Generalization transitions in hidden-layer neural networks for third-order feature discrimination

August Romeo*
Dipartimento di Fisica, II Università degli Studi di Roma, "Tor Vergata," Via E. Carnevale, 00173 Rome, Italy
(Received 4 June 1992; revised manuscript received 21 October 1992)

Stochastic learning processes for a specific feature detector are studied. This technique is applied to nonsmooth multilayer neural networks requested to perform a discrimination task of order 3 based on the T-block–C-block problem. Our system proves to be capable of achieving perfect generalization, after presenting finite numbers of examples, by undergoing a phase transition. The corresponding annealed theory, which involves the Ising model under external field, shows good agreement with Monte Carlo simulations.
PACS number(s): 87.10.+e, 05.90.+m, 05.20.−y

I. INTRODUCTION

A modern issue in neural network physics is understanding in what ways the training of a system can be made more efficient. The key to this matter is the generalization curve of the model, which gives us an idea of the rate of improvement in the response to new data as a function of the number of examples presented (see, e.g., [1, 2]). In this work we will apply this sort of analysis to a two-dimensional image-processing task, realized by grids of mutually overlapping feature detectors in a hidden-layer feed-forward network. The use of supervised-learning methods in problems of this sort [3, 4] has been one of the contributing elements for renewing the interest in vision-related functions by neural nets.

The particular design introduced in this paper has been motivated by the so-called T-C problem [5]. Finding solutions to this problem means creating network schemes suitable for telling the block-T from the block-C patterns in Fig. 1(a). The distinction must be invariant under translations and $\pi/2$-generated rotations. This task has been described as worthy of study for its nonobvious difficulty. All the patterns consist of five "+1" pixels, and within each of the four T-C pairs, the shapes differ from each other by just one square. In spite of what rotational invariance might suggest, considering simple distances is not enough, because both letters have the same "order-2 spectrum," i.e., the two sets of distances between all the pairs of points in each shape coincide. Further, there is the same number of occurrences for each distance value. Therefore, as any separation method must involve properties depending on triplets of squares, the task has been called "problem of order three" in [5].

Some solutions based on the "receptive field" principle have been obtained, by backpropagation, in a special hidden-layer network described in [6]. There, the hidden units form an array of replicated local detectors for mutually overlapping regions, and the learning process is restricted by constraints in order to preserve the repeated structure. In the present work we employ a similar hidden layer, but with a new type of feature detector. The local regions scanned will consist of just three squares—instead of nine—and the weights will be discrete, not continuous. Further, they will no longer be subject to replica constraints.

Our scheme, which we call the "H" solution, differs from those in a previous study [7] in two main aspects: each detector is sensitive to *several* local patterns and the model is more amenable to theoretical analysis. The first point suits the purpose of maintaining the required rotational invariance without having to use several detector "families" specialized in one feature. This is of interest when it comes to simplifying the architecture, as we will manage to perform the task employing only one family of replicated detectors. By lack of an exact quenched theory, an annealed approximation (AA) is made. Actually, a replica-symmetric or similar approach would also be possible but, for our aims, the annealed method suffices. At first sight, one may be surprised to see that it works quite well even at low temperatures. This is not so amazing in the light of a recent analysis [8] determining the types of problems for which the qualitative predictions of

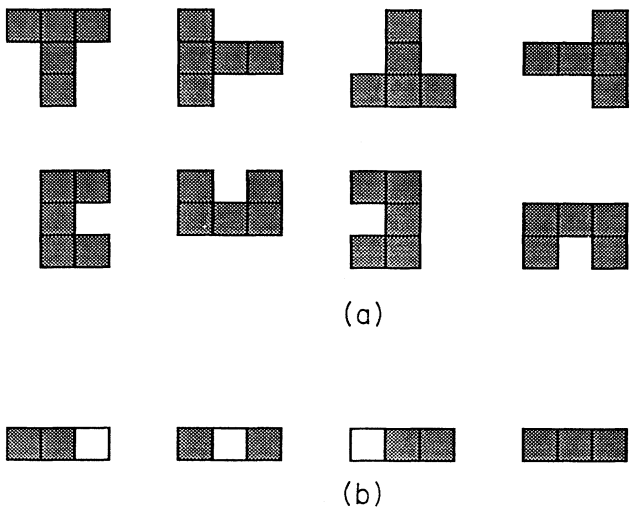

FIG. 1. (a) Shapes defining the T-C problem. (b) The four local patterns employed to distinguish both letters. Their sum of occurrences is 3 for a T block and 4 for a C block.

this technique continue to be valid at low temperature. In our case, it is the *realizability* of the rule studied that makes such an approach acceptable.

After evolution under the training dynamics, the system has to perform the right type of feature-counting, i.e., the task which solves the problem not just on the sets presented, but on the whole input space. Thus, we seek states of perfect generalization. In these situations the adequate training process is, rather than backpropagation, some sort of stochastic learning [1], which entails the controlled introduction of random errors during the learning phase.

Section II is a brief review of the formalism. The first model we introduce is described in Sec. III, together with the theory predicting its behavior and simulation results. A variant of that scheme is described in Sec. IV. Section V contains the final comments. A calculation of quite general applicability appears in the Appendix.

II. GENERALIZATION FUNCTIONS

Training a neural network means supplying it with examples, made of sets of p input patterns, $\{S^\mu, \mu = 1, \ldots, p\}$, where $S^\mu = (S_1^\mu, \ldots, S_M^\mu)$ for every μ, and their associated outputs $\mathcal{N}[S^\mu]$. Given a fixed architecture, each possible "network" is characterized by the weight matrix ω. The objective of any learning procedure is to find a mapping $\mathcal{N}_\omega[S]$ as similar as possible to the target rule $\mathcal{N}[S]$. This is often achieved by minimizing the training error or energy

$$E_t(\omega; \{S\}) = \sum_{\mu=1}^{p} \epsilon(\omega; S^\mu), \tag{1}$$

where $\epsilon(\omega; S) \propto (\mathcal{N}_\omega[S] - \mathcal{N}[S])^2$ measures the deviation of the existing $\mathcal{N}_\omega[S]$ from the ideal $\mathcal{N}[S]$. Nevertheless, as explained in recent works (see, e.g., [1], [2], [8], or [9]), the best way of assessing the efficiency of a network is to measure its performance on *any* possible input data, not necessarily in the training set, by means of the generalization error

$$\epsilon_g(\omega) = \langle\!\langle \epsilon(\omega; S) \rangle\!\rangle_S, \tag{2}$$

where $\langle\!\langle \cdots \rangle\!\rangle_S = \int d\mu(\{S\}) \cdots = \int \prod_{\mu=1}^{p} d\mu(S^\mu) \cdots$ denotes a *quenched* average over the whole distribution of example sets $\{S\}$, $d\mu(S)$ being a normalized measure. If, by randomness, the distributions of sets and of individual patterns coincide, this operation is done by just averaging over all the single S's. When the system is asked to deduce the *whole* target rule from the examples shown—as will be the case—the decrease of ϵ_g must be guaranteed, even at the expense of tolerating errors for some E_t's. That is the reason why stochastic learning is particularly appropriate in these situations. For processes of this nature, the resulting ω's have long-time distributions of the Gibbs type, with probabilities

$$P_{\{S\}}(\omega) = \frac{1}{Z_{\{S\}}} e^{-\beta E_t(\omega; \{S\})}, \tag{3}$$

where

$$Z_{\{S\}} = \int d\mu(\omega) e^{-\beta E_t(\omega; S)} \tag{4}$$

is the partition function, and $d\mu(\omega) = P_0(\omega) d\omega$ denotes the *a priori* normalized measure in weight space. $T = 1/\beta$ is the training temperature determining the allowed level of stochastic noise, in the sense that $T = 0$ forces the system to minimize E_t, while $T \to \infty$ permits any set of weights with equal probability. Except for $P_0(\omega)$, the above quantities depend on $\{S\}$, as they are referred to E_t, which varies for every chosen set.

The average training error over sets of p patterns is defined as

$$\epsilon_t(\beta, p) = \frac{1}{p} \langle\!\langle \langle E_t(\omega; \{S\}) \rangle_T \rangle\!\rangle_S, \tag{5}$$

which involves two averages: thermal $\langle \rangle_T$, i.e., using $P_{\{S\}}(\omega)$, and quenched, here limited to $\{S\}$'s of p elements. Interchanging the averages and using (3) we come to

$$\begin{aligned} \epsilon_t(\beta, p) &= \frac{1}{p} \left\langle\!\!\left\langle \frac{1}{Z_{\{S\}}} \int d\mu(\omega) e^{-\beta E_t(\omega; \{S\})} E_t(\omega; \{S\}) \right\rangle\!\!\right\rangle_S \\ &= -\frac{1}{p} \frac{\partial}{\partial \beta} \langle\!\langle \ln Z_{\{S\}} \rangle\!\rangle_S. \end{aligned} \tag{6}$$

Introducing the average free energy per weight $f(\beta, p)$

$$-N_\omega \beta f(\beta, p) = \langle\!\langle \ln Z_{\{S\}} \rangle\!\rangle_S, \tag{7}$$

with N_ω being the number of evolving weights, the above relation is expressed as

$$\epsilon_t(\beta, p) = -\frac{1}{\alpha} \frac{\partial}{\partial \beta} (\beta f(\beta, p)), \tag{8}$$

where $\alpha = p/N_\omega$ is the relative size of the training sets.

Analogously, the average generalization error for p-pattern sets is

$$\epsilon_g(\beta, p) = \langle\!\langle \langle \epsilon(\omega; S) \rangle_T \rangle\!\rangle_S. \tag{9}$$

The deviations of the typical values of E_t and $\epsilon(\omega; S)$ from their thermal and quenched averages are supposed to vanish as $N_\omega \to \infty$. $\epsilon_t(\beta, p)$ and $\epsilon_g(\beta, p)$, viewed as functions of p, are the learning and generalization curves, and encode the system's ability to learn and infer rules.

III. LINEAR H SOLUTION

Our system must work in such a way that it produces different outputs for the T and C shapes in Fig. 1(a). The input layer is a screen of $M = N \times N$ elements with ± 1 values, where binary images are formed. Since the order-2 spectra of both letters coincide, any local detector aimed at their discrimination must scan regions of three sites at least. On a square grid, no single three-pixel set can be invariant under $\pi/2$-generated rotations. Therefore, translationally repeated copies of a detector for one specific feature cannot be enough. However, we have found a solution of this nature with a replicated detector activated by *several* different bidimensional local patterns. These are four different features: rows of three

+1's, or with two +1's and one −1 in any of the three sites [Fig. 1(b)]. It is not difficult to verify—and this is our goal—that the total sum of occurrences of these combinations is three for any T, and four for any C, *regardless of orientation*, i.e., although every individual feature is not rotationally invariant, the sum of the four actually is. An analogous solution with columns instead of rows would equally work.

The hidden layer is a two-dimensional array of binary (0,1) units, which shall be called $H_{i,j}$. Every one of these neurons controls the region formed by the corresponding $S_{i,j}$ and its two horizontal nearest neighbors on the input layer. All the hidden units feed into a single integer-valued output neuron $\mathcal{N}$, that adds up the activation states of all of them (Fig. 2). This is why we call this model "linear." $H_{i,j}$ will be active—and thus have unit value—if its input region contains either three or two +1 pixels. By construction, the working of this system is translationally invariant.

The required detection is realized by the weights, threshold, and logistic function contained in the expression

$$H_{i,j}(S) = \Theta(S_{i,j-1} + S_{i,j} + S_{i,j+1} + \theta), \tag{10}$$

where $\Theta(x) = 0$ for $x \leq 0$ and 1 otherwise, and θ can be any number satisfying $0 < \theta \leq 1$. Bearing in mind that the S_{ij}'s are sign-valued variables, these activation states can be conveniently written as

$$H_{i,j}(S) = \tfrac{1}{2} + \tfrac{1}{4}\,(S_{i,j-1} + S_{i,j} + S_{i,j+1} - S_{i,j-1}S_{i,j}S_{i,j+1}). \tag{11}$$

The output unit will take on the value of their sum, i.e.,

$$\mathcal{N}[S] = \sum_{i,j} H_{i,j}(S), \tag{12}$$

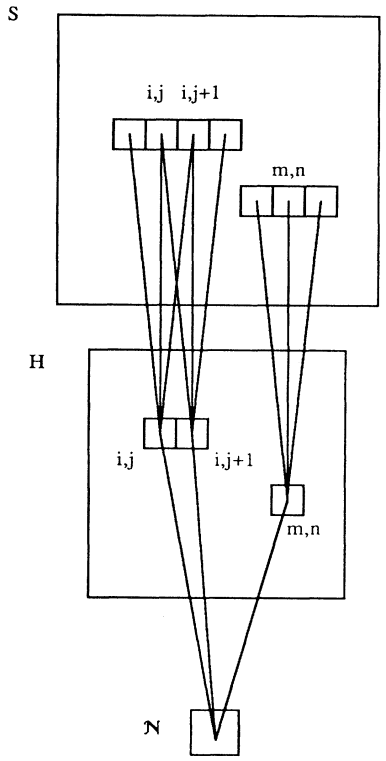

FIG. 2. Structure of the H network. A two-dimensional hidden layer scans all the possible three-square rows.

which is the total number of successful feature detections.

This task will be learned in networks with the same connection architecture, keeping all the thresholds and the weights from hidden to output layer fixed, but allowing the weights from input to hidden units to take on sign values, given by ±1 variables $A_{i,j}$, $B_{i,j}$, and $C_{i,j}$ entering in the way

$$\begin{aligned} H_{ABC\ i,j}(S) &= \Theta(A_{i,j}S_{i,j-1} + B_{i,j}S_{i,j} + C_{i,j}S_{i,j+1} + \theta) \\ &= \tfrac{1}{2} + \tfrac{1}{4}\,[A_{i,j}S_{i,j-1} + B_{i,j}S_{i,j} + C_{i,j}S_{i,j+1} - A_{i,j}B_{i,j}C_{i,j}S_{i,j-1}S_{i,j}S_{i,j+1}]\,. \end{aligned} \tag{13}$$

Then, the function evaluated by one of these $\{A, B, C\}$ networks is $\mathcal{N}_{ABC}[S] = \sum_{i,j} H_{ABC\ i,j}[S]$, and its error—referred to the target rule—when presenting a certain input pattern S reads

$$\begin{aligned} \epsilon(A,B,C;S) &= \frac{1}{N^2}(\mathcal{N}_{ABC}[S] - \mathcal{N}[S])^2 \\ &= \frac{1}{16N^2}\left[\sum_{i,j}[(A_{i,j}-1)S_{i,j-1} + (B_{i,j}-1)S_{i,j} + (C_{i,j}-1)S_{i,j+1} - (A_{i,j}B_{i,j}C_{i,j}-1)S_{i,j-1}S_{i,j}S_{i,j+1}]\right]^2 \end{aligned} \tag{14}$$

which, by summing over all the S^μ's of a given example set, yields the training energy $E_t(A,B,C;\{S\}) = \sum_{\mu=1}^{p} \epsilon(A,B,C;S^\mu)$. The generalization ability of such a model can be tested by starting from a network with random $A_{i,j}$'s, $B_{i,j}$'s, and $C_{i,j}$'s, and letting the system evolve by stochastic learning governed by the E_t's associated to training sets randomly drawn. The point is how hard it is now to reach the configuration $A_{i,j} = B_{i,j} = C_{i,j} = 1 \forall i,j$, which constitutes the exact solution of the problem on the whole input space. This difficulty is measured by the generalization error $\epsilon_g(A,B,C) = \langle\langle \epsilon(A,B,C;S)\rangle\rangle_S$. Since the S's are totally random and sign valued, formally we have $d\mu(S) = \prod_{i,j}[\frac{1}{2}\delta(S_{i,j}-1) + \frac{1}{2}\delta(S_{i,j}+1)]dS_{i,j}$, as a result

of which $\langle\langle\cdots\rangle\rangle_S = \frac{1}{2^{N^2}}\sum_{\{S_{i,j}=\pm1\}}\cdots$. Averaging over all the S patterns, we obtain

$$\epsilon_g(A,B,C) = \tfrac{1}{16}\,[14 - 6(\langle A\rangle_0 + \langle B\rangle_0 + \langle C\rangle_0) + 2(\langle AB\rangle_{0,-1} + \langle BC\rangle_{0,-1} + \langle AC\rangle_{0,-2}) - 2\langle ABC\rangle_{0,0,0}], \quad (15)$$

with the notations

$$\langle X\rangle_0 = \frac{1}{N^2}\sum_{i,j} X_{i,j},$$
$$\langle XY\rangle_{0,n} = \frac{1}{N^2}\sum_{i,j} X_{i,j}Y_{i,j+n}, \quad (16)$$
$$\langle XYZ\rangle_{0,0,0} = \frac{1}{N^2}\sum_{i,j} X_{i,j}Y_{i,j}Z_{i,j},$$

where X, Y, Z stand for any of A, B, C. In order to simplify the situation, we further restict the problem by looking at particular cases with just one evolving matrix

(1) $B_{i,j} = C_{i,j} = 1\forall i,j$. Only the A weights evolve

$$\epsilon_g(A,\mathbf{1},\mathbf{1}) = \tfrac{1}{4}(1 - \langle A\rangle_0), \quad (17)$$

where the matrix denoted as $\mathbf{1}$ has all its coefficients equal to 1. The dynamics is purely based on an "external field" equally influencing all the A weights, which would not interact among them.

(2) $B_{i,j} = 1\forall i,j$, $C_{i,j} = A_{i,j}$ evolving

$$\epsilon_g(A,\mathbf{1},A) = \tfrac{1}{8}(3 - 4\langle A\rangle_0 + \langle AA\rangle_{0,-2}) \equiv \epsilon_g(A). \quad (18)$$

This generalization error can be interpreted as a Hamiltonian for an Ising-fashion evolution in the A lattice with (i) an external field term which prevents symmetry under global sign flip, thus reminding us that $A_{i,j} = -1\forall i,j$ cannot be a solution; (ii) weight interaction between next-to-nearest (but not nearest) horizontal neighbors.

(3) $C_{i,j} = B_{i,j} = A_{i,j}$ evolving

$$\epsilon_g(A,A,A) = \tfrac{1}{8}(7 - 10\langle A\rangle_0 + 2\langle AA\rangle_{0,-1} + \langle AA\rangle_{0,-2}), \quad (19)$$

with the same characteristics as the previous case plus the usual nearest-neighbor interaction term $\sim \langle AA\rangle_{0,-1}$.

Taking into account both simplicity and physical interest, we have chosen to study model 2. Its resulting long-time behavior will not significantly differ from that of a system under the effective dynamics (18). Perfect generalization can be achieved by arriving, after evolution, at the state $A_{i,j} = 1\forall i,j$. Since the system is discrete, a discontinuous transition should not be ruled out. However, given that the ground state is unique, the typical competition effects will not be present. Therefore, what we certainly cannot expect is a phase transition involving sign symmetry breaking.

A. Annealed theory

The basic assumption of the annealed approximation (AA) method is the validity of replacing $\langle\langle \ln Z_{\{S\}}\rangle\rangle_S$ with $\ln\langle\langle Z_{\{S\}}\rangle\rangle_S$ in expression (7). In [8], the authors have shown the limitations and advantages of this technique. Although in general important deviations from the AA results are to be expected at low temperatures, there are at least two cases in which the qualitative behavior predicted is correct down to fairly low T's: *realizable* rules or networks with *Boolean* output. This property was found for perceptrons, but we feel it can be extended to our model, as the roles played by the weight vectors employed in that work and by our A matrix are completely analogous. The reason why the AA is applicable here is that, by construction, our rule is realizable, i.e., there exists an A^* such that $\epsilon(A^*;S) = 0\forall S$, namely $A^*_{i,j} = 1\forall i,j$.

We have supposed that the *a priori* random distribution of the sign-valued weights for the trained system is uniform, i.e.,

$$d\mu(A) = \prod_{i,j}[\tfrac{1}{2}\delta(A_{i,j} - 1) + \tfrac{1}{2}\delta(A_{i,j} + 1)]dA_{i,j},$$

which turns the partition function (4) into a discrete sum

$$Z_{\{S\}} = \int d\mu(A)e^{-\beta E_t(A;\{S\})} = \frac{1}{2^{N^2}}\sum_{\{A_{i,j}=\pm1\}} e^{-\beta E_t(A;\{S\})}. \quad (20)$$

Then, its quenched average is

$$\langle\langle Z_{\{S\}}\rangle\rangle_S = \frac{1}{2^{N^2}}\sum_{\{A_{i,j}=\pm1\}}\langle\langle e^{-\beta E_t(A;\{S\})}\rangle\rangle_S.$$

Taking advantage of the usual relation $\langle\langle e^{-\beta E_t(A;\{S\})}\rangle\rangle_S = \langle\langle e^{-\beta\epsilon(A;\{S\})}\rangle\rangle_S^p$, one arrives at

$$\langle\langle Z_{\{S\}}\rangle\rangle_S = \frac{1}{2^{N^2}}\sum_{\{A_{i,j}\}} e^{-p\,G_{\rm an}(A)}, \quad (21)$$

with

$$e^{-G_{\rm an}(A)} = \langle\langle e^{-\beta\epsilon(A;S)}\rangle\rangle_S. \quad (22)$$

$G_{\rm an}$ is sometimes called the annealed effective Hamiltonian. Here the error $\epsilon(A;S)$ is a function of the type (14), with $B_{i,j} = 1, C_{i,j} = A_{i,j}$, which reads

$$\epsilon(A;S) \equiv \epsilon(A,\mathbf{1},A;S) = \frac{1}{16N^2}\left[\sum_{i,j}(A_{i,j} - 1)(S_{i,j-1} + S_{i,j+1})\right]^2. \quad (23)$$

(Notice that inside the square brackets there are terms just linear in the S's.) In the Appendix we have found, as a general result for all the possible $\epsilon(A;S)$'s of a larger class including this case, and in the large-N limit

$$G_{\rm an}(A) = \tfrac{1}{2}\ln[1 + 2\beta\epsilon_g(A)]. \tag{24}$$

Both the type of function and the dependence on A through ϵ_g are corroborated by the particular calculations in [3], [8], and [4]. In our model, we have the ϵ_g given by (18), that we conveniently rewrite into the form $\epsilon_g(A) = \frac{1}{2}[\frac{3}{4} - m(A)] \equiv \epsilon_g[m(A)]$, where

$$m(A) = \frac{1}{N^2}\left[\sum_{i,j} A_{i,j} - \frac{1}{4}\sum_{i,j} A_{i,j}A_{i,j+2}\right] \tag{25}$$

can be regarded as a macroscopic property to which we shall associate an order parameter. Thus

$$\begin{aligned}\langle\langle Z_{\{S\}}\rangle\rangle_S &= \frac{1}{2^{N^2}}\sum_{\{A_{i,j}\}}\int_{-\infty}^{\infty} dm\; e^{-(p/2)\ln[1+\beta(3/4-m)]}\delta(m - m(A)) \\ &= \frac{1}{2^{N^2}}\int_{-\infty}^{\infty} dm \int_{-i\infty}^{i\infty}\frac{N^2 dk}{2\pi i}\; e^{-(p/2)\ln[1+\beta(3/4-m)]-N^2 km}\sum_{\{A_{i,j}\}} e^{N^2 km(A)},\end{aligned} \tag{26}$$

where a Fourier representation of the δ distribution has been introduced, and a simple variable change has taken place. Now, the sum over $\{A_{i,j}\}$ in the integrand is interpreted as the partition function for a discrete system with Hamiltonian $N^2 m(A)$ at "temperature" $1/k$, which is equivalent to a set of effectively one-dimensional Ising models subject to an external field and with the nearest-neighbor interaction replaced by a next-to-nearest neighbor one. Taking some care with the new index grouping, we apply the standard transfer-matrix method and, in the large-N limit, get

$$\begin{aligned}&\sum_{\{A_{i,j}\}} e^{N^2 km(A)} \\ &\simeq \left[e^{-k/4}\cosh k + \sqrt{e^{-k/2}\sinh^2 k + e^{k/2}}\right]^{N^2}.\end{aligned} \tag{27}$$

This sort of calculation using the knowledge of the weight partition function is similar to the method applied in [3], although no external-field term was present there. The same technique was employed in [4] for a two-dimensional Ising model with graphic methods. The annealed free energy per weight f must satisfy $\langle\langle Z_{\{S\}}\rangle\rangle_S \sim \int dm \int dk\; e^{-N^2\beta f(m,k)}$. We thus identify

$$\begin{aligned}\beta f(m,k) = &\frac{\alpha}{2}\ln[1 + \beta(\tfrac{3}{4} - m)] + km + \frac{k}{4} \\ &- \ln\left[\cosh k + \sqrt{\sinh^2 k + e^k}\right],\end{aligned} \tag{28}$$

where we have recalled $\alpha = p/N^2$, as $N_\omega = N^2$. In the thermodynamic limit ($N \to \infty$) the integral will be dominated by its saddle point, i.e., at any given α and β, the relevant values are those minimizing $f(m,k)$, which leads us to $\partial(\beta f)/\partial m = 0$, $\partial(\beta f)/\partial k = 0$. These two equations can be written

$$k = \frac{\alpha\beta/2}{1 + \beta(3/4 - m)} \equiv k(m), \tag{29}$$

$$\begin{aligned}m = &-\frac{1}{4} + \frac{\sinh k}{\sqrt{\sinh^2 k + e^k}} \\ &+ \frac{1}{2}\frac{e^k}{\cosh k\sqrt{\sinh^2 k + e^k} + \sinh^2 k + e^k}.\end{aligned} \tag{30}$$

As $\alpha\beta \to \infty$, $k \to \infty$ and thus $m \to 3/4$. Rising α means increasing the number of examples, and growth in β amounts to reducing the noise level. This value of m is expectable, as the limit in question should take us, sooner or later, to the "perfect generalization" solution $A_{i,j} = 1 \forall i,j$, which yields $m(A) = 3/4$. On the other hand, $\alpha\beta \to 0$ gives $k \to 0$ and therefore $m \to 0$. This is logical, too, as a state of maximal disorder—random $A_{i,j}$'s—gives $m(A) = 0$.

The study of the annealed free energy becomes much easier by solving one of the two saddle-point equations. Taking $k(m)$ as given by (29), we get

$$\beta f(m) \equiv \beta f(m, k(m)), \tag{31}$$

and then,

$$\begin{aligned}\langle\langle Z_{\{S\}}\rangle\rangle_S\big|_{\rm SP} &\sim \int dm\; e^{-N^2\beta f(m)}\Big|_{\rm SP} \\ &\sim e^{-N^2\beta f(m_{\rm SP})},\end{aligned}$$

where SP means that the saddle-point value is taken. Thus the properties of the system can be described by one single order parameter m. Since in the large-N limit stochastic fluctuations can be neglected, in most cases the system must converge to the m corresponding to the global minimum of $f(m)$. The existence of local minima, associated with metastable states, depends in general on the details of the problem and on the values of α and β. Here we see that $\beta f(m) \to 0$ as $\alpha\beta \gg 1$ and $m \to 3/4$. However, since $m = 3/4$ is the boundary of the domain for $f(m)$, some care is called for. Figure 3 shows the shape of $\beta f(m)$ for $\beta = 100$ and different values of α. As α goes away from zero, it is clearer that $\beta f(m) \to 0$ decreasing when $m \to 3/4$ on the left. In this sense, $\beta f(3/4) = 0$ can be considered as a "minimum." Below $\alpha_2 = 0.49$, another minimum—a true one—for a certain

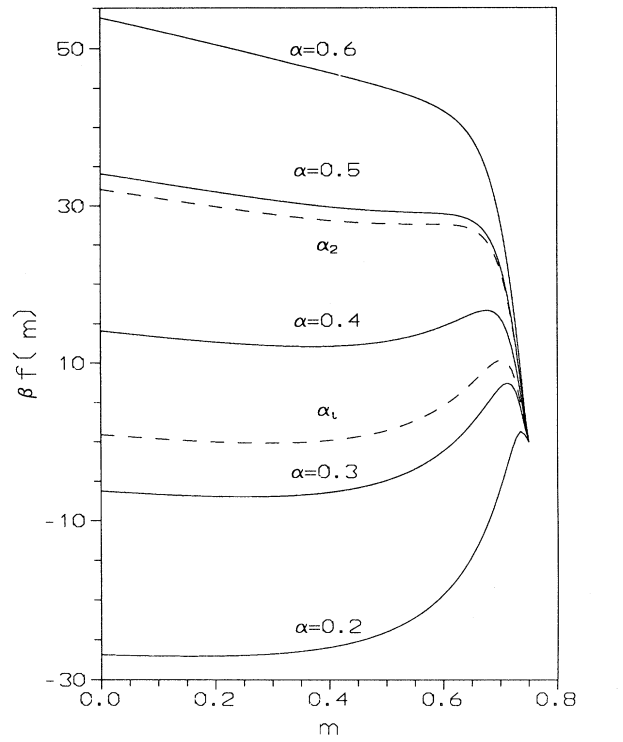

FIG. 3. Representation of $\beta f(m)$ (f being the free energy per weight) for the linear 'H' model as a function of the order parameter m for $\beta = 100$. The line $\alpha = \alpha_t$ marks the thermodynamic transition. In all cases $\beta f(3/4) = 0$, while the minimum at $m = m_0 < 3/4$ gives $\beta f(m_0) < 0$ for $\alpha < \alpha_t$ and $\beta f(m_0) > 0$ for $\alpha_t < \alpha < \alpha_2$. $\alpha = \alpha_2$ is the spinodal, above which the minimum at $m = m_0$ no longer exists.

$m = m_0 < 3/4$ exists. Within this range, for $\alpha < \alpha_t = 0.335$ the global minimum is at $m = m_0$, the one at $m = 3/4$ being simply local, while above α_t the situation is reversed.

Starting with random initial weights—i.e., $m \simeq 0$—when $\alpha < \alpha_t$ the system will evolve to settle down at m_0. If we rise α by supplying sets of more examples, then, on arriving at α_t, the value at which the network stabilizes will ideally shift from $m = m_0$ to $m = 3/4$, achieving the perfect generalization state. Therefore, a thermodynamic transition occurs at $\alpha = \alpha_t$. The value α_2, bounding the region in which more than one minima exist, is a *spinodal*. Above α_2 there is no local minimum for $m < 3/4$, and the system will converge fast to the state $m = 3/4$. At $\alpha = \alpha_2$, it is said to undergo a spinodal transition. This is of the greatest practical importance, since for moderately large networks the observable transitions are the spinodals rather than the thermodynamic ones. If the size tends to infinity, the time necessary to escape from a local minimum becomes exponentially large. As a result, the system gets stuck in the nearest metastable state as long as such states exist, which, by the above observation, happens until $\alpha = \alpha_2$ is reached. It is then—and not earlier at $\alpha = \alpha_t$—that the discontinuous drop signaling the transition takes place.

Next we outline how the values of $\epsilon_g(\beta, p)$ are predicted in this formalism. Interchanging averages in (9) and using (2) we can write $\epsilon_g(\beta, \alpha) = \langle \epsilon_g(A) \rangle_T$. Now, (21) suggests replacing the initial Gibbs probabilities with the "annealed" distribution

$$P_{\mathrm{AA}}(A) = \frac{1}{\langle\langle Z_{\{S\}} \rangle\rangle_S} e^{-p\, G_{\mathrm{an}}(A)}.$$

Doing so, we obtain the standard AA formula

$$\epsilon_g(\beta, \alpha) = \frac{1}{\langle\langle Z_{\{S\}} \rangle\rangle_S} \int d\mu(A)\; e^{-pG_{\mathrm{an}}(A)} \epsilon_g(A). \quad (32)$$

Manipulating this integral like the previous one, we arrive at

$$\epsilon_g(\beta, \alpha)_{\mathrm{SP}} \sim \frac{1}{\langle\langle Z_{\{S\}} \rangle\rangle_S} \int dm\; e^{-N^2 \beta f(m)} \epsilon_g[m] \Big|_{\mathrm{SP}}. \quad (33)$$

It is now clear that

$$\epsilon_g(\beta, \alpha)_{\mathrm{SP}} = \epsilon_g[m_{\mathrm{SP}}] = \tfrac{1}{2}(\tfrac{3}{4} - m_{\mathrm{SP}}). \quad (34)$$

In order to evaluate $\epsilon_g(\beta, \alpha)$ as a function of α for fixed β, increasing values of α until $\alpha = \alpha_2$ are taken. Each time, we solve numerically the saddle-point equation for m finding the minimum—within the interval $(0, 3/4)$—of (31), which is m_{SP}, and (34) gives us the value of ϵ_g for the α in question. The set of points thus obtained is shown in Fig. 4, for a certain value of β, as the curve drawn in a solid line. For the already mentioned practical reason, the transition worthy of being predicted is the spinodal. Taking this into account, we have increased α until α_2 selecting always the minimum $m = m_0$, even for $\alpha_t < \alpha < \alpha_2$, when it is just local. As for the thermodynamic transition itself, the curve would be the same but with the vertical line on $\alpha = \alpha_t$ instead of on α_2.

Further study of the function $\beta f(m)$ suggests that, within this theory, the passage to the sate of perfect generalization ceases to be a sharp transition when $T = 1/\beta$ is above a certain critical temperature T_c.

Though not theoretically evaluated in this work, the average training error can also be calculated. Departing from the general expression (8), we apply AA by saddle point as before, and arrive at $\epsilon_t(\beta, \alpha)_{\mathrm{SP}} = -(1/\alpha)(\partial/\partial\beta)(\beta f(m_{\mathrm{SP}}))$, which would give us the desired prediction in terms of m_{SP}.

B. Simulation

A number of training sessions have been simulated by means of a Metropolis-Monte Carlo program [10]. Starting from random $A_{i,j}$'s, sets of a $p = 1$ pattern are supplied, each of them giving rise to a new training energy $\sum_{\mu=1}^{p} \epsilon(A; S^\mu)$, with the $\epsilon(A; S)$ in (23). The weight-flip probabilities will depend on the set in question through E_t, and, of course, on β. A thermal average of $\epsilon_g(A)$ is computed for each set, and, afterwards, an average over all the p-pattern sets generated is also taken.

The method is repeated for increasing values of p, thus yielding new $\epsilon_g(\beta, \alpha)$'s ($\alpha = p/N^2$). In order to reduce the edge and size effects, we have used weight arrays subject to periodic boundary conditions maintained through the whole process of weight updating. Typical results of

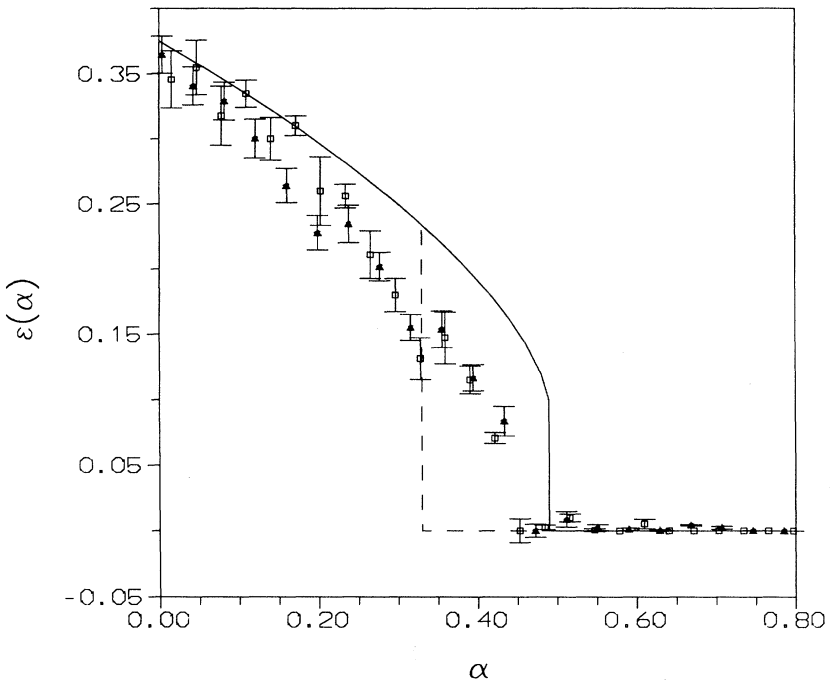

FIG. 4. Monte Carlo simulation depicting the generalization curve $\epsilon_g(\beta,\alpha)$ as a function of the relative size of the example set α corresponding to a fixed $\beta = 100$, for linear H networks of 64 and 256 evolving weights (squares and triangles, respectively). The solid line is the result of the annealed theory. The drop in the generalization error takes place between the thermodynamic transition value $\alpha = \alpha_t$, dashed line, and the spinodal $\alpha = \alpha_2$, vertical solid line, as expected.

processes at $\beta = 100$ are shown in Fig. 4. A transition is completed at the spinodal α_2, and there is good qualitative agreement with the AA prediction, even at such a low temperature. The points are closer to the theoretical curve for small α's, and, in spite of the slight deviations for intermediate values, the domain where the generalization error vanishes is accurately predicted.

IV. BOOLEAN H MODEL

Here we present a modified version of our scheme, where the linear output is replaced with a Boolean unit $\mathcal{B}$, setting

$$\mathcal{B}[S] = \Theta\left(\frac{1}{N}(\mathcal{N}[S] - \theta_L)\right), \qquad \mathcal{B}_A[S] = \Theta\left(\frac{1}{N}(\mathcal{N}_A[S] - \theta_L)\right), \tag{35}$$

for the solution and for the learning network, respectively. $\mathcal{N}[S]$ is given by (12) and (11) and $\mathcal{N}_A[S]$ by (12) and (13) for $C = A, B = \mathbf{1}$.

We can choose θ_L so that the values $\mathcal{N} = 3$, for T, and $\mathcal{N} = 4$, for C, be separated. This is achieved if $3 \le \theta_L < 4$, producing $\mathcal{B}$=0 (1) for T (C). The error function in this situation is appropriately written as

$$\epsilon(A;S) = \Theta\left(-\frac{1}{N}(\mathcal{N}_A[S] - \theta_L)\frac{1}{N}(\mathcal{N}[S] - \theta_L)\right). \tag{36}$$

Hence, the generalization error for a given A is now

$$\epsilon_g(A) = \langle\langle \epsilon(A;S)\rangle\rangle_S = \int dx \int dy\, \Theta(-xy) \int \frac{d\hat{x}}{2\pi} \int \frac{d\hat{y}}{2\pi}\, e^{i[\hat{x}x+\hat{y}y+(\hat{x}+\hat{y})(\theta_L/N-N/2)]} \langle\langle e^{-(i/4N)(\hat{x}\bar{\mathcal{N}}_A[S]+\hat{y}\bar{\mathcal{N}}[S])}\rangle\rangle_S. \tag{37}$$

δ distributions and their Fourier representations have been introduced. The new quantities $\frac{1}{4}\bar{\mathcal{N}}[S] = \mathcal{N}[S] - N^2/2$, $\frac{1}{4}\bar{\mathcal{N}}_A[S] = \mathcal{N}_A[S] - N^2/2$, contain only S-dependent terms. In regard to the last factor in the integrand, a calculation similar to those in Appendix A gives, as a result in the large-N limit,

$$\left\langle\left\langle \exp\left(-\frac{i}{4N}\sum_{i,j}[(\hat{x}+\hat{y})S_{i,j}(1 - S_{i,j-1}S_{i,j+1}) + (\hat{x}A_{i,j} + \hat{y})(S_{i,j-1} + S_{i,j+1})]\right)\right\rangle\right\rangle_S$$

$$\simeq \exp\{-\tfrac{1}{16}[m_1(A)\hat{x}^2 + 5\hat{y}^2 + m_2(A)\hat{x}\hat{y}]\}, \tag{38}$$

where

$$m_1(A) = \langle AA\rangle_{0,-2} + 2\langle A\rangle_0 + 2,$$
$$m_2(A) = 4 + 6\langle A\rangle_0, \tag{39}$$

in the notation (16).

Evaluating the integral (37), we arrive at

$$\epsilon_g(A) = \frac{1}{\pi}\arccos\left(\frac{m_2(A)}{2\sqrt{5m_1(A)}}\right) - \frac{32}{R(A)} I\left(\frac{\theta_L}{N} - \frac{N}{2}; \frac{10}{R(A)}, \frac{2m_1(A)}{R(A)}, \frac{2m_2(a)}{R(A)}\right)$$
$$\equiv \epsilon_g[m_1(A), m_2(A)], \tag{40}$$

where

$$R(A) = \sqrt{20m_1(A) - m_2^2(A)}, \tag{41}$$

and

$$I(t; a, b, c) \equiv \int_0^t dx \int_0^t dy\ e^{-ax^2 - by^2 + cxy}. \tag{42}$$

Even though this means changing the initial problem, the model becomes more amenable by choosing a threshold $\theta_L = N^2/2$, as the second term of $\epsilon_g(A)$ vanishes, leaving just

$$\epsilon_g(A) = \frac{1}{\pi}\arccos\left(\frac{2 + 3\langle A\rangle_0}{\sqrt{5(\langle AA\rangle_{0,-2} + 2\langle A\rangle_0 + 2)}}\right). \tag{43}$$

In view of this result, the same reasoning as in [3] for the "contiguity problem" is in order. The singular dependence of the derivative of $\epsilon_g(A)$ on $\langle A\rangle_0$, $\langle AA\rangle_{0,-2}$ at $\langle A\rangle_0 = \langle AA\rangle_{0,-2} = 1$ gives rise to a discontinuous transition from "high" ϵ_g to $\epsilon_g = 0$ *at any temperature.*

An AA approach would also be possible, but a bit more involved than in the linear model. We will just outline how to proceed. Since (36) is binary,

$$\langle\langle e^{-\beta\epsilon(A;S)}\rangle\rangle_S = \epsilon_g(A)e^{-\beta} + 1 - \epsilon_g(A). \tag{44}$$

Therefore the annealed effective Hamiltonian from (22) reads

$$G_{\rm an}(A) = -\ln[1 - (1 - e^{-\beta})\epsilon_g(A)]. \tag{45}$$

This time two order parameters are called for, as a result of which

$$\langle\langle Z_{\{S\}}\rangle\rangle_S \sim \int dm_1 \int dm_2 \int dk_1 \int dk_2\ e^{-N^2\beta f(m_1,m_2,k_1,k_2)}, \tag{46}$$

with the annealed free energy

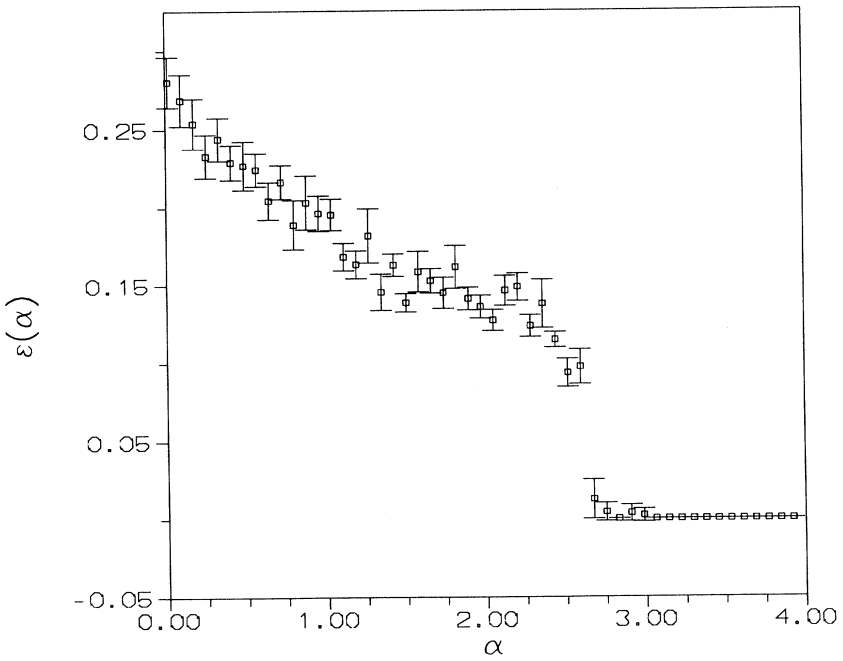

FIG. 5. Curve $\epsilon(\alpha) \equiv \epsilon_g(\beta, \alpha)$ (fixed β), for a Monte Carlo training session of a Boolean model of 64 weights at $\beta = 50$, showing the discontinuous order transition from a nonvanishing ϵ_g to $\epsilon_g = 0$ at a finite α.

$$\beta f(m_1,m_2,k_1,k_2)=\frac{\alpha}{2}\ln\{1-(1-e^{-\beta})\epsilon_g[m_1,m_2]\} +k_1m_1+k_2(m_2-1) -\ln[\sinh k_1+\sqrt{\cosh^2 k_1+e^{-4k_2}}]. \quad (47)$$

Again, the partition function of the Ising model with external field in the $N\to\infty$ limit has been used. The AA predictions would follow from the saddle-point equations for this f. The result of simulating a training session at $\beta=50$ is shown in Fig. 5, where the discontinuous order transition is visible.

V. CONCLUDING REMARKS

An alternative solution to the T-C problem has been found. Is is based on a structure of replicated detectors for overlapping regions, counting four different features whose sum is invariant under $\pi/2$-generated rotations. The model's architecture is relatively simple, as one array or "family" of such detectors in a single hidden layer is enough. The learning process leading to the corresponding feature-counting task has been studied taking advantage of the one-dimensional Ising model *with external field.* The results offer further evidence, in the line of [3, 4], that some properties discovered for single perceptrons [2, 8], and conjectured to hold for a more general class of networks, are valid in hidden-layer networks with discrete evolving weights.

In the linear H model, a transition from "poor" to perfect generalization takes place. Below a certain temperature, it is a sharp drop in the average of ϵ_g for a finite value of α. Following the discussion in [2], the asymptotic value of the minimal gap per example in the discontinuous spectrum of E_t is here of $O(1/M), M=N^2$, i.e., smaller than $O(1/\sqrt{M})$. Thus, the occurrence of the sharp transition depends on the particular values of T, as indicated by the existence of T_c. Although the one-dimensional Ising model by itself has no genuine phase transition, when embedded in this system as an entropic contribution to the free energy (28), it is capable of inducing one.

The discontinuous drop takes place at all temperatures after replacing the linear output with a Boolean unit, at least for a certain choice of the external threshold. This can be explained by a different asymptotic form of the minimal gap per example, which now is of $O(1/\sqrt{M})$, ultimately caused by the arccos function in (43). We therefore have a result analogous to those for the domain counter and the "contiguity problem" in [3] concerning hidden-layer networks. This type of alteration parallels the change in behavior when—considering systems with "Ising" weights—one goes from linear to Boolean perceptrons [11, 12], and is effected by a similar mechanism.

The annealed approximation turns out to be quite good, even at fairly low T's. This agrees with—and has also been encouraged by—the critical analysis carried out in [8], based on a study of $G_{\rm an}$, which gives grounds to foresee this behavior both for the case of realizable rules and for networks with a single Boolean output. Even though these conclusions were drawn by considering single-layer perceptrons, the analogous dependence of ϵ on the weights allows us to carry them over to our systems. The two models here presented—linear and Boolean—are about rules which are clearly realizable and, in addition, the second has a single binary output. Yet, we must not discard the possible presence of spin-glass effects manifested as noticeable slowing down in some of the simulations performed.

A disadvantage of our models is nonlocality in the sense that patterns containing the adequate number of features give the same result independently of their spatial separation (our objective was not recognition, but rather discrimination between different classes of fixed shapes). This issue has been partly dealt with, though for a different scheme, elsewhere [7]. However, a simple answer would be the addition of a preprocessor analyzing the two-spectrum and thus sieving out any pattern different from a T or C block.

The letter problem has just been a good motivation, as the sort of design introduced can be helpful for any task realizable by counting detections of similar kinds. The presence of discontinuous drops in the generalization error signals the possibility of attaining sharp increases in learning efficiency by a good choice of the number of examples and of the training temperature.

ACKNOWLEDGMENTS

I am grateful to G. Parisi for his contributions and support, to M. Virasoro for a fruitful discussion, to S. Fusi and P. del Giudice for reading typescripts of earlier versions, to C.J. Pérez-Vicente, A. Planes, and F. Ritort for critical observations, and to Ministry of Education and Science (Madrid) for FPIE financial support of the "Doctores y Tecnólogos" subprogram.

APPENDIX: EVALUATION OF $G_{\rm an}(\omega)$

In multilayer networks with random discrete weights and linear output, one often encounters error functions of the type

$$\epsilon(\omega;S)=\frac{1}{M^2}\left[\sum_{\substack{I,J=1\\ I\neq J}}^{M}\mathcal{A}_{IJ}S_IS_J+\sum_{I=1}^{M}\mathcal{B}_IS_I\right]^2. \quad (A1)$$

$\mathcal{A}$ can be taken to be symmetric, as the antisymmetric part does not contribute. $I,J,\ldots$ are general indices, and can label bidimensional sites. We assume that both the $\mathcal{A}_{IJ}$'s and $\mathcal{B}_I$'s are polynomial functions of the ω's, with coefficients of the order of one unit. Thus, inside $[\cdots]^2$ there are terms at most quadratic in the S's, of $O(M^0)$, and with randomly alternating signs.

First, consider the calculation of $\epsilon_g(\omega)=\langle\langle\epsilon(\omega;S)\rangle\rangle_S$. Using

$$\langle\langle S_I\rangle\rangle_S = 0, \quad \langle\langle S_I S_J S_K\rangle\rangle_S = 0,$$
$$\langle\langle S_I S_J\rangle\rangle_S = \delta_{IJ}, \tag{A2}$$
$$\langle\langle S_I S_J S_K S_L\rangle\rangle_S|_{I\neq J, K\neq L} = \tfrac{1}{2}(\delta_{IK}\delta_{JL} + \delta_{IL}\delta_{JK}),$$

we get

$$\epsilon_g(\omega) = \frac{1}{M^2}\left[\sum_{\substack{I,J\\ I\neq J}} \mathcal{A}_{IJ}^2 + \sum_I \mathcal{B}_I^2\right]. \tag{A3}$$

Now, we go to $G_{\rm an}$,

$$e^{-G_{\rm an}(\omega)} = \langle\langle e^{-\beta\epsilon(\omega;S)}\rangle\rangle_S$$
$$= \sum_{n=0}^{\infty} \frac{1}{n!}\left(-\frac{\beta}{M^2}\right)^n \times \left\langle\!\!\left\langle \left[\sum_{\substack{I,J\\ I\neq J}} \mathcal{A}_{IJ} S_I S_J + \sum_I \mathcal{B}_I S_I\right]^n \right\rangle\!\!\right\rangle_S . \tag{A4}$$

By multinomial expansions of $[\cdots]^n$ and employing (A2), we find

$$\left\langle\!\!\left\langle \left[\sum_{\substack{I,J=1\\ I\neq J}}^{M} \mathcal{A}_{IJ} S_I S_J + \sum_{I=1}^{M} \mathcal{B}_I S_I\right]^n \right\rangle\!\!\right\rangle_S$$
$$= \text{even terms in } \mathcal{A} \text{ and } \mathcal{B} \text{ of } \left[\sum_{\substack{I,J=1\\ I\neq J}}^{M} \mathcal{A}_{IJ} + \sum_{I=1}^{M} \mathcal{B}_I\right]^{2n}. \tag{A5}$$

Bearing in mind that the $\mathcal{A}_{IJ}$'s and $\mathcal{B}_I$'s are of $O(M^0)$, and assuming that they can have alternating signs as a result of the random variation of ω, then, by counting the number of terms of every type, we realize that the right-hand side is

$$(2n-1)!!\left[\sum_{\substack{I,J=1\\ I\neq J}}^{M} \mathcal{A}_{IJ}^2 + \sum_{I=1}^{M} \mathcal{B}_I^2\right]^n + O(M^{2n-1})$$
$$= (2n-1)!! M^{2n}[\epsilon_g(\omega)]^n + O(M^{2n-1}), \tag{A6}$$

$[(-1)!! \equiv 0]$ where the leading term is of $O(M^{2n})$. Inserting this into (A4), we are left with

$$e^{-G_{\rm an}(\omega)} = \sum_{n=0}^{\infty} \frac{(2n-1)!!}{n!}(-\beta)^n[\epsilon_g(\omega)]^n + O(M^{-1}). \tag{A7}$$

Recalling the square-root Taylor expansion

$$\frac{1}{\sqrt{1+x}} = \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n}\frac{(2n-1)!!}{n!}x^n,$$

one realizes that, in the $M \to \infty$ limit, the expression found is $e^{-G_{\rm an}(\omega)} = [1 + 2\beta\epsilon_g(\omega)]^{-1/2}$. Hence

$$G_{\rm an}(\omega) = \tfrac{1}{2}\ln[1 + 2\beta\epsilon_g(\omega)]. \tag{A8}$$

* Present address: Dept. MAiA, Fac. de Matemàtiques, Universitat de Barcelona, Granvia 585, 08071 Barcelona, Spain.

[1] D. Hansel and H. Sompolinsky, Europhys. Lett. **11**, 687 (1990).
[2] H. Sompolinsky, N. Tishby, and H.S. Seung, Phys. Rev. Lett. **65**, 1683 (1990).
[3] H. Sompolinsky and N. Tishby, Europhys. Lett. **13**, 567 (1990).
[4] I. Kocher and R. Monasson, Int. J. Neural Systems **2**, 115 (1991).
[5] M.L. Minsky and S.A. Papert, *Perceptrons* (MIT, Cambridge, MA, 1969) [expanded edition (1988)].
[6] D.E. Rumelhart and J.L. McClelland, *Parallel Distributed Processing* (MIT, Cambridge, MA, 1986), Vol. 1.
[7] A. Romeo (unpublished).
[8] H.S. Seung, H. Sompolinsky, and N. Tishby (unpublished).
[9] G. Parisi and F. Slanina, Europhys. Lett. **17**, 497 (1992).
[10] K. Binder, *Applications of the Monte Carlo Method in Statistical Physics* (Springer, Berlin, 1984).
[11] E. Gardner and B. Derrida, J. Phys. A **22**, 1983 (1989).
[12] G. Györgyi, Phys. Rev. A **41**, 7097 (1990).