Capacity of the multilayer perceptron with discrete synaptic couplings

Kazuo Nokura
Shonan Institute of Technology, Fujisawa 251, Japan
(Received 13 December 1993)

We study the multilayer perceptron with N discrete synaptic couplings. As a typical discrete set of values, we assume that synaptic couplings take $2L+1$ values k/L ($k=-L,\ldots,L$, L is an integer). Using the replica method, we study the space of solutions that implement prescribed P input-output relations. The property of the space of solutions is featured by two characteristic $\alpha \equiv P/N$ values; one is the point α_{AT} where the Almeida-Thouless instability takes place and the other is the point α_S where the entropy S of the replica symmetry (RS) theory vanishes. When the number of hidden units K is infinity, we find that the order of α_{AT} and α_S changes at $L=4$. For $L \geq 4$, the replica symmetry has to be broken with the finite entropy of solutions. For large L, we find that α_S is proportional to $(\ln L)^x$, where x is very close to $\frac{1}{2}$. The one-step replica symmetry breaking theory gives a smaller value for S than the RS theory does, but the difference is very small. For $K=3$, we find that α_S becomes larger than α_{AT} when $L \geq 2$. We also study numerically the space of solutions for $K=3$ using the least action algorithm modified for discrete coupling J. We find that some drastic change of the space of solutions really takes place around α_{AT}.

PACS number(s): 87.10.+e, 64.60.Cn

I. INTRODUCTION

Recently, many efforts have been made to study neural networks by the statistical mechanics approach. Of particular interest is the mechanism by which many neurons and synapses cooperate for macroscopic functions of neural systems. Among many models, the perceptron model is important, since it contains the basic concept of parallel processing of large numbers of inputs [1]. The simplest version of the perception is the single layer perceptron described as follows. Each input signal S_i ($i=1,2,\ldots,N$) is transmitted to the output unit through the synaptic coupling J_i. The output S_0 is given by the sign function of the local field $h=\sum_j J_j S_j$. Usually, we want to have prescribed outputs ξ_0^μ for prescribed inputs ξ_i^μ ($\mu=1,2,\ldots,P$). This is achieved by arranging $\{J_i\}$ by some learning algorithm. Thus the problem is to find the couplings which satisfy the relation

$$\xi_0^\mu = \mathrm{sgn}\left[\sum_j J_j \xi_j^\mu\right] \tag{1}$$

for all μ. The ratio $\alpha=P/N$ measures the number of patterns that the perceptron can memorize. There is an upper limit of α that allows the existence of solutions. This upper limit α_c, the critical capacity, is a fundamental quantity which features the ability of a machine and has been studied mainly with geometrical arguments [2,3].

A few years ago, Gardner introduced a statistical mechanics approach to this problem [4]. In this approach we use the replica method [5] and treat the couplings J_j as spin variables, and patterns ξ_j^μ as quenched random variables. Various modifications of the replica formulation allow us to study various aspects of neural networks.

Recently, this method was applied to more complicated multilayer perceptrons, i.e., parity machines [6] and committee machines [7,8]. These models have many hidden units which process the signals from input units and transmit their states to the output unit. For committee machines, the output S_0 is given by $\mathrm{sgn}(\sum_l \sigma_l)$, where σ_l is $\mathrm{sgn}(\sum_j J_{lj} S_j)$. Thus the problem is to find J_{lj}, such that

$$\xi_0^\mu = \mathrm{sgn}\left[\sum_{l=1}^{K} \sigma_{\mu l}\right], \qquad \sigma_{\mu l} = \mathrm{sgn}\left[\sum_j J_{lj} \xi_j^\mu\right] \tag{2}$$

for all μ, where K is the number of the hidden units. This is one of the simplest architectures that has hidden units. However, information about these models mainly comes from computer simulations. Also, it was pointed out that the problem of finding couplings for given patterns is NP complete [9]. In this respect, replica studies are significant in that they evaluate the properties of the machine.

According to [7], the behavior of the multilayer perceptron strongly depends on the type of synaptic couplings. First, the critical capacity α_c for continuous J is much larger than for Ising J. This feature is very different from the single layer perceptron, in which α_c varies only from 0.83 to 2.0, from Ising to continuous J [2,4,10]. We suspect that the ability of the multilayer perceptron is much more sensitive to the amount of information of each J_{lj} than the single layer perceptron. Second, in the continuous coupling, the replica symmetry breaking (RSB) takes place even for α smaller than α_c. This means that the space of solutions is divided into subregions above some α. On the contrary, in the Ising

case, there is no RSB for $\alpha<\alpha_c$ and the behavior is similar to the single layer case.

In this paper we study a two layer perceptron, introducing discrete synaptic coupling $J_{lj}=k/L$, where L is an integer and integer k varies from $-L$ to L. For the single layer perceptron, the replica symmetry (RS) theory for discrete coupling was studied in [11]. With discrete synaptic coupling like this, we can study the crossover from Ising-like to continuouslike behavior. We can also study the number of solutions or the entropy of solutions out of $(2L+1)^N$ configurations. It has been shown in [7] that, in the large K limit, the replica study is simplified significantly and α_c becomes infinity for continuous coupling. Thus, for discrete coupling with small $1/L$, we expect interesting behavior of α_c in the large K limit.

Section II is devoted to reviewing the replica calculations. We also study the number of solutions by the annealed approximation, which gives the upper bound of α_c. In Sec. III we discuss the solution of saddle point equations in the large L limit. Section IV is devoted to studying the numerical solution of RS and one-step RSB saddle point equations. Some results of the simulations for a small system are presented in Sec. V. Some comments and discussions will be given in Sec. VI.

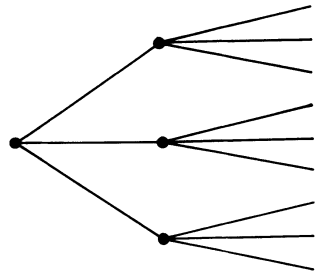

FIG. 1. Two layer perceptron with nonoverlapping input units, which we discuss in this paper. The number of hidden units is K. Each hidden unit has M input units. The total number of input synaptic couplings is $N=MK$.

II. FORMULATIONS FOR $K=1$, 3, AND ∞

A. Annealed approximation

One of the interesting features of discrete couplings is how the number of solutions depends on L and the architecture of the machine. From the argument in Sec. I, we expect that, in the single layer case, the α_c does not strongly depend on the mesh of coupling $1/L$ but does strongly depend on $1/L$ in the two layer case, especially for large K. To begin with, let us study the problem by the annealed approximation. In this paper, we concentrate on the nonoverlapping committee machine depicted in Fig. 1, in which the input units are divided into K groups of M input units and each hidden unit receives inputs from a corresponding group. ξ_j^μ are assumed to be ±1 with probability $\frac{1}{2}$. The volume of the space of solutions is given by

$$V=\sum_{\{J\}}\prod_{\mu}\theta\left[\sum_{l=1}^{K}\mathrm{sgn}(h_{\mu l})\right], \tag{3}$$

where $\theta(x)$ is 0 for $x<0$, 1 for $x\geq 0$, and

$$h_{\mu l}=\sum_j J_{lj}\xi_j^\mu\Big/\sqrt{M}\ . \tag{4}$$

If we set $K=1$, the expression reduces to the single layer case. With the annealed approximation, i.e., the ξ_i^μ average of V, we find

$$\langle V\rangle=\int_{-\infty}^{\infty}\int_0^{\infty}\prod_\mu\frac{dX_\mu d\Lambda_\mu}{2\pi}\exp\left[-i\sum_\mu X_\mu\Lambda_\mu\right]\cos^K X_\mu\prod_l\frac{dE_l dq_d^l}{2\pi}\exp\left[M\left[E_l q_d^l+\ln\sum_J\exp(-E_l J^2)\right]\right], \tag{5}$$

where we have used the δ function $\delta(Mq_d^l-\sum_j J_{lj}^2)$. In the saddle point approximation for large M, we obtain $E_l=0$ and

$$q_d^l=\sum_J J^2\exp-E_lJ^2/\sum_J\exp-E_lJ^2=(L+1)/3L\ ,$$

which is the average over the homogeneous distribution. Thus we obtain

$$\langle V\rangle=(\tfrac{1}{2})^P(2L+1)^N \tag{6}$$

for all odd K. Thus the entropy per synapse is given by $S=-\alpha\ln2+\ln(2L+1)$. By demanding $\langle V\rangle=1$, we can find the upper bound of the critical capacity, $\alpha_a\equiv P_a/N=\ln(2L+1)/\ln2$. This relation implies that the critical capacity is proportional to the bit number of a coupling. However, it is well known that for a single layer perceptron the set of patterns should be linear separable and $\alpha_c\leq 2.0$ [2,4]. In the following sections, we will discuss what happens for the two layer case by the replica method.

B. Replica method

This section is devoted to studying a two layer perceptron by the replica method. We will follow the formulation presented in [7]. The typical value of V is given by $\exp\langle\ln V\rangle$, where $\langle\ \rangle$ means the ξ_i^μ average. To evaluate $\langle\ln V\rangle$, we use the replica method,

$$\langle\ln V\rangle=\lim_{n\to0}\frac{\langle V^n\rangle-1}{n}\ . \tag{7}$$

In the saddle point approximation, the constraint from each pattern contributes independently to $\langle\ln V\rangle$ if we assume that the ξ_i^μ are independent of each other. However, hidden units couple strongly because of the con-

straints $\theta(\sum_l \sigma_l)$. This makes it difficult to study the problem for general K. An interesting finding in [7] is that, when $K \to \infty$, the expression for $\langle \ln V \rangle$ is quite simplified and becomes very similar to the single layer case.

Let us briefly review the derivation of $\langle \ln V \rangle$ for $K=1$, 3, and ∞. It is convenient to express $\langle V^n \rangle$ by the distribution function of local fields in the form

$$\langle V^n \rangle = \int \prod_{\mu l \alpha} dh_{\mu l}^{\alpha} \prod_{\mu \alpha} \theta \left(\sum_l \operatorname{sgn}(h_{\mu l}^{\alpha}) \right) P\{h_{\mu l}^{\alpha}\} . \quad (8)$$

Since we have assumed that the ξ_j^{μ} are independent of each other and each hidden unit has different inputs, $P\{h_{\mu l}^{\alpha}\}$ becomes a product over l and μ, which is given by

$$P\{h_{\mu l}^{\alpha}\} = \prod_{\mu l} P(h_{\mu l}^{\alpha}) = \sum_{\{J\}} \prod_{\mu l} \left\langle \prod_{\alpha} \delta \left[h_{\mu l}^{\alpha} - \sum_j J_{lj}^{\alpha} \xi_{lj}^{\mu} \Big/ \sqrt{M} \right] \right\rangle . \quad (9)$$

The average over ξ_j^{μ} induces couplings in the replica space. Introducing overlaps among J_{lj}^{α}, i.e., $q_{\alpha\beta}^{l}$ and q_{α}^{l}, we have

$$P(h_{\mu l}^{\alpha}) = \int \exp \left[i \sum_{\alpha} h_{\mu l}^{\alpha} y_{\mu l}^{\alpha} - \tfrac{1}{2} \sum_{\alpha} q_{\alpha}^{l} y_{\mu l}^{\alpha} y_{\mu l}^{\alpha} - \tfrac{1}{2} \sum_{\alpha \neq \beta} q_{\alpha\beta}^{l} y_{\mu l}^{\alpha} y_{\mu l}^{\beta} \right] \prod_{\alpha} \frac{dy_{\mu l}^{\alpha}}{2\pi}$$
$$\times \sum_J \prod_{\alpha < \beta} \delta \left[M q_{\alpha\beta}^{l} - \sum_j J_{lj}^{\alpha} J_{lj}^{\beta} \right] \prod_{\alpha} \delta \left[M q_{\alpha}^{l} - \sum_j J_{lj}^{\alpha} J_{lj}^{\alpha} \right] \prod_{\alpha < \beta} dq_{\alpha\beta}^{l} \prod_{\alpha} dq_{\alpha}^{l} . \quad (10)$$

A typical continuous coupling is achieved by the spherical constraint $q_{\alpha}^{l}=1$. In our case, q_{α}^{l}will be determined by the saddle point equation. The δ functions are further replaced by the integral form with the integral variables E_{α}^{l} and $F_{\alpha\beta}^{l}$. In the saddle point approximation for large M, q_{α}^{l}, $q_{\alpha\beta}^{l}$, E_{α}^{l}, and $F_{\alpha\beta}^{l}$ are assumed to be the saddle point value which gives the extremum of $\langle \ln V \rangle$. The simplest assumption is the symmetric one among replica indices and hidden units. Setting $q_{\alpha\beta}^{l}=q$, $q_{\alpha}^{l}=q_d$, $iE_{\alpha}^{l}=E$, and $-iF_{\alpha\beta}^{l}=F$, we obtain

$$\langle V^n \rangle = \prod_{\mu} \left\{ \int \prod_{l} Dt_{\mu l} \int \prod_{\alpha} \theta \left[\sum_{l=1}^{K} \operatorname{sgn}(h_{\mu l}^{\alpha}) \right] \prod_{l\alpha} \exp \left[-\frac{1}{2} \frac{(\sqrt{q}\, t_{\mu l} - h_{\mu l}^{\alpha})^2}{q_d - q} \right] \frac{dh_{\mu l}^{\alpha}}{\sqrt{2\pi (q_d - q)}} \right\} \exp NG , \quad (11)$$

where $Dt = \exp(-\frac{1}{2}t^2)/\sqrt{2\pi}$ and

$$G = nEq_d - \tfrac{1}{2}n(n-1)qF + \ln \int Du \left[\sum_J \exp \left[-\frac{2E+F}{2} J^2 + \sqrt{F}\, uJ \right] \right]^n . \quad (12)$$

By introducing the integral form for the θ function, we obtain

$$\langle V^n \rangle = \prod_{\mu} \left[\int \prod_{l} Dt_{\mu l} \int_{-\infty}^{\infty} \int_{0}^{\infty} \prod_{\alpha} \frac{dX_{\mu}^{\alpha} d\Lambda_{\mu}^{\alpha}}{2\pi} \exp \left[-i \sum_{\alpha} X_{\mu}^{\alpha} \Lambda_{\mu}^{\alpha} \right] \prod_{l\alpha} [\exp(iX_{\mu}^{\alpha}) H(X_{\mu l}) + \exp(-iX_{\mu}^{\alpha}) H(-X_{\mu l})] \right] \exp(NG) , \quad (13)$$

where $H(X) = \int_X^{\infty} \exp(-\frac{1}{2}x^2) dx / \sqrt{2\pi}$ and $X_{\mu l} = \sqrt{q}\, t_{\mu l} / \sqrt{q_d - q}$.

To proceed further, we should specify the value of K. For $K=1$ and 3, the integrals of X_{μ}^{α} and Λ_{μ}^{α} are straightforward. Dropping indices μ, the results for $K=1$ and 3 are given by

$$\langle \ln V/N \rangle = \alpha \mathcal{A} + Eq_d + \tfrac{1}{2}qF + \int Du \ln f(2E+F, \sqrt{F}\, u) , \quad (14)$$

$$f(x,y) = \sum_J \exp \left[-\frac{x}{2} J^2 + yJ \right] , \quad (15)$$

where

$$\mathcal{A} = \int Dt \ln H(X) , \quad (16)$$

with $X = X_1$ for $K=1$, and

$$\mathcal{A} = \int \prod_{l=1}^{3} Dt_l \ln(H_1 H_2 + H_1 H_3 + H_2 H_3 - 2H_1 H_2 H_3) , \quad (17)$$

for $K=3$, where $H_l = H(X_l)$.

In the $K \to \infty$ limit, the nontrivial result is obtained if we assume X_{μ}^{α} is order $1/\sqrt{K}$. This assumption allows us to expand $\exp(\pm iX_{\mu}^{\alpha})$ in (13). Picking up the terms up to K^{-1} we get the summations $\sum_l F_l/\sqrt{K}$ and $\sum_l F_l^2/K$ in the exponential, where $F_l = 2H_l - 1$. To the largest order of K, we can replace the latter by $Q = \int F^2 Dt$ and the former by $\sqrt{Q}\, T$, where T obeys the Gaussian distribution with mean 0 and variance 1. In this way, the integral over K variables

reduces to one variable, and we obtain $\mathcal{A}=\int DT \ln H(X)$ with $X=\sqrt{Q}\,T/\sqrt{1-Q}$, where $Q=1-(2/\pi)\arccos(q/q_d)$.

The replica symmetry breaking ansatz for $q^l_{\alpha\beta}$ and $F^l_{\alpha\beta}$ can be treated in the same way. For the one-step RSB, we assume that the matrix $q^l_{\alpha\beta}$ has elements q_1 in diagonal blocks of size $m\times m$ and elements q_0 among different blocks. Setting $F^l_{\alpha\beta}$ in the same manner, we obtain

$$\langle \ln V/N\rangle=\frac{\alpha}{m}\int DT \ln\int DT_I H(X)^m+Eq_d+\tfrac{1}{2}[(1-m)q_1F_1+mq_0F_0]$$
$$+\frac{1}{m}\int Du \ln\int Du_I f(2E+F_1,\sqrt{F_0}u+\sqrt{F_1-F_0}u_I)^m , \tag{18}$$

where

$$X=\frac{\sqrt{Q_0}T+\sqrt{Q_1-Q_0}T_I}{\sqrt{1-Q_1}} , \tag{19}$$

where $Q_1=1-(2/\pi)\arccos(q_1/q_d)$ and $Q_0=1-(2/\pi)\arccos(q_0/q_d)$. It is easy to write the one-step RSB expressions for $K=1$ and 3. If we set $q_d=1$ and replace J sum with J integral in (15), we obtain the expression for the spherical constraint case. The Almeida-Thouless (AT) instability can be studied in the same way as the spherical constraint case studied in [7], since the fluctuations of q^l_α and E^l_α do not contribute to this instability [12].

III. STUDY OF THE SADDLE POINT EQUATIONS IN THE SMALL q_d-q AND LARGE L LIMIT

To obtain the solutions of the saddle point equation, we should do some numerical works, which will be presented in Sec. IV. In this section, we study the solution in the small q_d-q and large L limit. In the $q_d\to q$ limit, we can take the singular part of the equations and study the behavior of α_S, which we believe is qualitatively correct apart from this limit. Let us first study the RS solution. The one-step RSB solution will be discussed in Appendix B. The saddle point equation for RS is given by

$$F=\frac{\alpha}{1-Q}\int DT\frac{1}{H(X)^2}\left[\frac{dH}{dX}\right]^2\frac{2}{\pi}\frac{1}{\sqrt{q_d^2-q^2}} ,$$
$$E=-\frac{1}{2}\frac{q}{q_d}F ,$$
$$q_d=\int du\langle J^2\rangle , \tag{20}$$
$$q=\int du\langle J\rangle^2 ,$$

where

$$\langle\cdots\rangle=\frac{\sum_J\cdots\exp\left[-\frac{2E+F}{2}J^2+\sqrt{F}uJ\right]}{\sum_J\exp\left[-\frac{2E+F}{2}J^2+\sqrt{F}uJ\right]} . \tag{21}$$

In Appendix A, we present the details of the evaluations for small $\Delta q_d\equiv q_d-q$ and large L. The result is given by

$$F=\alpha C\frac{1}{\Delta q_d^{3/2}} ,$$
$$2E+F=\alpha C'\frac{1}{\sqrt{\Delta q_d}} ,$$
$$q_d=C'' , \tag{22}$$
$$\Delta q_d=C'''\frac{1}{2E+F} ,$$

where all C's are constants without singularities. These equations are derived under two assumptions. The first is that $(2E+F)/\sqrt{F}$ does not strongly depend on Δq_d and tends to a limiting value r for $\Delta q_d\to 0$. The solution $(2E+F)/\sqrt{F}=\sqrt{C'C'''/C}$ justifies this assumption. The second is that the summation over J in (15) can be approximated by the integral even if $2E+F$ becomes large. To see if it is true, we should study the change of the terms in (15) under the change $J\to J+1/L$. If this value $(2E+F)/L\sim\sqrt{F}/L$ is small, this approximation will be good. Solving the above equations, we obtain $\alpha=C'''/(C'\sqrt{\Delta q_d})$. The α which gives $(2E+F)/L\sim 1$ is of the order of $\sqrt{L}$, which is much larger than the upper limit $\alpha_a=\ln(2L+1)/\ln 2$. Thus, for $\alpha<\alpha_a$, we can work with the integral approximation.

Using these solutions, we obtain the entropy, i.e., the value of $\langle \ln V/N\rangle$ at the saddle point, given by

$$S_{\rm RS}=-B\alpha^2+\ln L , \tag{23}$$

where B is a complicated function of r and q_d, which is given in Appendix A. $S_{\rm RS}$ becomes 0 when $\alpha=\alpha_S\equiv\sqrt{\ln L/B}$. Thus α_S increases as L increases, but it is much smaller than $\alpha_a=\ln(2L+1)/\ln 2$.

In Appendix B, we studied the saddle point equation of the one-step RSB. We found that the solution $\alpha\propto 1/\sqrt{q_d-q_1}$ is consistent with the saddle point equation. As discussed in Appendix B, this result implies also that $S_{\rm RSB}$ becomes zero at the point $\alpha_{S(\rm RSB)}\propto\sqrt{\ln L}$. In Sec. IV we present the numerical solution of the saddle point equations and find that the ratio $\alpha_S/\sqrt{\ln L}$ is nearly constant for rather small L but continues to increase weakly as L increases. We will also find that the difference of entropy between RS and RSB is quite small, which implies $\alpha_S\sim\alpha_{S(\rm RSB)}$.

To finish this section, we should comment on the single layer perceptron, i.e., $K=1$. In this case, Q is given by q/q_d. Repeating the same argument for the small Δq_d

limit, we find that F and $2E+F$ are proportional to $1/\Delta q_d^2$ and $1/\Delta q_d$, respectively. Thus α becomes a constant for $\Delta q_d \to 0$. This means that $2E+F$ diverges as α tends to this value and the integral approximation for $f(x,y)$ becomes wrong. As we will see in Sec. IV, S of the single layer perceptron rapidly becomes zero just below $\alpha=2.0$, which is the critical capacity of the continuous coupling.

IV. NUMERICAL STUDY OF THE SADDLE POINT EQUATION

A. L dependence of α_{AT} and α_S

According to [7], the behavior of a multilayer perceptron strongly depends upon the type of coupling, that is, for continuous J, α_c is infinity and the AT instability takes place for finite α, but for Ising J, α_c is very close to the single layer case and there is no AT instability for $\alpha<\alpha_c$. In this section, we first study the L dependence of two characteristic α's. One is the point α_S where S_{RS} vanishes. This α is studied in Sec. III in the large L limit. The other is the point α_{AT} where the AT instability of the RS solution takes place. α_{AT} is determined by the equation

$$1-\alpha_{AT}K\gamma_P\gamma_J=0\ , \tag{24}$$

where γ_P and γ_J are given in Appendix C. If $\alpha_S<\alpha_{AT}$, α_{AT} is fictitious and we can identity $\alpha_S=\alpha_c$. If $\alpha_S>\alpha_{AT}$, RSB takes place with positive entropy and we should study the entropy of the RSB solution to find α_c.

In Fig. 2, the L dependence of α_{AT} and α_S is presented for $K=3$ and ∞. At small L, their order reverses and the RSB region appears for larger L. α_{AT} of $K=3$ and ∞ tend to 1.6 and 2.72, respectively, for large L. These values are slightly smaller than in the spherical constraint case. On the other hand, α_S increases rapidly as L increases. α_S for $K=3$ should have an upper limit, which is not reached in the range of the studied L. Although α_S for large L is only an approximation of α_c, it gives a good idea about α_c since, as discussed below, S_{RSB} is very close to S_{RS}.

To study the L dependence of α_S, we present $\alpha_S/\sqrt{\ln L}$ and $\alpha_S/\ln L$ for $K=\infty$ in Fig. 3. $\alpha_S/\sqrt{\ln L}$ seems to be nearly constant for small L. However, we found that $\alpha_S/\sqrt{\ln L}$ continues to increase very weakly up to $L=32$, which is the maximum L we have studied. This behavior is slightly different from the results obtained in Sec. III. The discussion in Sec. III is based upon the assumption that Δq_d is very small. However, we cannot control Δq_d at $\alpha \sim \alpha_S$. As we will discuss in Sec. IV B, Δq_d is not so small for $\alpha \sim \alpha_S$.

The behavior of entropy for $K=1$, 3, and ∞ is presented in Fig. 4 for $L=8$. It changes systematically, although each line is obtained from different saddle point equations. The point at which S_{RS} becomes zero is $\alpha=1.84$, 2.83, and 3.40, respectively. The arrows in the figure indicate the AT points. Above these points, we

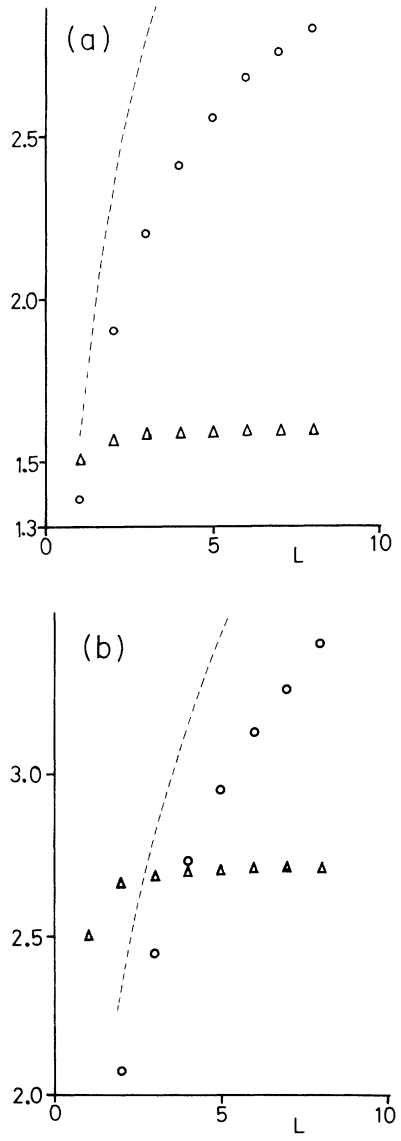

FIG. 2. L dependence of α_S and α_{AT} for (a) $K=3$ and (b) $K=\infty$. Δ for α_S and $\circ$ for α_{AT}. The crossover from Ising-like to continuous-like behavior takes place near $L\sim 2$ and 4, respectively. The broken line is the upper bound of α_c given by $\alpha_a=\ln(2L+1)/\ln 2$.

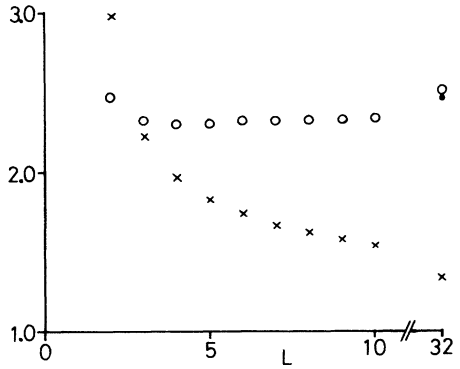

FIG. 3. L dependence of $\alpha_S/\ln L$, $\times$, and $\alpha_S/\sqrt{\ln L}$, $\circ$. The same data as in Fig. 2 are used. For $L=32$, $\bullet$ is $\alpha_{S(RSB)}/\sqrt{\ln L}$.

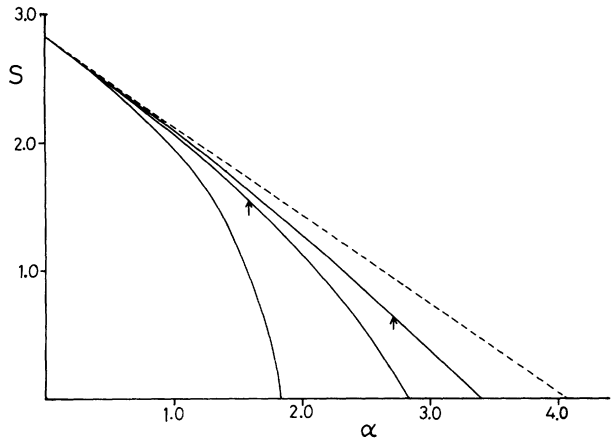

FIG. 4. α dependence of entropy S derived by RS theory, for $L=8$ and $K=1$, 3, and ∞ from bottom. The broken line is $S_a=-\alpha\ln 2+\ln 17$. The arrows indicate the points where the AT instability takes place.

found that the RSB entropy is always smaller than that of RS. However, the difference is too small to be presented in the same figure. For $K=\infty$, we have studied the difference up to $L=32$ and the largest difference is of order 10^{-2} near the point $S_{\mathrm{RSB}}=0$. The value $\alpha_{s(\mathrm{RSB})}/\sqrt{\ln L}$ for $L=32$ is shown by a dot in Fig. 3.

B. α dependence of order parameters

In Fig. 5 we present the α dependence of order parameters $K=\infty$ and $L=8$. RSB takes place at $\alpha_{\mathrm{AT}}\sim 2.72$, with $q_d=0.371$ and $q=0.227$. For $\alpha<\alpha_S$, the qualitative behavior of the order parameters is very similar to the spherical case [7]. For the wide range of α, the self-overlap q_d is very close to $(L+1)/3L=0.375$, which is the average of J^2 over the homogeneous distribution. To evaluate α_S analytically in Sec. III, we assumed that Δq_d was small enough at $S_{\mathrm{RS}}=0$. However, the numerical results showed that Δq_d is not so small for the RS solution, i.e., $q_d=0.365$ and $q=0.249$ at $\alpha=3.4$. We suppose that this is the main reason why $\alpha_S/\sqrt{\ln L}$ depends upon L weakly, even for large L. On the other hand, for RSB, we found $q_d=0.356$, $q_1=0.323$, and $q_0=0.220$ at $\alpha=3.4$, which gives a rather small q_d-q_1. Thus we expect that the small q_d-q_1 approximation is better for RSB than for RS in this region. As commented upon in Sec. IV A, S_{RSB} is slightly smaller than S_{RS}, giving smaller $\alpha_{S(\mathrm{RSB})}$. This seems to suppress the L dependence of $\alpha_{S(\mathrm{RSB})}/\sqrt{\ln L}$ for large L.

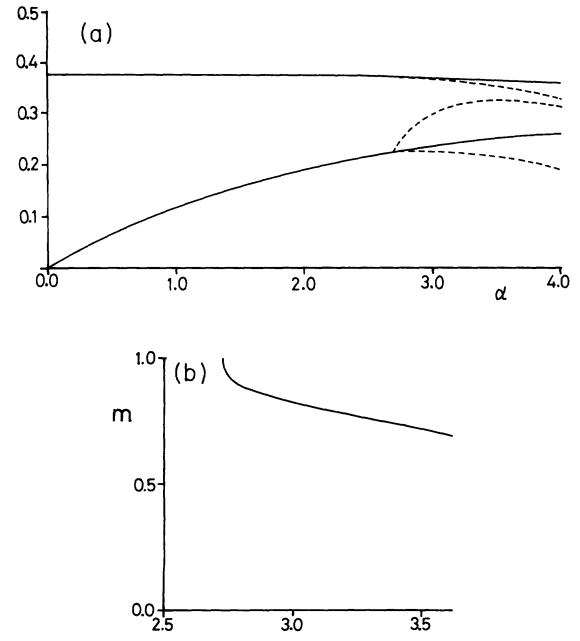

FIG. 5. α dependence of order parameters (a) q_d, q_1, q_0, and (b) $m=1-\sum_\alpha P_\alpha^2$ for $L=8$ and $K=\infty$. RSB takes place at $\alpha=\alpha_{\mathrm{AT}}\sim 2.72$. m starts from 1 at $\alpha=\alpha_{\mathrm{AT}}$ and decreases rapidly.

One interesting aspect is, as pointed out in [7], that the parameter m starts from 1 at the AT point and decreases rapidly. This is very different from the one-step RSB of the infinite range spin glass model [13]. According to the general theory of RSB [5,10], m is related to the weight of the pure state P_α by $m=1-\sum_\alpha P_\alpha^2$. This relation and the behavior of our m imply that, just after RSB takes place, the space of the solution breaks into many pure states with a small weight, but most of them become negligible as α increases.

V. REPLICA STUDIES AND SIMULATIONS FOR $K=3$

In this section, we describe the simulations for small systems. We concentrate on the $K=3$ case. The numerical study of the RSB phase is not easy because the space of solutions has many subregions, which are properly described by the overlap function [5]. The study of this function is outside the scope of the present paper. Here we want to observe some qualitative change of the space of solutions around α_{AT}.

Let us first review the replica studies for $K=3$. α_S, α_{AT}, and S_{RS} have been presented in Figs. 2 and 4. α_{AT} tends to 1.6, while α_S continues to increase for the studied L. We have studied the $L=8$ case, which has a wide region of RSB. In Fig. 6, the α dependence of the order parameters is shown with the results of the simulations, which we shall discuss below. At α_{AT}, we found $q_d\sim 0.37$ and $q\sim 0.22$ in the RS theory. This figure also shows the one-step RSB solution denoted by ○. Although they are only qualitative because of the numerical uncertainty, the behavior of the order parameters, including m, is very similar to the $K=\infty$ case.

The general approach in obtaining solution couplings is by simulated annealing (SA). Taking the number of wrong outputs as a cost function, we studied the system by this method. However, we found that it is not efficient in our case, especially when α is close to the AT point; usually, the cost function decreases to some small integer value but takes an enormous amount of time to become zero or does not become zero during the simulation time. This is mainly because the cost function is a step function, and diffusion in the configuration space is very slow.

Another choice is the least action algorithm (LAA)

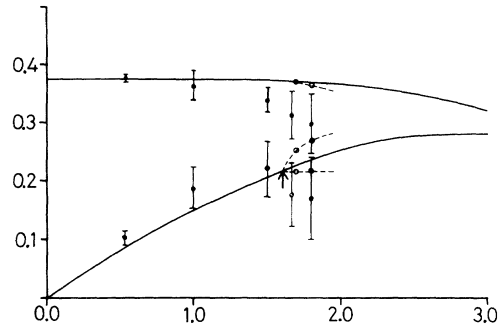

FIG. 6. The results of replica calculation for $K=3$ and $L=8$, and the results of simulations for $K=3$, $L=8$, and $M=5$. In this case, replica theory implies $\alpha_{AT}\sim 1.6$. In the simulations, the number of samples, i.e., the number of sets of patterns, is 10 for $P=8$, 15, 23. For $P=25$ and 27, we studied 30 samples but we did not obtain the solutions for 5 and 4 samples, respectively. The error bar is the sample fluctuation. With 10^5 MC steps and $T=1/3.0$, about 10^3-10^4 configurations contribute to the averages for $P=27$. The RSB solutions are given by ○. The broken lines are to guide the eye.

modified for discrete J. In general, the least action algorithm goes as follows [3,7]. First, we assume some initial values for the couplings. Then we see the output of a pattern ν and, if it is right, go to the next pattern. If it is wrong, we find the hidden unit l, which has the negative local field with the smallest absolute value. Then, for this unit l, we do the Hebb learning, i.e., $J_{lj}\to J_{lj}+1/\sqrt{M}\,\xi_j^\nu$ $(j=1,2,\ldots,M)$. However, when couplings are discrete, change of this order is not allowed. If we adopted $(1/L)\xi_j^\nu$ for the change instead of $(1/\sqrt{M})\xi_j^\nu$, it would make the change of $h_{l\nu}$ much larger than $h_{l\nu}$ itself. We suppose that natural modification is to introduce the stochastic processes, that is, we choose J_{lj} with probability $1/\sqrt{M}$ and $J_{lj}\to J'_{lj}=J_{lj}+1/L\xi_j^\nu$, if the condition $|J'_{lj}|\le 1$ is satisfied. With this modification, the change of $h_{l\nu}$ becomes of order $1/L$ with respect to $h_{l\nu}$. We assume that the initial couplings take random values k/L with $k=-L,\ldots,L$. To avoid the ambiguity coming from the zero local field, we add a small random threshold to each hidden unit. This does not modify the results of replica calculations. One period of checking every pattern once is called a session. After obtaining the solution, we do the Monte Carlo simulation (MCS) to do the average over the solution space.

We have studied $\overline{q_d}$ and $\bar{q}$, which are defined by

$$\overline{q_d}=\sum_{\mathrm{MC}}\sum_{lj}J_{lj}^2/(NN_s)\,, \tag{25}$$

$$\bar{q}=\sum_{lj}\left[\sum_{\mathrm{MC}}J_{lj}/N_s\right]^2/N\,, \tag{26}$$

where N_s means the number of MC configurations that satisfy the solution condition and $\sum_{\mathrm{MC}}$ means the summation over them. For finite temperature, N_s is smaller than the number of MC steps. Let us describe the parameters of our simulations. We have chosen $(K,M)=(3,5)$ and $L=8$. The number of LAA sessions needed to find a solution is about $10^2\sim 10^3$ if we can obtain the solution. The MC average is performed with 10^5 MC steps and temperature $T=1/3.0$. If the temperature is very low or zero, the MC average is not effective, especially for $\alpha\sim\alpha_{AT}$. This means that $\bar{q}$ is very close to $\overline{q_d}$, and we get little information about the space of the solutions. If the temperature is high, almost all configurations are out of the space of the solutions. With $T=1/3.0$, we found that about 10–30% of the MC steps are in the solution space.

Figure 6 shows the numerical results of $\overline{q_d}$ and $\bar{q}$. For $\alpha<\alpha_{AT}$, the results of the numerical simulation are slightly larger for $\bar{q}$ and smaller for $\overline{q_d}$ than the replica results for q and q_d, yet the agreement seems to be good except near the AT point.

For $\alpha>\alpha_{AT}$ we should be careful to compare the numerical results with the RSB theory, since the space of the solutions is divided into many subregions. If the configuration of the couplings diffuses only within each subregion, our numerical value $\bar{q}$ corresponds to q_1 of the replica theory. This situation is realized by zero temperature. In our simulations, the temperature is rather high and we have observed many nonsolution configurations in the MC steps. We also know that m is close to 1 for $\alpha\sim\alpha_{AT}$. As discussed above, this means that there are many subregions of similar weight. Thus we suppose that the coupling configuration diffuses among many subregions and the contributions to $\bar{q}$ from the same subregions are small. For this reason, we suppose that our numerical results for $\bar{q}$ correspond to q_0 rather than q_1 of the replica theory. In Fig. 6, $\bar{q}$ stops increasing at some point between $\alpha=\frac{23}{15}$ and $\frac{25}{15}$ and the sample fluctuation becomes quite large for $\alpha\ge\frac{25}{15}$. This implies that the structure of the space of solutions changes drastically between these α. This agrees with the replica theory which predicts the AT instability at $\alpha=1.6$. For $\alpha\ge\frac{25}{15}$, the deviation from the RSB results is rather large. We suspect that it is due to the finite size effect, yet there is a possibility of more steps of RSB in the replica theory.

VI. DISCUSSION

We have studied the two layer perceptron with discrete synaptic couplings. The behavior of order parameters for large L is similar to the continuous models, yet our model does not allow the large value of couplings, as discussed in the single layer model [11], and is different quantitatively from the spherical constraint case. For $K=\infty$, assuming large L and small Δq_d, we studied the most singular terms of the saddle point equation, giving the relation $\alpha_S\propto\sqrt{\ln L}$. The numerical study of the saddle point equation reveals that Δq_d is not so small and $\alpha_S/\sqrt{\ln L}$ is a weakly increasing function of L. Entropy from the one-step RSB theory was also studied, giving results very close to the RS results. We suspect that the further introduction of RSB steps will not alter the results strongly. This point is an open question.

For both $K=3$ and ∞, we found that the order of α_{AT} and α_S changes at rather small $L=2-4$. There appears a RSB region for a smaller mesh $1/L$. We expect a similar property for moderate K. This situation is quite different

from the infinite range spin glass model, in which we can understand the structure of the configuration space by cooperative flips of Ising variables [14]. In this respect, it will be interesting to study the relation between L and the number of steps of RSB required to give a stable replica theory. We should note that at least another one step of RSB should be introduced because of the negative entropy for $\alpha > \alpha_{S(\mathrm{RSB})}$. Finite temperature formulation may be required if there is no solution at all.

Figure 4 shows the α dependence of the entropy for $L=8$. It is illuminating to consider how this figure changes when L changes. As L increases, the entropy of each K increases like $\ln L$ for small α. For large α, however, the $K=1$ and 3 cases have upper limits of α beyond which the entropy is negative for any L. This upper limit is 2.0 for $K=1$ and about 3.0 for $K=3$ [7]. For $K=\infty$, because of the relation $\alpha_S \propto \sqrt{\ln L}$, entropy is expected to become positive for any α if L is large enough. For moderate K, we expect that α_c increases like $\sqrt{\ln L}$ and tends to the upper bound $\ln K$ [3] for finite L.

We have used the LAA that is modified for discrete couplings to obtain the solutions. It works well for moderate α. In general, it is very difficult to study the solutions numerically, especially for α close to the critical capacity α_c. An exhaustive search would be the most reliable method, but it would require an enormous amount of computer time. The observations about Fig. 4 given above suggest an interesting possibility about the space of the solutions. Imagine that one studies the system with a mesh $1/L$ and does not find any solution but instead finds some configurations which have a very small cost function. Then a question arises: To obtain the solutions with a smaller mesh, say $1/2L$, are the configurations obtained above good starting points for obtaining the solutions for a mesh $1/2L$? We suppose that it is natural to expect that two configurations correlate strongly.

In this paper, we did not discuss the committee machine with overlapping inputs. When hidden units share inputs, inner products of couplings among hidden units appear as order parameters in the replica theory. However, it was shown that these order parameters are negative and of order $1/K$ in the RS ansatz [8]. This is reasonable because, to achieve large capacity, hidden units will tend to project input signals onto planes which are as different as possible. Thus, in the large K limit, we expect that our arguments will not be strongly modified for the overlapping input case, at least in the RS ansatz.

Finally, we want to comment on the internal representation of the system. Having the hidden units, the two layer perceptron has many internal representations, $\sigma_{\mu l}=\mathrm{sgn}(h_{\mu l})$, for a given pattern ξ_j^ν. For $\alpha<\alpha_{\mathrm{AT}}$, the space of the solutions is connected and a set of $\sigma_{\nu l}$ will change gradually when couplings change gradually. The average $[\sigma_{\nu l}]$, where [ ] is the average over the solution J, gives some idea about the fluctuation of the internal representation. For the RSB, the space of the solutions is divided into many subregions. Corresponding to each subregion, a pattern will have different internal representations that are not continuously connected. In the framework of replica theory, the correlation among different internal representations is given by $C_{\alpha\beta}=\sum_l [\sigma_{\nu l}]_\alpha [\sigma_{\nu l}]_\beta / K$. This value will have hierarchical structure if replica symmetry breaks hierarchically. This is an interesting possibility of perceptrons with the hidden units.

APPENDIX A

In this appendix, we discuss the $q_d \to q$ and L-large limit of the RS saddle point equations. In the $q_d \to q$ limit, $1-Q$ tends to $(2/\pi)\sqrt{2(1-q/q_d)}$. Using the approximation $H(X)\sim 1$ for $X<0$ and $H(X)\sim \exp(-\frac{1}{2}X^2)/(\sqrt{2\pi}X)$ for $X>0$, F and E in (20) are reduced to the form

$$F=\alpha\frac{\pi}{8}\left[\frac{q_d}{2}\right]^{1/2}\frac{1}{(\Delta q_d)^{3/2}}\ , \tag{A1}$$

$$2E+F=\alpha\frac{\pi}{8}\left[\frac{1}{2q_d}\right]^{1/2}\frac{1}{\sqrt{\Delta q_d}}\ . \tag{A2}$$

To discuss the q_d and q in (20), it is convenient to start with the last term of (14), which is given by

$$\ln z=\int Du\,\ln f(2E+F,\sqrt{F}u)\ , \tag{A3}$$

$$f(2E+F,\sqrt{F}u)=\sum_J \exp-\frac{2E+F}{2}J^2+\sqrt{F}uJ\ . \tag{A4}$$

If the change of a term in this sum under the change $J\to J+1/L$ is small enough, we approximate this sum by the integral

$$f(2E+F,\sqrt{F}u)=L\int_{-1}^{1}dJ\exp-\frac{2E+F}{2}J^2+\sqrt{F}uJ\ . \tag{A5}$$

The largest contribution comes from the point $J=u\sqrt{F}/A$ if $|u|\le A/\sqrt{F}$, $J=1$ if $u>A/\sqrt{F}$ and $J=-1$ if $u<-A/\sqrt{F}$, where we denote $A=2E+F$. Using these expressions, we obtain

$$\ln z=(1-2I_2)\frac{F}{2A}+2I_1\sqrt{F}+I_0 A+\ln L\ , \tag{A6}$$

where

$$I_k=\int_{A/\sqrt{F}}^{\infty}u^k Du\ . \tag{A7}$$

By differentiating $\ln z$, we can obtain q_d and q in terms of A and F. The results are complicated functions of A and $r\equiv A/\sqrt{F}$. If we assume that r tends to some constant, we obtain the expressions in (22). To study the entropy, we should go back to $\ln V/N$ and express it by q_d and Δq_d. Using (A2) and noting $q_d E+(1/2)qF=0$, we obtain

$$\ln V/N=-\alpha C\frac{1}{\sqrt{\Delta q_d}}+f_1(y)\frac{1}{\Delta q_d}+f_2(y)\frac{\sqrt{\alpha}}{\Delta q_0^{3/4}}+f_3(y)\frac{\alpha}{\Delta q_d^{1/2}}+\ln L\ , \tag{A8}$$

where $y=\sqrt{\alpha\sqrt{\Delta q_d}}$ and $f_k(y)$ are given by the integrals I_k. We can easily notice that this equation has the ex-

tremum with the form $\alpha=a/\sqrt{\Delta q_d}$, where a is a constant given by the extremum condition. Using this a, we finally obtain

$$S=-B\alpha^2+\ln L \ , \tag{A9}$$

where

$$B=-\alpha^2\left[\frac{C}{a}-f_1(\sqrt{a})\frac{1}{a^2}-f_2(\sqrt{a})\frac{1}{a^{3/4}} -f_3(\sqrt{a})\frac{1}{a}\right] . \tag{A10}$$

APPENDIX B

In this appendix, we discuss the saddle point equation of the one-step RSB theory. As in Appendix A, we concentrate on the limit $\Delta q_d \equiv q_d - q_1 \to 0$. From (18), we obtain

$$\begin{aligned} E&=-\alpha\frac{1}{2}\frac{m}{1-Q_1}\int DT\frac{H_B^2}{H_A^2}\frac{2}{\pi}\frac{1}{\sqrt{q_d^2-q_0^2}}\frac{q_0}{q_d} \\ &\quad -\alpha\frac{1}{2}\frac{1-m}{1-Q_1}\int DT\frac{H_C}{H_A}\frac{2}{\pi}\frac{1}{\sqrt{q_d^2-q_1^2}}\frac{q_1}{q_d} , \\ F_1&=\alpha\frac{1}{1-Q_1}\int DT\frac{H_C}{H_A}\frac{2}{\pi}\frac{1}{\sqrt{q_d^2-q_1^2}} , \\ F_0&=\alpha\frac{1}{1-Q_1}\int DT\frac{H_B^2}{H_A^2}\frac{2}{\pi}\frac{1}{\sqrt{q_d^2-q_0^2}} , \end{aligned} \tag{B1}$$

$$\begin{aligned} q_d&=\int Du\frac{1}{P}\int Du_I f^m\langle J^2\rangle , \\ q_1&=\int Du\frac{1}{P}\int Du_I f^m\langle J\rangle^2 , \\ q_0&=\int Du\frac{1}{P^2}\left[\int Du_I f^m\langle J\rangle\right]^2 , \end{aligned} \tag{B2}$$

and

$$\tfrac{1}{2}(q_1F_1-q_0F_0)=\frac{\alpha}{m}\int DT\frac{1}{+}H_A\int DT_I H^m\ln H-\frac{\alpha}{m^2}\int DT\ln H_A+\frac{1}{m}\int Du\frac{1}{P}\int Du_I f^m\ln f-\frac{1}{m^2}\int Du\ln P , \tag{B3}$$

where

$$H_A=\int H^m DT_I,\quad H_B=\int H^{m-1}\frac{dH}{dX}DT_I ,\quad H_C=\int H^{m-2}\left[\frac{dH}{dX}\right]^2 DT_I,\quad P=\int f^m Du_I ,$$

The $\Delta q_d \equiv q_d - q_1 \to 0$ limit of H_A, H_B, and H_C are given by

$$H_A=H(X_0)+J_0,\quad H_B=J_1 ,\quad H_C=J_2 , \tag{B4}$$

where $X_0=\sqrt{Q_0}T/\sqrt{1-Q_0}$ and

$$J_k=\int_{-X_0}^{\infty}(\sqrt{2\pi}X)^{k-m}\exp-\tfrac{1}{2}c(\sqrt{Q_0}T+\sqrt{Q_1-Q_0}T_I)^2 DT_I . \tag{B5}$$

Here we assume that the ratio $c\equiv m/(1-Q_1)$ has a finite limit. This means m goes to 0 like $1-Q_1$. With this assumption, H_A tends to some constant, $H_B \propto 1/\sqrt{1-Q_1}$, and $H_C \propto 1/(1-Q_1)$. The singular parts of E, F_1, and F_0 are therefore

$$\begin{aligned} E&=\alpha\left[e_1\frac{1}{(1-Q_1)^2}\frac{1}{\sqrt{\Delta q_d}}+e_2\frac{1}{1-Q_1}\right] , \\ F_1&=\alpha f_1\frac{1}{(1-Q_1)^2}\frac{1}{\sqrt{\Delta q_d}} , \\ F_0&=\alpha f_0\frac{1}{(1-Q_1)^2} , \end{aligned} \tag{B6}$$

where e_1, e_2, f_1, and f_0 are constants which depend upon c, q_0, q_d. We should note that the singular part of $2E+F_1$ is proportional to $\alpha\sqrt{\Delta q_d}/(1-Q_1)^2 \sim \alpha/\sqrt{\Delta q_d}$, which is less singular than either E or F_1.

To discuss the behavior of Δq_d, it is convenient to study the last term of (18), which we denote $\ln z_{\rm RSB}$. Repeating the same analysis as was done for the RS case, we obtain

$$\ln z_{\rm RSB}=\frac{1}{m}\int Du\ln(a+b_1\exp m\sqrt{F_0}u+b_2\exp -m\sqrt{F_0}u)+\ln L , \tag{B7}$$

where, setting $A=2E+F_1$ and $\Delta F_1=F_1-F_0$,

$$\begin{aligned} a&=\int_{R_-}^{R_+}Du_I\exp\frac{1}{2}\frac{m}{A}(\sqrt{F_0}u+\sqrt{\Delta F_1}u_I)^2 , \\ b_1&=\int_{R_+}^{\infty}Du_I\exp-\tfrac{1}{2}mA+m\sqrt{\Delta F_1}u_I , \\ b_2&=\int_{-\infty}^{R_-}Du_I\exp-\tfrac{1}{2}mA-m\sqrt{\Delta F_1}u_I , \end{aligned} \tag{B8}$$

where $R_\pm=(\pm A-\sqrt{F_0}u)/\sqrt{\Delta F_1}$. In the $\Delta q_d \to 0$ limit of $\ln z_{\rm RSB}$, we only need to estimate the largest term in

the logarithm. Using (B6) and $m \propto \sqrt{\Delta q_d}$, we estimate

$$m\sqrt{F_0}\sim\sqrt{\alpha},\quad \frac{m}{A}F_1\sim\frac{1}{\sqrt{\Delta q_d}},\quad mA\sim\alpha\ ,$$
$$m\sqrt{F_1}\sim\sqrt{\alpha}/\Delta q_d^{1/4},\quad \frac{A}{\sqrt{F_1}}\sim\sqrt{\alpha}\Delta q_d^{1/4}, \tag{B9}$$
$$\frac{\sqrt{F_0}}{\sqrt{F_1}}\sim\Delta q_d^{1/4}\ .$$

To identify the largest terms, we should make some assumption on α. A natural assumption is $\alpha\propto 1/\sqrt{\Delta q_d}$, which is suggested by the RS theory. Here we only show the consistency of this solution. With this assumption, $R_\pm$ reduces to $\pm A/\sqrt{F_1}$, which is some constant, and a, b_1, and $b_2=b_1$ are estimated to be the largest values of the integrants. Thus the largest term in $\ln z_{\rm RSB}$ is proportional to $\alpha/m+\text{const}\sqrt{\alpha}/(\Delta q_d^{1/4}m)$, or $1/(\sqrt{\Delta q_d}m)$. Using the relation $q_dE+\frac{1}{2}mq_0F_0+\frac{1}{2}(1-m)q_1F_1=0$ and setting $m=m_0\sqrt{\Delta q_d}$, we obtain

$$\ln V/N=-\alpha\frac{C_{\rm RSB}}{m_0\sqrt{\Delta q_d}}+\left[\frac{\alpha}{\sqrt{\Delta q_d}}+\text{const}\frac{\sqrt{\alpha}}{\Delta q_d^{3/4}}\right],$$
$$\text{or}\ \frac{1}{\Delta q_d}+\ln L\ , \tag{B10}$$

where $C_{\rm RSB}=\int DT\ln(1/H_A)$. In the second line, we omitted the nonsingular coefficients. Either case gives the relation $\alpha\propto 1/\sqrt{\Delta q_d}$. Thus the dominant α dependence in $\ln V/N$ is α^2, which leads to $\alpha\propto\sqrt{\ln L}$ for $S_{\rm RSB}=0$. Finally, using the relations (B9), we can check that the relations $m\propto\sqrt{\Delta q_d}$ and $\alpha\propto 1/\sqrt{\Delta q_d}$ are consistent with the relation (B3), which, with (B2), will give the constants c, q_d, and q_0 for the $\Delta q_d\to 0$ limit.

APPENDIX C

In this appendix, we present the formula for γ_J and γ_P used to evaluate $\alpha_{\rm AT}$. The details of the derivation are the same as were done in [7] (see also [15]). In the following equations, order parameters are implicitly assumed to be the RS solutions for a given α. For both $K=3$ and ∞, γ_J in (24) is given by

$$\gamma_J=\int Du(\langle J^2\rangle-\langle J\rangle^2)^2\ , \tag{C1}$$

where $\langle\ \rangle$ means the average (21). γ_P is formally given by

$$\gamma_P=\gamma_0+(K-1)\gamma_1\ , \tag{C2}$$

where

$$\gamma_0=\int\prod_{l=1}^{K}Dt_l((\overline{y_1^2})^2-2(\overline{y}_1)^2\overline{y_1^2}+(\overline{y}_1)^4)\ ,$$
$$\gamma_1=\int\prod_{l=1}^{K}Dt_l((\overline{y_1y_2})^2-2\overline{y}_1\overline{y}_2\overline{y_1y_2}+(\overline{y}_1)^2(\overline{y}_2)^2)$$

$\overline{y_l^py_m^q}$ is given by $\int(y_l)^p(y_m)^qDW/\int DW$, where the weight DW, which comes from each replica, is given by

$$DW=\prod_{l=1}^{K}\frac{dy_l dh_l}{2\pi}\exp\left[\sum_l-\tfrac{1}{2}(q_d-q)y_l^2+i\sqrt{q}\,t_ly_l-iy_lh_l\right]\times\theta\left[\sum_l\operatorname{sgn}(h_l)\right]\ . \tag{C3}$$

Using these formulas for $K=\infty$, we obtain the leading contribution to $K\gamma_P$ as

$$K\gamma_P=\frac{1}{1-Q}\frac{1}{q_d^2-q^2}\frac{2}{\pi}\left\{\frac{q}{\sqrt{q_d^2-q^2}}\langle W^2\rangle_T+\frac{2}{\pi}\frac{1}{1-Q}\left[\frac{Q}{1-Q}\langle T^2W^2\rangle_T+2\left[\frac{Q}{1-Q}\right]^{1/2}\langle TW^3\rangle_T+\langle W^4\rangle_T\right]\right\}, \tag{C4}$$

where $W=H'(X)/H(X)$ with $H'(X)=-\exp(-\frac{1}{2}X^2)/\sqrt{2\pi}$ and $\langle\cdots\rangle_T$ means $\int\cdots DT$. For $K=3$, $K\gamma_P$ is given by

$$K\gamma_P=\frac{3}{(q_d-q)^2}\left[\frac{q}{q_d+q}\langle(1-\Sigma_1)^2W_1^2\rangle_t+\frac{(q_d-q)(q_d-2q)}{(q_d+q)(q_d+2q)}\langle(1-\Sigma_1)^4W_1^4\rangle_t+2\langle\Sigma_1^2\Sigma_2^2W_1^2W_2^2\rangle_t\right], \tag{C5}$$

where $\langle\cdots\rangle_t$ means $\int\cdots\prod_{l=1}^{3}Dt_l$, and $W_l=H'(X_l)/H(X_l)$, $\Sigma_1=H_2H_3/\Sigma$, and $\Sigma_2=H_1H_3/\Sigma$, where $\Sigma=H_1H_2+H_1H_3+H_2H_3-2H_1H_2H_3$.

[1] M. Minsky and S. Papert, *Perceptrons,* expanded ed. (MIT Press, Cambridge, MA, 1990).
[2] T. Cover, IEEE Trans. Electron. Comput. **14**, 326 (1965).
[3] G. J. Mitchison and R. M. Durbin, Biol. Cyb. **60**, 345 (1989).
[4] E. Gardner, J. Phys. A **21**, 257 (1988).
[5] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[6] E. Barkai, D. Hansel, and I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990).
[7] E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992).

[8] A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayer, and A. Zippelius, Phys. Rev. A **45**, 7590 (1992).
[9] S. Judd, in *Proceedings of the IEEE First Conference on Neural Networks,* edited by M. Caudill and C. Butler (SOS Printing, San Diego, 1987), Vol. II, p. 685.
[10] W. Krauth and M. Mezard, J. Phys. (Paris) **50**, 3057 (1989).
[11] H. Gutfreund, and Y. Stein, J. Phys. A **23**, 2613 (1990).
[12] J. R. L. de Almeida and D. J. Thouless, J. Phys. A **11**, 983 (1978).
[13] G. Parisi, J. Phys. A **13**, 1101 (1980).
[14] K. Nokura, J. Phys. A **22**, 93 (1989); T. Uezu and K. Nokura, Phys. Rev. B **46**, 898 (1992).
[15] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).