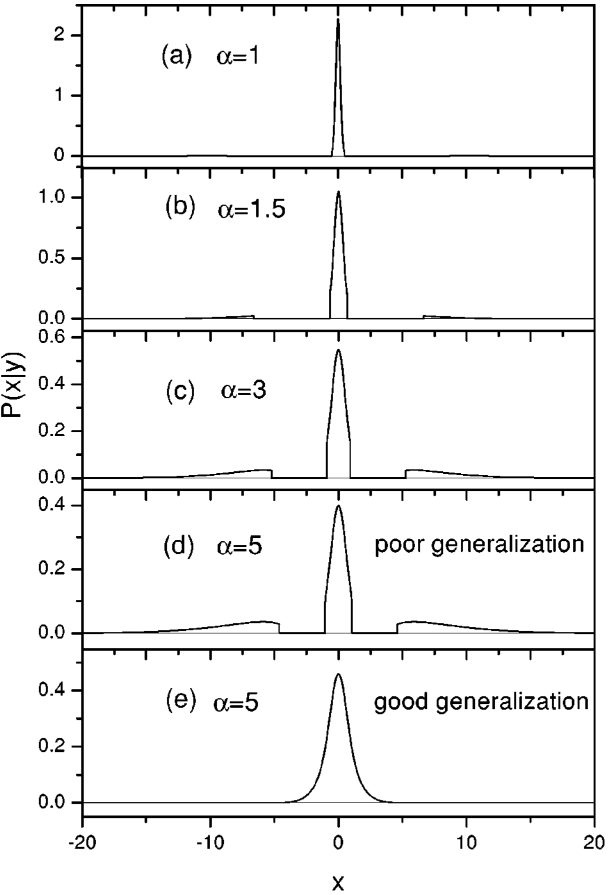

FIG. 6. The theoretical prediction of the student activation distributions at $T=5$ and $\lambda=0.001$ for different sizes of the training set α, where $\tilde{y}=0$. When $\alpha=5$, the system has two states. Respectively, (d) and (e) are the distributions when the system is in the poor and good generalization states.

V. PHASE TRANSITIONS

Another consequence of nonlinearity is the existence of two stable solutions of R, q, and γ to the mean-field equations (27), (28), and (18) for a given set of parameters. Studying the behaviors of the curves of λ versus γ such as those in Fig. 7, we find two critical parameters $\alpha_c^*(T)$ and $\alpha_0(T)$ for a given noise temperature T. The three accompanying cases of phase behavior are illustrated in Fig. 8.

Case 1: $\alpha<\alpha_c^*(T)$. λ is a monotonic decreasing function of γ. Hence for any weight decay strength, there is a unique local susceptibility. Numerical results in this region show that the magnitude of student weight vector q increases with decreasing weight decay λ. As shown in Fig. 8, there is no phase transition.

Case 2: $\alpha_c^*(T)\leqslant\alpha<\alpha_0(T)$. At $\alpha=\alpha_c^*(T)$, multiple solutions of γ for a given λ start to appear near the inflection point of the curve. The solution with the smallest γ corresponds to the *good generalization* solution with small q and ε_g. The solution with the largest γ corresponds to the *poor generalization* solution with large q and ε_g. In between the two solutions, there is a third, unstable, solution, which can be considered as the barrier separating the two stable solutions in the energy landscape. When α increases beyond $\alpha_c^*(T)$, the intermediate range of λ where multiple solutions exist becomes increasingly wide.

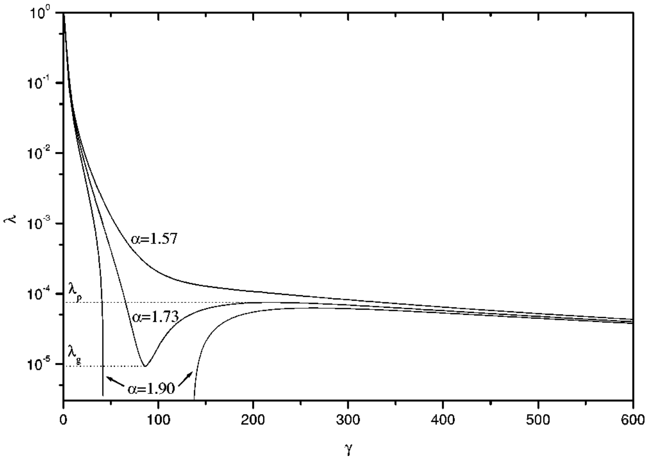

FIG. 7. The dependence of local susceptibility γ on the weight decay strength λ for different α at $T=1$, with $\alpha_c^*(1)=1.65$ and $\alpha_0(1)=1.74$. All curves approach $\lambda=0$ when γ goes to infinity.

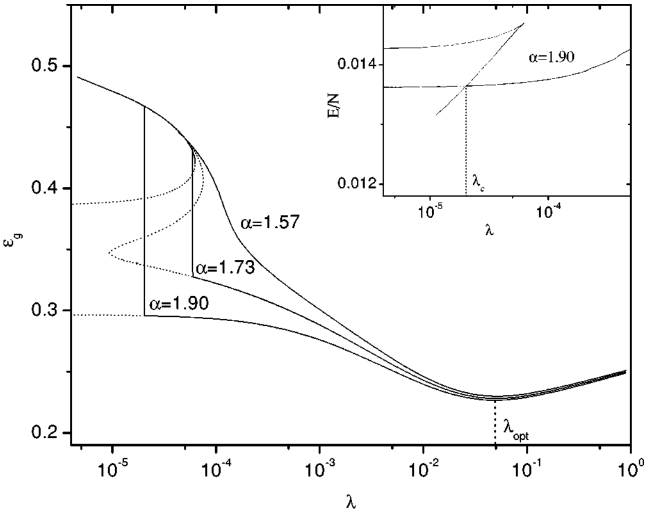

FIG. 8. The dependence of generalization error ε_g on the weight decay strength λ for values of α belonging to three different kinds of phase behavior when the noise temperature $T=1$. Inset illustrates the three branches of energy curve for $\alpha=1.90$.

At very large λ, the good generalization state is the only stable solution. When λ decreases, a metastable state with poor generalization appears at the *spinodal point* $\lambda_p(\alpha,T)$. When λ decreases further, the globally stable state switches from the good generalization state to the poor at $\lambda_c(\alpha,T)$. As shown in the inset of Fig. 8, the energy curve has two stable branches that cross at $\lambda_c(\alpha,T)$, where a first-order transition occurs, with possible hysteretic effects. On further decreasing of λ, the metastable state of good generalization disappears at another *spinodal point* $\lambda_g(\alpha,T)$. Hence $\alpha_c^*(T)$ is a *critical point* where discontinuous transition first appears.

Case 3: $\alpha>\alpha_0(T)$. At $\alpha=\alpha_0(T)$, the *spinodal point* λ_g of the good generalization state vanishes. Hence both poor generalization and good generalization solutions coexist for λ below λ_p down to zero, as shown in Fig. 8. Here the example set is large enough to provide information about the teacher such that the good generalization solution exists even in the absence of weight decay, although it may be metastable.

The existence of the discontinuous transition when λ changes, accompanied by the hysteretic effects, is verified by the simulation of a sample in Fig. 9. It is interesting to observe a third state with intermediate q and ε_g. The existence of such intermediate states is not uncommon in simulations, although transitions between the poor and good generalization are mostly direct, as predicted by the theory. Considering the stability condition (22) for the parameters used in Fig. 9, we find that the perturbative cavity solution is stable in the good generalization phase, but unstable in the poor one. This strongly implies that multiple metastable states exist in the poor generalization phase, contributing to the cascading transition observed in Fig. 9.

Similarly, discontinuous transitions occur when α increases for a given λ. In the learning curves in Fig. 10, we see that the student may even learn worse for more training examples if they are not sufficient. Only after sufficient examples are fed to the student will ε_g decrease asymptotically on increasing α, showing a bump at intermediate α. For smaller weight decay, there is a discontinuous transition from a good to a poor generalization state at a critical example size $\alpha_c(\lambda,T)$. Discontinuous transitions on changing α and λ are also observed in the high temperature limit in multilayer networks learning clean examples [25]. For increasing λ, the bump smoothes out and the position of α with maximum ε_g shift towards 1. The position of the maximum also depends on the noise temperature *T*. For small values of *T*, the maximum stays near $\alpha=1$. For larger noise, the maximum could move to higher values of α, which implies that more examples are required for the student to really learn some essence of the teacher's rule when the noise is stronger. Similar overtraining behavior is also found in linear networks learning unrealizable tasks [13,14], but no phase transition is found there.

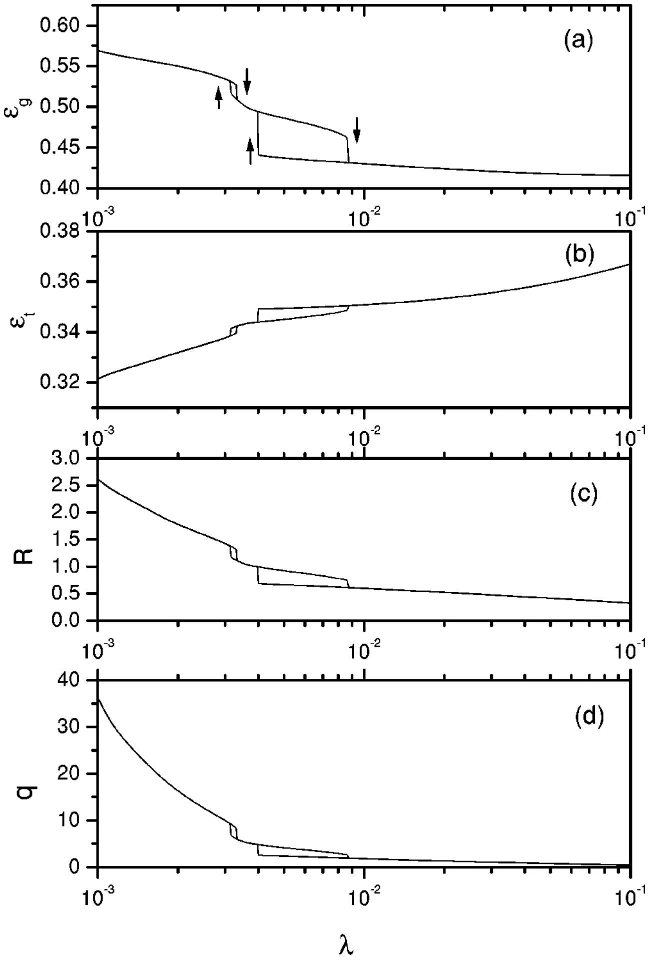

FIG. 9. Variations of ε_g, ε_t, R, and q of a sample in simulation when the weight decay strength λ changes at $T=5$ and $\alpha=4$. The number of input nodes $N=50$. The arrows in (a) denote the routes of changing λ.

At the parameters used in Fig. 11(a), sample averaged simulations show that theory and simulation agree satisfactorily on both sides of the bump. However, theory predicts a relatively abrupt change of ε_g for α around 1.6, which is not observed in the simulation. This discrepancy may be partly due to the finite size effects, but we cannot preclude that

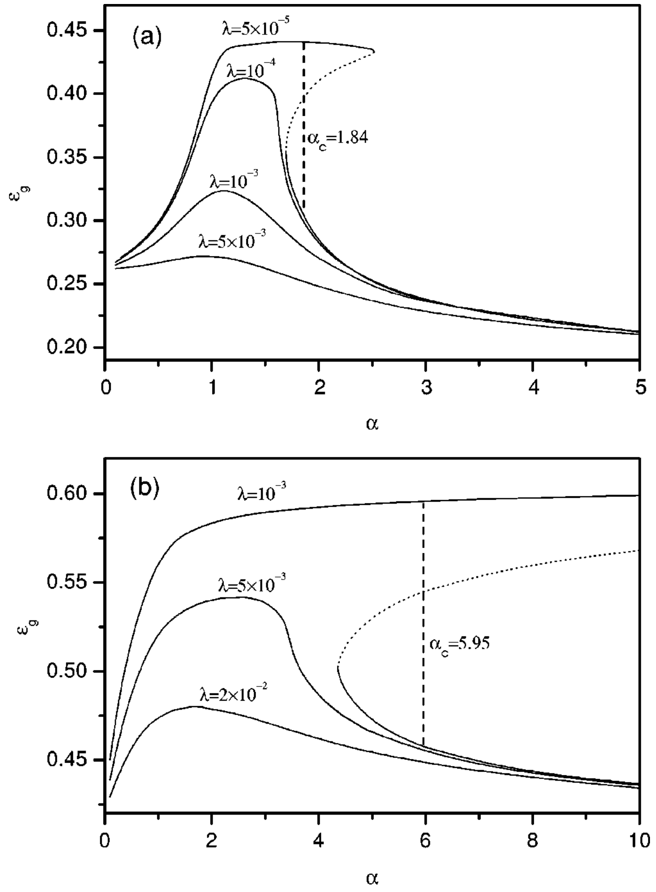

FIG. 10. The learning curves for different weight decay strengths: (a) $T=1$; (b) $T=5$.

effects of rough energy landscape (RSB) also contribute.

This discrepancy between theory and sample averaged simulations is also observed at the parameters used in Fig. 11(b) when discontinuous transitions exist. Hysteretic effects are shown by the different values of the transition points in the upward and downward directions of changing α, given by $\alpha_c^u(\lambda,T)=4.84$ and $\alpha_c^d(\lambda,T)=4.29$, respectively. The theoretical prediction of $\alpha_c(\lambda,T)$ is obtained in Fig. 12 from the intersection of the two branches of the energy curve. However, this prediction of $\alpha_c^t(\lambda,T)=5.95$ is higher than the position of hysteresis. Again, we attribute the discrepancy to finite size and the rough energy landscape.

We can interpret the effects of a rough energy landscape from the comparison between theoretical and simulation results. For small α, Fig. 6 shows that although a band gap exists in the activation distribution, the statistical weight of the outlying bands is only very small, thus the correction due to a rough energy landscape is minor, as can be seen in Fig. 11(b). When the size of the training set increases, the increasing weight of the outlying bands implies stronger effects of rough energy landscapes, which may account for the lowering of the critical α of the discontinuous transition in simulations when compared with the prediction of a smooth energy landscape. It is found that the smooth ansatz is stable for the branch of good generalization state in Fig. 12, but unstable for the poor one. Hence the introduction of the

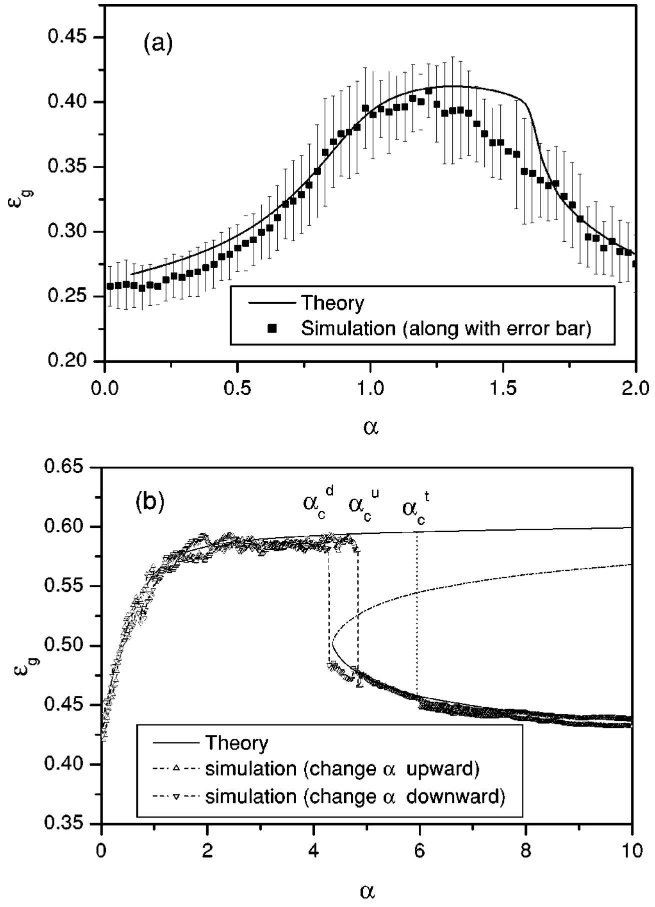

FIG. 11. Simulation versus theoretical results for the generalization error ε_g on changing α. For $T=1$ and $\lambda=0.0001$ (a), the simulation result is the average over 14 samples. For $T=5$ and $\lambda=0.001$ (b), the simulation result is the average over 20 samples on decreasing and increasing α. In all simulations, the number of input nodes $N=150$.

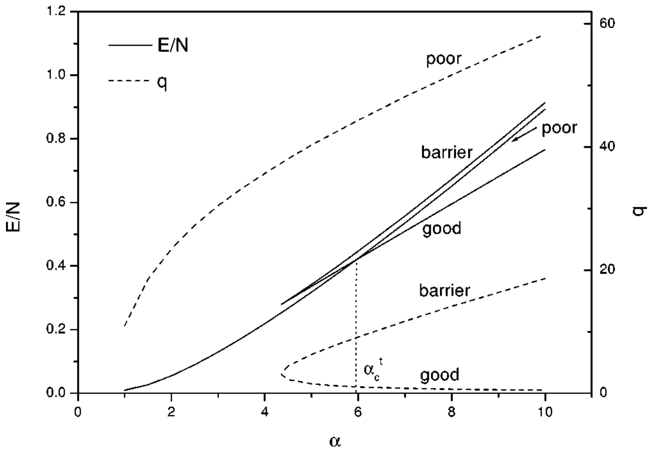

FIG. 12. The energy per nodes E/N (solid line) and the magnitude of student vector q (dotted line) versus the size α of training set, where $T=5$ and $\lambda=0.001$. The phase transition point is determined from the crossover of the two stable branches of the energy curve, $\alpha_c^t=5.95$, and the spinodal point of the good generalization state is at $\alpha_g^{sp}=4.36$.

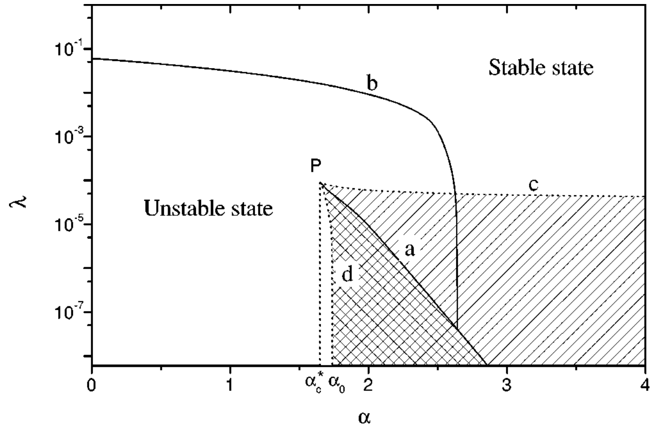

FIG. 13. The phase diagram for nonlinear perceptrons learning noisy examples when $T=1$. P is the critical point with $\alpha=\alpha_c^*=1.65$. Line d terminates at $\alpha=\alpha_0=1.74$ when λ approaches zero.

roughening effects will modify the energy curve of the poor generalization state, while that of the good one remains unchanged, thus shifting the position of the crossing point. The lowering of the α value in simulations implies that the energy of the poor generalization state is higher when we change from a smooth picture to a rough one. This is consistent with previous results that RSB increases the energy of similar perceptrons with discrete outputs [19,21].

The full phase diagram is drawn in Fig. 13 for a given noise temperature *T*. Above and below the *thermodynamic transition* line, line *a*, the perceptron is in the good and poor generalization phase respectively. Line *a* ends at the critical point *P*, where $\alpha=\alpha_c^*(T)$. The values of q and ε_g change discontinuously when the global parameters move across line *a*, but continuously when they move around point *P* without crossing line *a*. Line *b* denotes the stability line separating the regimes of smooth and rough energy landscapes. The rough regime covers the entire region left of the stability line as well as the entire poor generalization phase below line *a*. Here the position of line *a* is estimated assuming smooth energy landscape. Simulations such as those in Fig. 11(b) indicate that the effects of rough energy landscapes may shift its position leftwards. Line *c* is the *spinodal* line of the poor generalization phase, where $\lambda=\lambda_p(\alpha,T)$. The poor generalization phase is metastable in the shaded region bounded by the lines *c* and *a*. Similarly, line *d* is the *spinodal* line of the good generalization phase, where $\lambda=\lambda_g(\alpha,T)$, with the good generalization phase being metastable between lines *d* and *a*. When λ approaches zero, the abscissa of line *d* approaches $\alpha_0(T)$. Both lines *c* and *d* are computed in the smooth ansatz only, with roughening effects neglected.

It is interesting to consider the change of learning strategy in different regions of Fig. 13. In the region bounded by lines *c* and *d*, more than one learning strategies are competing against each other, corresponding to different local minima in energy. To the left of line *b*, all states adopt learning strategies that sacrifice a fraction of examples, but those with large q (poor generalization as shown in Fig. 12) sacrifice a significantly large fraction. To the right of line *b*, the competition takes places between states with large q, which sacrifice a fraction of examples, and those with small q (good generalization as shown in Fig. 12), which use a more even-handed strategy with no band gaps separating the activations. Around the phase transition line *a*, the globally minimal state switches on increasing α, from one with sacrificial strategy to a more even-handed one. This discontinuous change in learning strategies is illustrated from Figs. 6(a)–(d) to 6(e), where the phase transition line *a* is crossed over on increasing α for a given λ.

Outside the region with multiple states, the magnitude q of the student weight vector decreases above line *c* (since weight decay becomes strong) or to the left of the line *d* (since examples are not enough). Hence the weight vector is not flexible enough to allow for multiple strategies. In general, the fraction of sacrificed examples is smaller in this region. This reduces the difference between the strategies of sacrificing and not sacrificing the examples. As a result, all states to the left of line *b* learn with a single sacrificial strategy and to the right of it with a single even-handed strategy.

VI. CONCLUSION AND REMARKS

We have studied the supervised learning of noisy examples in nonlinear and differentiable perceptrons using the cavity method, yielding predictions identical to the replica method, yet providing a more physical interpretation. The mean-field equations enable us to study the macroscopic behavior of the system. An example is the optimal weight decay λ_{opt} that minimizes the generalization error, as illustrated in Fig. 8, analogous to previous studies in the linear perceptrons [15]. However, the emphasis of this paper is on phenomena attributable to the conflicting information inherent in noisy data, and the nonlinearity of the student perceptron. We have demonstrated the existence of band gaps in the activation distribution, separating preferred and sacrificed examples. It is an indication of the extent of information competition and the roughness of the energy landscape, corresponding to the effect of RSB in the replica approach. The more prominent the band gaps, the more significant the effects of rough energy landscapes. When tuning up the weight decay or increasing the size of the training set, a phase transition occurs in the student perceptron from a poor generalization state with a long weight vector to a good generalization state with a short weight vector. The phase transition is accompanied by a change in the learning strategy from sacrificial to even handed. We present the phase diagram of this system, together with the boundaries of the gapped regime and of the metastable region. The relation between band gaps and the picture of a rough energy landscape was discussed in a previous study [29]. Here we further show where this consideration is most necessary.

We remark that the preferential or sacrificial effects are common in many other learning systems, such as multilayer perceptrons [29] and weight pruning networks [34]. They create metastable states that cause the hysteretic behavior as shown in our simulations (see Figs. 9 and 11). The presence of metastable states prevent the convergence of dynamical learning process to the ground state. Hence it is an important issue in the practical implementation of learning dynamics.

We have illustrated that the cavity method can be used to analyze systems laden with conflicting information. It can be applied to other systems such as support vector machines (SVM) when examples are noisy and insufficient [35]. SVM learning of clean examples has recently been studied using the replica theory [36]. However, since the functional form of the energy is different, band gaps may not be present. Nevertheless, a cavity analysis of SVMs could offer new valuable insights.

ACKNOWLEDGMENTS

We thank H. Nishimori and Y. Kabashima for valuable discussions. This work is supported by the Research Grant Council of Hong Kong (HKUST6157/99P).

APPENDIX A: THE CAVITY ACTIVATION AND LOCAL SUSCEPTIBILITY

From Eqs. (3) and (7) and the definitions of t_0 and x_0, we obtain

$$x_0-t_0=\frac{1}{\lambda}(O_0-f_0^0)(f_0^0)'+\frac{1}{\lambda N}\sum_{\mu j}[(O_\mu-f_\mu^0)(f_\mu^0)'-(O_\mu-f_\mu)f_\mu']\xi_j^\mu\xi_j^0. \tag{A1}$$

Expanding the last term to first order, and assuming that x_μ is a well defined function of t_μ, we arrive at

$$\begin{aligned}x_0-t_0={}&\frac{1}{\lambda}(O_0-f_0^0)(f_0^0)'+\frac{1}{\lambda N\sqrt{N}}\\&\times\sum_{\mu j}[-(f_\mu')^2+(O_\mu-f_\mu)f_\mu'']\frac{\partial x_\mu}{\partial t_\mu}\xi_j^\mu\xi_j^0\\&\times\sum_{k(\neq j)}(J_k^{0\setminus\mu}-J_k^{\setminus\mu})\xi_k^\mu+\frac{1}{\lambda N\sqrt{N}}\sum_{\mu j}[-(f_\mu')^2\\&+(O_\mu-f_\mu)f_\mu'']\frac{\partial x_\mu}{\partial t_\mu}(J_j^{0\setminus\mu}-J_j^{\setminus\mu})(\xi_j^\mu)^2\xi_j^0,\end{aligned} \tag{A2}$$

where $J_k^{0\setminus\mu}$ and $J_k^{\setminus\mu}$ denote the student weights trained with training sets without example μ, and, respectively, with and without example 0. Note that $\Sigma_{k(\neq j)}(J_k^{0\setminus\mu}-J_k^{\setminus\mu})\xi_k^\mu/\sqrt{N}\approx t_\mu^0-t_\mu\sim O(N^{-1/2})$ and is uncorrelated with ξ_j^μ. Neglecting the dependence of $[-(f_\mu')^2+(O_\mu-f_\mu)f_\mu''](\partial x_\mu/\partial t_\mu)$ on $\xi_k^\mu\xi_j^\mu$ that is of order N^{-1}, we conclude that the second term on the right-hand side of Eq. (A2) is of order $N^{-1/2}$ and hence negligible. In the last term, $(\xi_j^\mu)^2$ is uncorrelated with $(J_j^{0\setminus\mu}-J_j^{\setminus\mu})\xi_j^0$, and hence can be replaced by its average value of 1. For the remaining summation over $j,\Sigma_j(J_j^{0\setminus\mu}-J_j^{\setminus\mu})\xi_j^0/\sqrt{N}$ reduces to $x_0^{\setminus\mu}-t_0^{\setminus\mu}$. Assuming that the change in the activation difference $x-t$ of examples 0 due to the removal of example μ is small, $x_0^{\setminus\mu}-t_0^{\setminus\mu}$ further reduces to x_0-t_0. Thus

$$x_0-t_0=\frac{1}{\lambda}(O_0-f_0^0)(f_0^0)'+\frac{1}{\lambda N}\times\sum_\mu[-(f_\mu')^2+(O_\mu-f_\mu)f_\mu'']\frac{\partial x_\mu}{\partial t_\mu}(x_0-t_0). \tag{A3}$$

Defining the local susceptibility γ by

$$\gamma^{-1}=\lambda+\frac{1}{N}\sum_\mu[(f_\mu')^2-(O_\mu-f_\mu)f_\mu'']\frac{\partial x_\mu}{\partial t_\mu}, \tag{A4}$$

we arrive at Eq. (8). Applying the same cavity argument to example μ,t_μ and x_μ should also be related by Eq. (8). This simplifies Eq. (A4) to

$$\gamma^{-1}=\lambda+\frac{\gamma^{-1}}{N}\sum_\mu\left(\frac{\partial t_\mu}{\partial x_\mu}-1\right)\frac{\partial x_\mu}{\partial t_\mu}, \tag{A5}$$

from which Eq. (9) follows.

APPENDIX B: THE STABILITY CONDITION

In obtaining Eq. (A2), the validity of the perturbative expansion in Eq. (A1) is subject to the condition that the fluctuation $\Delta_J=\Sigma_j(J_j^0-J_j)^2$ is finite. Subtracting Eq. (7) by Eq. (3), multiplying both sides by $J_j^0-J_j$ and summing over j, we obtain

$$\Delta_J=\frac{1}{\lambda\gamma}(x_0-t_0)^2-\frac{1}{\lambda\gamma}\sum_\mu\left(1-\frac{\partial x_\mu}{\partial t_\mu}\right)\frac{\partial x_\mu}{\partial t_\mu}(t_\mu^0-t_\mu)^2, \tag{B1}$$

where Eq. (8) is adopted, and x_μ is assumed to be a well defined function of t_μ. The factor $(t_\mu^0-t_\mu)^2$ in Eq. (B1) can be expanded as $\Sigma_{jk}(J_j^{0\setminus\mu}-J_j^{\setminus\mu})(J_k^{0\setminus\mu}-J_k^{\setminus\mu})\xi_j^\mu\xi_k^\mu/N$ and is only related with $(1-\partial x_\mu/\partial t_\mu)(\partial x_\mu/\partial t_\mu)$ in the order $O(N^{-1})$. Therefore, the average over μ of the former and latter's product can be replaced by the product of their averages. Since $(J_j^{0\setminus\mu}-J_j^{\setminus\mu})(J_k^{0\setminus\mu}-J_k^{\setminus\mu})$ are uncorrelated with examples $\boldsymbol{\xi}^\mu$, only terms with $j=k$ contribute to the average over μ. Then $\Sigma_\mu(t_\mu^0-t_\mu)^2$ becomes $\Delta_J^{\setminus\mu}$. Assuming that the change in Δ_J due to the removal of example μ is small, this further reduces to Δ_J and renders (B1) to

$$\Delta_J=\frac{1}{\lambda\gamma}(x_0-t_0)^2-\frac{1}{\lambda\gamma N}\sum_\mu\left(1-\frac{\partial x_\mu}{\partial t_\mu}\right)\frac{\partial x_\mu}{\partial t_\mu}\Delta_J.$$

Using Eq. (8), this can be further reduced to Eq. (21).

APPENDIX C: THE EFFECTS OF A GAP ON MEAN- FIELD EQUATIONS

When there is a gap in the distribution $P(x|\tilde{y})$, x_μ is no longer a differentiable function of t_μ, the mean-field equations (14), (17), and (18) are subject to modification.

In Eq. (A1), the summation over μ now includes different situations depending on the value of the cavity field t_μ. For those examples with t_μ and t_μ^0 located on the same side of the gap, the analysis is similar to that in Appendix A. However, if t_μ is close to the gap position t_g, then when the new example 0 is included in the training set, the change of cavity activation $\Delta t_\mu \equiv \Sigma_k (J_k^{0\setminus\mu} - J_k^{\setminus\mu})\xi_k^\mu/\sqrt{N}$ may give rise to large value of $(O_\mu - f_\mu^0)(f_\mu^0)' - (O_\mu - f_\mu)f'_\mu$ as the generic activation x_μ changes from $x_<(t_\mu)$ to $x_>(t_\mu^0)$ or from $x_>(t_\mu)$ to $x_<(t_\mu^0)$. We distinguish the following cases to calculate the summation in Eq. (A1).

The first case corresponds to $t_g - \Delta t_\mu < t_\mu < t_g$. Among the p examples, this happens with probability $\delta(t_\mu - t_g(\tilde{y}_\mu))\Delta t_\mu \theta(\Delta t_\mu)$. Its contribution to the summation in Eq. (A1) is

$$\sum_{\{\text{Case 1}\}} = \frac{1}{\lambda N} \sum_{\mu j} \delta(t_\mu - t_g(\tilde{y}_\mu))\Delta t_\mu \theta(\Delta t_\mu) \times \{[f(\tilde{y}_\mu) - f(x_\mu^>)]f'(x_\mu^>) - [f(\tilde{y}_\mu) - f(x_\mu^<)]f'(x_\mu^<)\}\xi_j^\mu \xi_j^0. \tag{C1}$$

Similarly, the second case corresponds to $t_g < t_\mu < t_g - \Delta t_\mu$, with the contribution

$$\sum_{\{\text{Case 2}\}} = \frac{1}{\lambda N} \sum_{\mu j} \delta(t_\mu - t_g(\tilde{y}_\mu))(-\Delta t_\mu)\theta(-\Delta t_\mu) \times \{[f(\tilde{y}_\mu) - f(x_\mu^<)]f'(x_\mu^<) - [f(\tilde{y}_\mu) - f(x_\mu^>)]f'(x_\mu^>)\}\xi_j^\mu \xi_j^0. \tag{C2}$$

Combining them together, we have the total contribution from the gap

$$\sum_{\{\text{Gap}\}} = \frac{\alpha}{\lambda}(x_0 - t_0) \int dy P(y) \int d\tilde{y} P(\tilde{y}|y) \times \int dt P(t|\tilde{y}) \delta(t - t_g(\tilde{y}))\{[f(\tilde{y}) - f(x_>(t,\tilde{y}))] \times f'(x_>(t,\tilde{y})) - [f(\tilde{y}) - f(x_<(t,\tilde{y}))] \times f'(x_<(t,\tilde{y}))\}. \tag{C3}$$

Simplifying the integrals, we have

$$\sum_{\{\text{Gap}\}} = \frac{\alpha}{\lambda} \frac{x_0 - t_0}{\sqrt{2\pi\left(q - \frac{R^2}{1+T^2}\right)}} \int Du \times \exp\left[-\frac{\left(t_g - \frac{R}{\sqrt{1+T^2}}u\right)^2}{2\left(q - \frac{R^2}{1+T^2}\right)}\right] \times \{[f(\tilde{y}) - f(x_>(t_g,\tilde{y}))]f'(x_>(t_g,\tilde{y})) - [f(\tilde{y}) - f(x_<(t_g,\tilde{y}))]f'(x_<(t_g,\tilde{y}))\}, \tag{C4}$$

with $\tilde{y} = \sqrt{1+T^2}u$. Therefore, we obtain the self-consistent equation (27) for γ and reproduce the function $t(x)$ in Eq. (8), where x is related to u and v by Eq. (15). The positions of band gap $t_g, x_<$ and $x_>$ are determined using the Maxwell's construction discussed in Sec. IV. Following Eqs. (27), (8), and (15), we get the equation of R with extra terms (28) and equation of q without extra term (18), after elaborate work on integrating by parts.

APPENDIX D: CONDITION FOR MAXWELL'S CONSTRUCTION

For a given teacher output $f(\tilde{y})$, x is a multivalued function of t when $t'(x) < 0$ at the inflection point $t''(x) = 0$. For the sigmoid function $f(x) = [1 + e^{-x}]^{-1}$, this implies

$$\frac{1}{2\gamma} < \frac{f^2(1-f)^2(1-3f+3f^2)}{1-6f+6f^2},$$
$$f(\tilde{y}) = \frac{f(4-15f+12f^2)}{1-6f+6f^2}, \tag{D1}$$

where f represents $f(x)$ at the inflection point. Thus, we obtain the condition of Maxwell's construction (30) if we define W as the parametric function of γ via

$$W = \frac{(2\mathfrak{f}-1)(1-12\mathfrak{f}+12\mathfrak{f}^2)}{2(1-6\mathfrak{f}+6\mathfrak{f}^2)},$$
$$\gamma = \frac{1-6\mathfrak{f}+6\mathfrak{f}^2}{2\mathfrak{f}^2(1-\mathfrak{f})^2(1-3\mathfrak{f}+3\mathfrak{f}^2)}. \tag{D2}$$

The function of $W(\gamma)$ for $\gamma > 0$ is plotted in Fig. 2.

[1] J.J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).

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

[3] H.S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1991).

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

[5] M. Mézard, J. Phys. A **22**, 2181 (1989).

[6] K.Y.M. Wong, Europhys. Lett. **30**, 245 (1995).

[7] M. Opper and O. Winther, Phys. Rev. Lett. **76**, 1964 (1996).

[8] E. Gardner, Europhys. Lett. **4**, 481 (1987); J. Phys. A **21**, 257 (1988).

[9] G. Györgyi, Phys. Rev. Lett. **64**, 2957 (1990).

[10] D. Amit, H. Gutfreund, and H. Sompolinsky, Ann. Phys.

(Leipzig) **173**, 30 (1987).

[11] K.Y.M. Wong and D. Sherrington, Phys. Rev. E **47**, 4465 (1993).

[12] M. Opper and W. Kinzel, in *Models of Neural Networks III*, edited by E. Domany, J. L. Vanhemmen, and K. Schulten (Springer, Berlin, 1996), p. 25.

[13] S. Bös, W. Kinzel, and M. Opper, Phys. Rev. E **47**, 1384 (1993).

[14] S. Bös, Phys. Rev. E **58**, 833 (1998).

[15] A.P. Dunmur and D.J. Wallace, J. Phys. A **26**, 5767 (1993).

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

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

[18] M. Bouten, J. Phys. A **27**, 6021 (1994).

[19] P. Majer, A. Engel, and A. Zippelius, J. Phys. A **26**, 7405 (1993).

[20] W. Whyte, D. Sherrington, and K.Y.M. Wong, J. Phys. A **28**, 7105 (1995).

[21] W. Whyte and D. Sherrington, J. Phys. A **29**, 3063 (1996).

[22] A. Krogh, J. Phys. A **25**, 1119 (1992); A. Krogh and J.A. Hertz, *ibid.* **25**, 1135 (1992).

[23] H. Schwarze, and J. Hertz, in *Neural Information Processing Systems*, edited by S. Hanson, J. Cowan, and L. Giles (Morgan-Kaufmann, San Mateo, CA, 1993), Vol. 5, p. 523.

[24] B. Schottky and U. Krey, J. Phys. A **30**, 8541 (1997).

[25] M. Ahr, N. Biehl, and E. Schlösser, J. Phys. A **32**, 5003 (1999).

[26] W. Kinzel, Philos. Mag. B **77**, 1455 (1998).

[27] A.H.L. West and D. Saad, J. Phys. A **30**, 3471 (1997).

[28] D.J. Thouless, P.W. Anderson, and R.G. Palmer, Philos. Mag. **35**, 593 (1977).

[29] K. Y. M. Wong, in *Neural Information Processing Systems*, edited by M. C. Mozer, M. I. Jordan, and T. Pestsche (MIT Press, Cambridge, MA, 1997) Vol. 9, p. 302; K. Y. M. Wong, S. Li, and P. Luo, in *Advanced Mean-field Methods: Theory and Practice*, edited by M. Opper and D. Saad (MIT Press, Cambridge, MA, 2001), p. 99.

[30] *Advanced Mean-field Methods: Theory and Practice*, edited by M. Opper and D. Saad (MIT Press, Cambridge, MA, 2001).

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

[32] J.R.L. de Almeida and D.J. Thouless, J. Phys. A **11**, 983 (1978).

[33] G. Parisi, Phys. Rev. Lett. **50**, 1946 (1983).

[34] K. Y. M. Wong, in *Theoretical Aspects of Neural Computation*, edited by K. Y. M. Wong, I. King, and D. Y. Yeung (Springer, Singapore, 1998), p. 93.

[35] V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 2000).

[36] R. Dietrich, M. Opper, and H. Sompolinsky, Phys. Rev. Lett. **82**, 2975 (1999).