Statistical mechanics of learning from examples

H. S. Seung and H. Sompolinsky
Racah Institute of Physics and Center for Neural Computation, Hebrew University, Jerusalem 91904, Israel and AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974

N. Tishby*
AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974
(Received 16 May 1991; revised manuscript received 23 September 1991)

Learning from examples in feedforward neural networks is studied within a statistical-mechanical framework. Training is assumed to be stochastic, leading to a Gibbs distribution of networks characterized by a temperature parameter T. Learning of realizable rules as well as of unrealizable rules is considered. In the latter case, the target rule cannot be perfectly realized by a network of the given architecture. Two useful approximate theories of learning from examples are studied: the high-temperature limit and the annealed approximation. Exact treatment of the quenched disorder generated by the random sampling of the examples leads to the use of the replica theory. Of primary interest is the generalization curve, namely, the average generalization error ϵ_g versus the number of examples P used for training. The theory implies that, for a reduction in ϵ_g that remains finite in the large-N limit, P should generally scale as αN, where N is the number of independently adjustable weights in the network. We show that for smooth networks, i.e., those with continuously varying weights and smooth transfer functions, the generalization curve asymptotically obeys an inverse power law. In contrast, for nonsmooth networks other behaviors can appear, depending on the nature of the nonlinearities as well as the realizability of the rule. In particular, a discontinuous learning transition from a state of poor to a state of perfect generalization can occur in nonsmooth networks learning realizable rules. We illustrate both gradual and continuous learning with a detailed analytical and numerical study of several single-layer perceptron models. Comparing with the exact replica theory of perceptron learning, we find that for realizable rules the high-temperature and annealed theories provide very good approximations to the generalization performance. Assuming this to hold for multilayer networks as well, we propose a classification of possible asymptotic forms of learning curves in general realizable models. For unrealizable rules we find that the above approximations fail in general to predict correctly the shapes of the generalization curves. Another indication of the important role of quenched disorder for unrealizable rules is that the generalization error is not necessarily a monotonically increasing function of temperature. Also, unrealizable rules can possess genuine spin-glass phases indicative of degenerate minima separated by high barriers.

PACS number(s): 87.10+e, 02.50+s, 05.20−y

I. INTRODUCTION

In recent years, many attempts have been made to train layered feedforward neural networks to perform computational tasks, such as speech recognition [1] and generation [2], handwriting recognition [3], and protein structure prediction [4]. These networks have also been used as models for neurobiological systems [5, 6], and have been employed as metaphors for cognitive processes such as learning, generalization, and concept formation [7].

Learning in neural networks, as well as in other parametric models [8], has also attracted considerable theoretical interest. The activity in this area has centered on two issues. The first is the question of representation, or *realizability*. Given a network of some architecture and size, is there a set of weights that makes the network perform the desired task? The second is the question of *learning*. Given that such a network exists, can its structure and parameters be found with a reasonable amount of time, computational resources, and training data?

Here we focus on the question of learning. We further restrict our scope to supervised learning from examples, which relies on a training set consisting of examples of the target task. The training algorithm uses the examples to find a set of network weight values that perform the task well. The most widely used class of training algorithms works by optimizing a suitable cost function that quantifies the error on the training set.

Such learning algorithms have several potential difficulties. The algorithms may become trapped in local minima that are far from optimal. Furthermore, finding good minima may require prohibitively long convergence times. Finally, there is no guarantee that good performance on a training set also leads to good performance on novel inputs. This last issue, the ability of adaptive systems to *generalize* from a limited number of examples, is the focus of the present work. Understanding the determinants of generalization ability is crucial for devising machine learning strategies, as well as for obtaining insight into learning processes in biological systems.

Our study is based on a statistical-mechanical (SM)

formulation of learning in neural networks. The training procedure is assumed to be stochastic, leading to a Gibbs distribution of network weights. The performances of the system on the training set as well as on novel inputs are calculated as appropriate *thermal* averages on the Gibbs distribution in weight space and *quenched* averages on the sampling of examples. These averages provide an accurate account of the *typical* behavior of large networks.

The currently dominant approach in computational learning theory is based on Valiant's learning model and on the notion of *probably almost correct* (PAC) learning [9, 10]. The main achievements of this approach are general bounds on the probability of error on a novel input for a given size of the training set [11, 12], as well as classification of learning problems according to their time complexity [9, 13]. Most of these (sample complexity) combinatorial bounds depend on the specific structure of the model and the complexity of the task through only a single number, known as the Vapnik-Chervonenkis dimension [14–16]. Generally, they are independent of the specific learning algorithm or distribution of examples. The generality of the PAC approach is also its main deficiency, since it is dominated by the worst case, atypical behavior. Our statistical-mechanical approach thus differs considerably from the PAC learning theory in that it can provide precise quantitative predictions for the typical behavior of specific learning models.

The SM formalism can also be applied to certain learning models for which few PAC results are yet known. Despite recent works which extend the original PAC framework [12, 17], most PAC theorems apply to *realizable* tasks, namely tasks that can be performed perfectly by the network, given enough examples. In many real life problems the target task can only be approximated by the assumed architecture of the network, so the task is *unrealizable*. In addition, many of the PAC learning results are limited to networks with threshold decision elements, although in many applications analog neurons are used. The SM approach is close in its spirit, though not in its scope and results, to the Bayesian information-theoretic approach, recently applied also to continuous networks [17, 18].

A SM approach to learning from examples was first proposed by Carnevali and Patarnello [19], and further elaborated by Tishby, Levin, and Solla [20, 21]. Del Giudice, Franz, and Virasoro, and Hansel and Sompolinsky applied spin-glass theory to study perceptron learning of a classification task [22]. Gardner and Derrida [23] and Györgyi and Tishby [24, 25] have used these methods for studying learning of a perceptron rule. Related models have been studied in Refs. [26, 27]. However the extent of applicability of results gained from these specific toy models to more general circumstances has remained unknown.

Recently an interesting attempt to characterize generic generalization performance has been put forward by Schwartz *et al.* [28]. This work suffers from two basic deficiencies. First, the analysis relies on an approximation whose validity has not been addressed. In fact this approximation is closely related to the well-known *annealed approximation* (AA) in the statistical mechanics of random systems. Although the AA simplifies enormously the theoretical analysis of these complex systems, in most interesting cases it is known to be unreliable, sometimes even in its qualitative predictions. The second problem is that no attention has been given to the dependence of performance on system size. In fact, the behavior of large systems may be quite different from that of *small-size* ones, and its analysis is more involved.

In the present study we attempt to characterize the generic behaviors of learning from examples in large layered networks. In particular we investigate the expected rate of improvement of the generalization with an increasing number of examples, denoted by the *generalization curve*. The PAC theory bounds the generalization curve by an inverse power law. Such a gradual improvement has also been observed in computer experiments of supervised learning [20, 29]. In other cases, however, one observes a rather sharp improvement when a critical number of examples is reached [20, 28, 30].

These seemingly conflicting behaviors have analogies in psychological studies of animal learning. The dichotomy between gradual and sudden learning is at the heart of the long-standing controversy between the behaviorist [31] and the gestalt [32] approaches to learning in the cognitive sciences. In this debate the underlying assumption has been that a learning process that is based on incremental modifications of the internal structure of the system can yield only gradual improvements in performance. The sudden appearance of concept understanding was therefore related to preexisting strong biases towards the learned concept, or to mysterious holistic learning mechanisms.

In the present study we show that in large systems, a sudden emergence of good generalization ability can arise even within the framework of incremental microscopic training algorithms. We analyze the conditions under which such *discontinuous transitions to perfect learning* occur. Also, we study the asymptotic forms of learning curves in cases where they are smooth. Other issues addressed in this work include (i) the consequences of the annealed approximation for learning in large networks and the scope of its validity, (ii) the properties of learning unrealizable rules, (iii) the possible emergence of spin-glass phenomena associated with the frustration and randomness induced by the random sampling of examples, (iv) how the nonlinearities inherent in the network operation affect its performance, and (v) the effect of stochastic training (noise in the learning dynamics) on generalization performance.

We address these issues by combining general results from the SM formulation of learning with detailed analytical and numerical studies of specific models. The specific examples studied here are all of learning in a single-layer perceptron models, which are significantly poorer in computational capabilities than multilayer networks. Even these simple models exhibit nontrivial generalization properties. Indeed, even the realization of random dichotomies in a perceptron with *binary* weights is a hard problem both theoretically and computationally (see, e.g., Krauth and Mézard [33] and also [34]). Here we study learning from examples in a perceptron with

real-valued weights as well as with binary weights. Some of the results found here for the perceptron models have been recently shown to exist in two-layer models also [35, 36]. Furthermore, a perceptron with strong constraints on the range of values of its weights can be thought of as representing a nonlinearity generated by a multilayer system.

In Sec. II we present two useful approximations to the SM of learning from examples: a high-temperature theory, and the above-mentioned annealed approximation. Several general consequences of these approximations as well as their range of validity are discussed. We then present the full theory, based on the replica method of averaging over quenched disorder, and derive from it some general results. In Sec. III we derive an inverse power law for the learning curves in the case of smooth networks, where the training energy is a differentiable function of the weights.

Learning curves of nonsmooth networks do not have a single universal shape. In order to elucidate the possible behavior of such networks, we study in Sec. IV perceptron learning models where both the target rules and the trained networks are single-layer perceptrons. In Sec. V we focus on specific examples of realizable perceptron rules. We study in detail the case of perceptrons with binary weights, where discontinuous transitions in learning performance occur. In addition, we investigate the spin-glass phases that exist in these models at low temperatures and small number of examples per weight.

The annealed approximation has proved to yield qualitatively correct predictions for most of the properties of the realizable perceptron models. In Sec. VI we show that this is not the case for unrealizable rules. We investigate two models of unrealizable perceptron rules where the architecture of the trained perceptron is not compatible with the target rule. Spin-glass phases are found in the unrealizable models, even at large number of examples per weight. Also the generalization error as a function of temperature may have a minimum at nonzero T, demonstrating the phenomenon of overtraining. Section VII summarizes the results and their implications. A preliminary report on some of this work appeared previously in Ref. [37].

II. GENERAL THEORY

A. Learning from examples

We consider a network with M input nodes S_i ($i = 1, \ldots, M$), N synaptic weights W_i ($i = 1, \ldots, N$), and a single output node $\sigma = \sigma(\mathbf{W}; \mathbf{S})$. The quantities $\mathbf{S}$ and $\mathbf{W}$ are M- and N-component vectors denoting the input states and the weight states, respectively. For every $\mathbf{W}$, the network defines a map from $\mathbf{S}$ to σ. Thus the weight space corresponds to a class of functions, constrained by the architecture of the network. Learning can be thought of as a search through weight space to find a network with desired properties.

In supervised learning, the weights of the network are tuned so that it approximates as closely as possible a target function $\sigma_0(\mathbf{S})$. One way of achieving this is to provide a set of *examples* consisting of P input-output pairs $(\mathbf{S}^l, \sigma_0(\mathbf{S}^l))$, with $l = 1, \ldots, P$. We assume that each input $\mathbf{S}^l$ is chosen at random from the entire input space according to some normalized *a priori* measure denoted $d\mu(\mathbf{S})$. The examples can be used to construct a *training energy*

$$E(\mathbf{W}) = \sum_{l=1}^{P} \epsilon(\mathbf{W}; \mathbf{S}^l) , \tag{2.1}$$

where the *error function* $\epsilon(\mathbf{W}; \mathbf{S})$ is some measure of the deviation of the network's output $\sigma(\mathbf{W}; \mathbf{S})$ from the target output $\sigma_0(\mathbf{S})$. The error function should be zero whenever the two agree, and positive everywhere else. A popular choice is the quadratic error function

$$\epsilon(\mathbf{W}; \mathbf{S}) = \frac{1}{2}[\sigma(\mathbf{W}; \mathbf{S}) - \sigma_0(\mathbf{S})]^2 . \tag{2.2}$$

Training is usually accomplished by minimizing the energy, for example via gradient descent

$$\frac{\partial \mathbf{W}}{\partial t} = -\nabla_{\mathbf{W}} E(\mathbf{W}) . \tag{2.3}$$

The training energy measures the network's performance on a limited set of examples, whereas the ultimate goal is to find a network that performs well on all inputs, not just those in the training set. The performance of a given network $\mathbf{W}$ on the whole input space is measured by the *generalization function*. It is defined as the average error of the network over the whole input space, i.e.,

$$\epsilon(\mathbf{W}) = \int d\mu(\mathbf{S})\, \epsilon(\mathbf{W}; \mathbf{S}) . \tag{2.4}$$

We distinguish between learning of *realizable rules* and *unrealizable rules*. Realizable rules are those target functions $\sigma_0(\mathbf{S})$ that can be completely realized by at least one of the networks in the weight space. Thus in a realizable rule there exists a weight vector $\mathbf{W}^*$ such that

$$\epsilon(\mathbf{W}^*, \mathbf{S}) = 0 \quad \text{for all } \mathbf{S} , \tag{2.5}$$

or, equivalently, $\epsilon(\mathbf{W}^*) = 0$. An *unrealizable* rule is a target function for which

$$\epsilon_{\min} = \min_{\mathbf{W}} \epsilon(\mathbf{W}) > 0 . \tag{2.6}$$

Unrealizable rules occur in two basic situations. In the first, the data available for training are corrupted with noise, making it impossible for the network to reproduce the *data* exactly, even with a large training set. This case has been considered by several authors, including Refs. [24] and [25]. Here we will not address this case explicitly. A second situation, which will be considered, is when the network architecture is restricted in a manner that does not allow an exact reproduction of the target rule itself.

B. Learning at finite temperature

We consider a stochastic learning dynamics that is a generalization of Eq. (2.3). The weights evolve according to a relaxational Langevin equation

$$\frac{\partial \mathbf{W}}{\partial t} = -\nabla_{\mathbf{W}} E(\mathbf{W}) - \nabla_{\mathbf{W}} V(\mathbf{W}) + \boldsymbol{\eta}(t), \tag{2.7}$$

where $\boldsymbol{\eta}$ is a white noise with variance

$$\langle \eta_i(t)\eta_j(t')\rangle = 2T\delta_{ij}\delta(t-t') \ . \tag{2.8}$$

We have added also a potential $V(\mathbf{W})$ that represents possible constraints on the range of weights. This term depends on the assumptions about the *a priori* distribution of $\mathbf{W}$ and does not depend on the examples. The above dynamics tends to decrease the energy, but occasionally the energy may increase due to the influence of the thermal noise. At $T = 0$, the noise term drops out, leaving the simple gradient descent equation (2.3). The above equations are appropriate for continuously varying weights. We will also consider weights that are constrained to discrete values. In such cases the analog of (2.7) is a discrete-time Monte Carlo algorithm, similar to that used in simulating Ising systems [38].

In simulated annealing algorithms for optimization problems, thermal noise has been used to prevent trapping in local minima of the energy [39]. The temperature is decreased slowly so that eventually at $T \approx 0$ the system settles to a state with energy near the global energy minimum. Although thermal noise could play the same role in the present training dynamics, it may play a more essential role in achieving good learning. Since the ultimate goal is to achieve good generalization, reaching the global minimum of the training energy may not be necessary. In fact, in some cases training at fixed finite temperature may be advantageous, as it may prevent the system from *overtraining*, namely finding an accurate fit to the training data at the expense of good generalization ability. Finally, often there are many nearly degenerate minima of the training error, particularly when the available data set is limited in size. In these cases it is of interest to know the properties of the ensemble of solutions. The stochastic dynamics provides a way of generating a useful measure, namely a Gibbs distribution, over the space of the solutions.

In the present work, we study only long-time properties. As is well known, Eq. (2.7) generates at long times a Gibbs probability distribution. In our case it is

$$\mathcal{P}(\mathbf{W}) = Z^{-1} e^{-\beta E(\mathbf{W})} \ , \tag{2.9}$$

where the variance of the noise in the training procedure now becomes the temperature $T = 1/\beta$ of the Gibbs distribution. The normalization factor Z is the partition function

$$Z = \int d\mu(\mathbf{W}) \exp[-\beta E(\mathbf{W})], \tag{2.10}$$

and we have incorporated the contribution from $V(\mathbf{W})$ into the *a priori* normalized measure in weight space, $d\mu(\mathbf{W})$. The powerful formalism of equilibrium statistical mechanics may now be applied to calculate thermal averages, i.e., averages with respect to $P(\mathbf{W})$. They will be denoted by $\langle\,\rangle_T$. In the thermodynamic limit, such average quantities yield information about the typical performance of a network, governed by the above measure, independent of the initial conditions of the learning dynamics.

Even after the thermal average is done, there is still a dependence on the P examples $\mathbf{S}^l$. Since the examples are chosen randomly and then fixed, they represent *quenched* disorder. Thus to explore the typical behavior we must perform a second, quenched average over the distribution of example sets, denoted by $\langle\!\langle\,\rangle\!\rangle \equiv \int \prod_l d\mu(\mathbf{S}^l)$.

The *average training* and *generalization errors* are given by

$$\epsilon_t(T,P) \equiv P^{-1} \langle\!\langle\, \langle E(\mathbf{W})\rangle_T \,\rangle\!\rangle \ , \tag{2.11}$$

$$\epsilon_g(T,P) \equiv \langle\!\langle\, \langle \epsilon(\mathbf{W})\rangle_T \,\rangle\!\rangle \ . \tag{2.12}$$

The free energy F and entropy S of the network are given by

$$F(T,P) = -T \langle\!\langle \ln Z \,\rangle\!\rangle, \tag{2.13}$$

$$S(T,P) = -\left\langle\!\!\left\langle \int d\mu(\mathbf{W})\, \mathcal{P}(\mathbf{W}) \ln \mathcal{P}(\mathbf{W}) \right\rangle\!\!\right\rangle. \tag{2.14}$$

They are related by the identity

$$F = P\epsilon_t - TS \ . \tag{2.15}$$

Knowing F, the expected training error can be evaluated via

$$\epsilon_t = \frac{1}{P}\frac{\partial(\beta F)}{\partial \beta} \ , \tag{2.16}$$

and the entropy by

$$S = -\frac{\partial F}{\partial T} \ . \tag{2.17}$$

The graphs of $\epsilon_g(T,P)$ and $\epsilon_t(T,P)$ as functions of P will be called *learning curves.*

Formally our results will be exact in the thermodynamic limit, i.e., when the size of the network approaches infinity. The relevant scale is the total number of degrees of freedom, namely the total number of (independently determined) synaptic weights N. For the limit $N \to \infty$ to be well defined we envisage that the problem at hand as well as the network architecture allow for a uniform scaleup of N. However, our results should provide a good approximation to the behavior of networks with a fixed large size.

The correct thermodynamic limit requires that the energy function be extensive, i.e., proportional to N. The consequences of this requirement can be realized by averaging Eq. (2.1) over the example sets, yielding

$$\langle\!\langle\, E(\mathbf{W}) \,\rangle\!\rangle = P\epsilon(\mathbf{W}). \tag{2.18}$$

Hence, assuming that $\epsilon(\mathbf{W})$ is of order 1, the number of examples should scale as

$$P = \alpha N, \tag{2.19}$$

where the proportionality constant α remains finite as N grows. This scaling guarantees that both the entropy and the energy are proportional to N. The balance between the two is controlled by the noise parameter T, which remains finite in the thermodynamic limit. A formal derivation of this scaling is given below using the replica method.

Using the definitions Eqs. (2.11) and (2.12) and the convexity of the free energy, one can show that

$$\epsilon_t(T,\alpha) \leq \epsilon_g(T,\alpha) \tag{2.20}$$

for all T and α (see Appendix A). We will show below that, as the number of examples P increases, the deviations of the energy function from its average form, Eq. (2.18), become increasingly small. This implies that for any fixed temperature, increasing α leads to $\epsilon_g \to \epsilon_{\min}, \epsilon_t \to \epsilon_{\min}, \quad \alpha \to \infty$.

C. High-temperature limit

A simple and interesting limit of the learning theory is that of high temperatures. This limit is defined so that both T and α approach infinity, but their ratio remains constant:

$$\beta\alpha = \text{finite} , \quad \alpha \to \infty , \quad T \to \infty . \tag{2.21}$$

In this limit E can simply be replaced by its average Eq. (2.18), and the fluctuations δE, coming from the finite sample of randomly chosen examples, can be ignored. To see this we note that δE is of order $\sqrt{P}$. The leading contribution to βF from the term $\beta\delta E$ in Z is proportional to $\beta^2 \langle\!\langle (\delta E)^2 \rangle\!\rangle \approx N\alpha\beta^2$. This is down by a factor of β compared to the contribution of the average term, which is of the order $N\alpha\beta$. Thus, in this limit, the equilibrium distribution of weights is given simply by

$$\mathcal{P}_0(\mathbf{W}) = Z^{-1} \exp[-N\beta\alpha\epsilon(\mathbf{W})] , \tag{2.22}$$

where

$$Z_0 = \int d\mu(\mathbf{W}) \exp[-N\beta\alpha\epsilon(\mathbf{W})] . \tag{2.23}$$

The subscript 0 signifies that the high-temperature limit is the zeroth order term of a complete high-temperature expansion, derived in Appendix B.

In the high-T limit, it is clear from Eq. (2.22) that all thermodynamic quantities, including the average training and generalization errors, are functions only of the *effective temperature* T/α. It should be emphasized that the present case is different from most high-temperature limits in statistical mechanics, in which all states become equally likely, regardless of energy. Here the simultaneous $\alpha \to \infty$ limit guarantees nontrivial behavior, with contributions from both energy and entropy. In particular, as the effective temperature T/α decreases, the network approaches the optimal ("ground state") weight vector $\mathbf{W}^*$, which minimizes $\epsilon(\mathbf{W})$. This behavior is similar to the $T =$ finite, $\alpha \to \infty$ limit of (2.58) below.

It is sometimes useful to discuss the microcanonical version of the statistical mechanics of learning in the high-T limit. Equation (2.23) can be written as

$$Z_0 = \int d\epsilon \exp[-N\beta\alpha f(\epsilon)] , \tag{2.24}$$

where the free energy per weight of all networks whose generalization error equals ϵ is

$$f(\epsilon) = \epsilon - \frac{T}{\alpha} s(\epsilon) . \tag{2.25}$$

The function $s(\epsilon)$ is the entropy per weight of all the networks with $\epsilon(\mathbf{W}) = \epsilon$, i.e.,

$$s(\epsilon) = N^{-1} \ln \int d\mu(\mathbf{W})\, \delta(\epsilon(\mathbf{W}) - \epsilon) . \tag{2.26}$$

In the large-N limit the expected generalization error is simply given by

$$\beta\alpha = \frac{\partial s}{\partial \epsilon} . \tag{2.27}$$

Thus the properties of the system in the high-T limit are determined by the dependence of the entropy on generalization error.

From the theoretical point of view, the high-T limit simply characterizes models in terms of an effective energy function $\epsilon(\mathbf{W})$ which is often a rather smooth function of $\mathbf{W}$. The smoothness of the effective energy function also implies that the learning process at high temperature is relatively fast. One does not expect to encounter many local minima, although a few large local minima may still remain, as will be seen in some of the models below. Another feature of learning at high temperature is the lack of a difference between the expected training and generalization errors, i.e., $\epsilon_g - \epsilon_t$. From Eq. (2.22) and the definitions Eqs. (2.11) and (2.12) it follows that $\epsilon_t = \epsilon_g$ in the high-T limit. Of course the price that one pays for learning at high temperature is the necessity of a large training set, as α must be at least of order T.

D. The annealed approximation

Another useful approximate method for investigating learning in neural networks is the annealed approximation, or AA for short. It consists of replacing the average of the logarithm of Z, Eq. (2.13), by the logarithm of the average of Z itself. Thus the annealed approximation for the average free energy F_{an} is

$$-\beta F_{\text{an}} = \ln \langle\!\langle Z \rangle\!\rangle . \tag{2.28}$$

Using the convexity of the logarithm function, it can be shown that the annealed free energy is a lower bound for the true quenched value,

$$F_{\text{an}} \leq F . \tag{2.29}$$

Whether this lower bound can actually serve as a good approximation will be examined critically in this work.

Using Eqs. (2.10) and (2.1) one obtains

$$\langle\!\langle Z \rangle\!\rangle = \int d\mu(\mathbf{W}) e^{-PG_{\text{an}}(\mathbf{W})} , \tag{2.30}$$

$$G_{\rm an}(\mathbf{W}) = -\ln \int d\mu(\mathbf{S})\, e^{-\beta\epsilon(\mathbf{W};\mathbf{S})} . \tag{2.31}$$

The generalization and training errors are approximated by

$$\epsilon_g = \frac{1}{\langle\langle Z \rangle\rangle} \int d\mu(\mathbf{W})\epsilon(\mathbf{W}) e^{-PG_{\rm an}(\mathbf{W})} , \tag{2.32}$$

$$\epsilon_t = \frac{1}{\langle\langle Z \rangle\rangle} \int d\mu(\mathbf{W}) \frac{\partial G_{\rm an}(\mathbf{W})}{\partial\beta} e^{-PG_{\rm an}(\mathbf{W})} . \tag{2.33}$$

1. Single Boolean output

A particularly simple case is that of an output layer consisting of a single Boolean output unit. In this case $\epsilon(\mathbf{W};\mathbf{S}) = 1$ or 0 only, so that

$$G_{\rm an}(\mathbf{W}) = -\ln[1-(1-e^{-\beta})\epsilon(\mathbf{W})] . \tag{2.34}$$

Since $G_{\rm an}$ depends on $\mathbf{W}$ only through $\epsilon(\mathbf{W})$, which is of order 1, we can write a microcanonical form of the AA, analogous to what was done for the high-T limit in Eqs. (2.24)–(2.27). The annealed partition function $\langle\langle Z \rangle\rangle$ takes the form

$$\langle\langle Z \rangle\rangle = \int d\epsilon \, \exp N[G_0(\epsilon) - \alpha G_{\rm an}(\epsilon)] , \tag{2.35}$$

where

$$G_{\rm an}(\epsilon) \equiv -\ln[1-(1-e^{-\beta})\epsilon] \tag{2.36}$$

$$G_0(\epsilon) \equiv N^{-1} \ln \int d\mu(\mathbf{W})\delta(\epsilon - \epsilon(\mathbf{W})) . \tag{2.37}$$

The function $NG_0(\epsilon)$ is the logarithm of the density of networks with generalization error ϵ. At finite temperature, it is different from the annealed entropy $S_{\rm an} \equiv -\partial F_{\rm an}/\partial T$, which is the logarithm of the density of networks with training error ϵ. However, since $\epsilon_t = \epsilon_g$ in the high-temperature limit, NG_0 approaches $S_{\rm an}$ as $T \to \infty$.

In the thermodynamic limit ($N \to \infty$), the integral (2.35) is dominated by its saddle point. Thus at any given α and T the value of the average generalization error is given by minimizing the free energy $f(\epsilon)$, where $-\beta f \equiv G_0 - \alpha G_{\rm an}$. This leads to the implicit equation

$$\left.\frac{\partial G_0}{\partial\epsilon}\right|_{\epsilon=\epsilon_g} = \frac{\alpha(1-e^{-\beta})}{1-(1-e^{-\beta})\epsilon_g} , \tag{2.38}$$

which is analogous to the high-T result Eq. (2.27). It is interesting to note that in this case the AA predicts a simple relation between the training and generalization errors. Using Eq. (2.33) above, one obtains

$$\epsilon_t = \frac{e^{-\beta}\epsilon_g}{1-(1-e^{-\beta})\epsilon_g}, \tag{2.39}$$

where ϵ_g is the average generalization error given by (2.38), or, equivalently,

$$\epsilon_g = \frac{\epsilon_t}{e^{-\beta}+(1-e^{-\beta})\epsilon_t}. \tag{2.40}$$

To the extent that the annealed approximation is valid, this relation could be used in actual applications to estimate the generalization error from the measured training error.

2. The annealed approximation as a dynamics in example space

The above annealed results only approximate the learning procedure described in Eq. (2.7). However, they can be viewed as the exact theory for a dynamic process where both the weights and the examples are updated by a stochastic dynamics, similar to Eq. (2.7), involving the same energy function, i.e.,

$$\frac{\partial\mathbf{W}}{\partial t} = -\nabla_{\mathbf{W}} E + \eta(t) , \tag{2.41}$$

$$\frac{\partial\mathbf{S}^l}{\partial t} = -\nabla_{\mathbf{S}^l} E + \eta_l(t). \tag{2.42}$$

Here E is a function of both $\mathbf{W}$ and $\mathbf{S}^l$. This dynamic process leads to a Gibbs probability distribution both in weight space and in input space

$$\mathcal{P}_{\rm an}(\mathbf{W};\mathbf{S}^l) = Z_{\rm an}^{-1} e^{-\beta E(\mathbf{W};\mathbf{S}^l)} , \tag{2.43}$$

where

$$Z_{\rm an} = \int d\mu(\mathbf{W}) \int \prod_{l=1}^{P} d\mu(\mathbf{S}^l) \exp[-\beta E(\mathbf{W})] , \tag{2.44}$$

which is exactly the annealed partition function.

From the perspective of Eqs. (2.41) and (2.42) the AA represents the behavior of a system with a distorted measure of the input space. The fact that we will find it to be a good approximation in several nontrivial cases reflects the robustness of the performance of the networks in these cases to moderate distortions of the input measure. The effect of reducing temperature is also clear. The larger β is, the larger the distortion of the *a priori* input measure due to the Gibbs factor. Consequently, one expects that deviations from the AA may be important at low T.

3. How good is the annealed approximation?

First we note that $G_{\rm an} \to \beta\epsilon(\mathbf{W})$ as $\beta \to 0$. Thus the AA is valid at high temperatures, since it reduces to the high-T limit described above. At lower temperatures the AA does seems to incorporate some of the effects of quenched disorder, in that ϵ_t is generally less than ϵ_g, in accord with Eq. (2.20). This is in contrast to the high-T limit, in which $\epsilon_g = \epsilon_t$. On the other hand, the results of the AA are in general not exact at finite temperature.

To obtain some insight into the quality of the AA at finite temperatures we examine its behavior in the limit of large α. From Eq. (2.34) it follows that in the AA the asymptotic value of the generalization error is

$$\lim_{\alpha\to\infty} \epsilon_g(T,\alpha) = \epsilon(\mathbf{W}^\dagger) , \tag{2.45}$$

where $\mathbf{W}^\dagger$ minimizes $G_{\rm an}$. In general, this vector is not necessarily the same as the vector $\mathbf{W}^*$, which minimizes

$\epsilon(\mathbf{W})$. Hence there is no guarantee that the AA correctly predicts the value of the optimal generalization error or the values of the optimal weights, except for two special cases. One is the case of a realizable rule for which $\epsilon(\mathbf{W}^*;\mathbf{S}) = 0$ for all inputs $\mathbf{S}$. Clearly the minimum of $G_{\rm an}$ in Eq. (2.31) then occurs at $G_{\rm an}(\mathbf{W}^*) = 0$. The second is the case of a network whose output layer consists of a single Boolean output unit, as discussed above. From (2.34) it is evident that the minimum of $G_{\rm an}$, in this case, coincides with the minimum of ϵ_g, and hence $\mathbf{W}^\dagger = \mathbf{W}^*$.

With respect to the training error, the AA for unrealizable rules is also inadequate: the correct limit $\epsilon_t \to \epsilon_{\rm min}$ is typically violated, even for the Boolean case, and the limit $\epsilon_t \to \epsilon_g$ does not hold either. In particular, in the $T \to 0$ limit the annealed training error approaches

$$\lim_{T\to 0} \epsilon_t(T,\alpha) = \min_{\mathbf{W},\mathbf{S}} \epsilon(\mathbf{W};\mathbf{S}) \tag{2.46}$$

since annealing both the weights and examples at zero temperature minimizes the training energy with respect to all variables. Often the right-hand side is zero, so that the AA predicts spuriously $\epsilon_t(T=0,\alpha) = 0$ for all α.

E. The replica method

To evaluate the correct behavior at all T one has to evaluate quenched averages such as Eq. (2.13) and its derivatives. Such averages are commonly studied using the replica method [40]. The average free energy is written as

$$-\beta F = \langle\!\langle \ln Z \rangle\!\rangle = \lim_{n\to 0} \frac{1}{n} \ln \langle\!\langle Z^n \rangle\!\rangle \ . \tag{2.47}$$

One first evaluates Eq. (2.47) for integral n and then analytically continues to $n = 0$. Using Eqs. (2.1) and (2.10) we obtain

$$\langle\!\langle Z^n \rangle\!\rangle = \int \prod_{\sigma=1}^{n} d\mu(\mathbf{W}^\sigma) \exp(-N\alpha \mathcal{G}_r[\mathbf{W}^\sigma]) \ , \tag{2.48}$$

where the replicated Hamiltonian is

$$\mathcal{G}_r[\mathbf{W}^\sigma] = -\ln \int d\mu(\mathbf{S}) \exp\left(-\beta \sum_{\sigma=1}^{n} \epsilon(\mathbf{W}^\sigma;\mathbf{S})\right) . \tag{2.49}$$

The average generalization error (2.12) can be rewritten using replicas as

$$\begin{aligned}\epsilon_g(T,\alpha) &= \lim_{n\to 0} \left\langle\!\!\left\langle Z^{n-1} \int d\mu(\mathbf{W}) \epsilon(\mathbf{W}) e^{-\beta E(\mathbf{W})} \right\rangle\!\!\right\rangle \\ &= \lim_{n\to 0} \int \prod_{\sigma=1}^{n} d\mu(\mathbf{W}^\sigma) \epsilon(\mathbf{W}^1) e^{-N\alpha \mathcal{G}_r[\mathbf{W}^\sigma]}\end{aligned} \tag{2.50}$$

and the average training error (2.11) as

$$\epsilon_t(T,\alpha) = \lim_{n\to 0} \frac{1}{n} \int d\mu(\mathbf{W}^\sigma) \frac{\partial \mathcal{G}_r[\mathbf{W}^\sigma]}{\partial \beta} e^{-N\alpha \mathcal{G}_r[\mathbf{W}^\sigma]} \ . \tag{2.51}$$

The simplicity of the replica formulation lies in the fact that only the *number* of examples remains as a simple prefactor in Eq. (2.48). All other example dependence has been removed, so that the replicated Hamiltonian $\mathcal{G}_r$ depends only on the form of $\epsilon(\mathbf{W};\mathbf{S})$ and on the nature of the *a priori* measure on the input space $d\mu(\mathbf{S})$. Equations (2.48) and (2.49) also make explicit the scaling of the problem. Since $\epsilon(\mathbf{W};\mathbf{S})$ is defined to be of order 1, $\mathcal{G}_r$ itself is of order 1 times n. Thus, as the integral on the weight space in Eq. (2.48) is nN dimensional, where N is the number of degrees of freedom in weight space, P must scale as N.

The AA can be obtained from Eq. (2.47) by setting $n = 1$ instead of taking the limit $n \to 0$. The replicated Hamiltonian $\mathcal{G}_r[\mathbf{W}]$ with $n = 1$ reduces to the annealed expression $G_{\rm an}(\mathbf{W})$, Eq. (2.31).

1. *Replica theory and the high-T limit*

The replica theory provides a simple derivation of the high-T limit described in Sec. II C. Since $\mathcal{G}_r$ is an intensive quantity independent of P, the high-T limit can be derived by simply expanding it in powers of β. The leading terms are

$$\begin{aligned}P\mathcal{G}_r[\mathbf{W}^\sigma] = N\Bigg(&\alpha\beta \sum_{\sigma=1}^{n} \epsilon(\mathbf{W}^\sigma) \\ &-\frac{1}{2}\alpha\beta^2 \sum_{\sigma,\rho=1}^{n} g(\mathbf{W}^\sigma,\mathbf{W}^\rho) + O(\beta^3)\Bigg) \ ,\end{aligned} \tag{2.52}$$

where

$$\begin{aligned}g(\mathbf{W}^\sigma,\mathbf{W}^\rho) = &\int d\mu(\mathbf{S})\, \epsilon(\mathbf{W}^\sigma;\mathbf{S}) \epsilon(\mathbf{W}^\rho;\mathbf{S}) \\ &-\epsilon(\mathbf{W}^\sigma)\epsilon(\mathbf{W}^\rho) \ .\end{aligned} \tag{2.53}$$

Note that g measures the correlations in the errors of two different weights on the same example.

The general form of Eq. (2.52) is similar to that of a spin-glass replica Hamiltonian [41, 42]. The first term is the one that survives the high-T limit. It represents the nonrandom part of the training energy. Taking into account only this contribution leaves the different replicas uncoupled, and hence F reduces to its high-T limit

$$-\beta F \to \ln \int d\mu(\mathbf{W}) e^{-N\beta\alpha\epsilon(\mathbf{W})} \ , \tag{2.54}$$

in which the training energy becomes proportional to the generalization function, i.e., $E(\mathbf{W}) \to P\epsilon(\mathbf{W})$. As T decreases the second term of Eq. (2.52) becomes important. This term is a coupling between different replicas which originates from the randomness of the examples.

2. Spin glasses and replica symmetry breaking

In some cases, the coupling between replicas produce only minor changes in the learning curves. In others, such terms can lead to the appearance of qualitatively different phases at low temperatures. These phases are conveniently described by the properties of the matrix

$$Q_{\mu\nu} = \frac{1}{N}\langle \mathbf{W}^\mu \cdot \mathbf{W}^\nu \rangle \ , \tag{2.55}$$

which measures the expected overlap of the weights of two copies of the system. Since the replicated Hamiltonian (2.49) is invariant under permutation of the replica indices, one naively would expect that $Q_{\mu\nu} = q$ for all $\mu \neq \nu$. The physical interpretation of q would then be

$$q = N^{-1}\langle\!\langle\, \langle \mathbf{W} \rangle_T \cdot \langle \mathbf{W} \rangle_T \,\rangle\!\rangle \ . \tag{2.56}$$

It is known as the Edwards-Anderson parameter in spin-glass theory [40]. The high-temperature phase indeed possesses this *replica symmetry*. However, as the temperature is lowered, a spontaneous *replica symmetry breaking* (RSB) can occur, signaling the appearance of a spin-glass phase. In this phase, the expected values of correlations among different replicas do depend on the replica indices.

Formally, the spin-glass phase is characterized by a nontrivial dependence of quantities such as $Q_{\mu\nu}$ on the replica indices. Physically, the spin-glass phase is marked by the existence of many degenerate ground states of the energy (or free energy) which are well separated in configuration space. The different values of $Q_{\mu\nu}$ represent the distribution of overlaps among pairs of these ground states. This degeneracy is not connected with any simple physical symmetry, but is a result of strong frustration in the system. Furthermore, these degenerate ground states occupy disconnected regions in configuration space that are separated by energy barriers that diverge with N. Such barriers are important in the context of learning, since they lead to anomalously slow learning dynamics [43–45].

F. The large-α limit

The replica formalism can also be used to investigate the behavior at a large number of examples, i.e., the $\alpha \to \infty$ limit. From Eq. (2.48) it is clear that the free energy and the training and generalization errors are all weight space integrals that are dominated by the minimum of $\mathcal{G}_r$, as $\alpha \to \infty$. Denoting this minimum by $\mathbf{W}^\sigma = \mathbf{W}^*$, we find

$$\begin{aligned} \mathcal{G}_r^{\min} &= -\ln \int d\mu(\mathbf{S}) \exp[-\beta n \epsilon(\mathbf{W}^*;\mathbf{S})] \\ &= \beta n \epsilon(\mathbf{W}^*) + O(n^2). \end{aligned} \tag{2.57}$$

This implies that $\mathbf{W}^*$ minimizes both the generalization error $\epsilon(\mathbf{W})$ and $\mathcal{G}_r$ in the $n \to 0$ limit. Hence we conclude that for any fixed temperature the training and generalization errors both approach the optimal value $\epsilon_{\min} = \epsilon(\mathbf{W}^*)$ as $\alpha \to \infty$,

$$\epsilon_g \to \epsilon_{\min}, \quad \epsilon_t \to \epsilon_{\min}, \quad \alpha \to \infty. \tag{2.58}$$

In the following section, α will be used as a control parameter in a saddle-point expansion to calculate the approach to the optimum for smooth networks.

III. SMOOTH NETWORKS

We define smooth networks to be those with continuously varying weights and error functions $\epsilon(\mathbf{W};\mathbf{S})$ that are at least twice differentiable with respect to $\mathbf{W}$ in the vicinity of the optimal weight vector $\mathbf{W}^*$. According to this definition, whether a network is smooth depends on both the smoothness of the weight space and the smoothness of the error function $\epsilon(\mathbf{W};\mathbf{S})$. In a smooth network, neither the output neuron nor the hidden neurons are saturated at the optimal $\mathbf{W}^*$. We now use the replica formalism to derive the asymptotic shape of the learning curves in these networks.

As stated above, the integrals over the weight space are dominated, as $\alpha \to \infty$, by the optimal weight vector $\mathbf{W}^*$, which minimizes both $\mathcal{G}_r$ (in the $n \to 0$ limit) and $\epsilon(\mathbf{W})$. At finite large α, the leading corrections to ϵ_g come from the immediate neighborhood of $\mathbf{W}^*$. In a smooth network we can expand $\mathcal{G}_r$ in powers of

$$\delta W_i^\sigma \equiv W_i^\sigma - W_i^* \ . \tag{3.1}$$

The linear terms vanish since $\mathbf{W}^*$ is a minimum of $\mathcal{G}_r$. The leading corrections are

$$\mathcal{G}_r \approx \mathcal{G}_r^{\min} + \frac{1}{2} \sum_{i,j,\sigma,\rho} \delta W_i^\sigma A_{ij}^{\sigma\rho} \delta W_j^\rho \ , \tag{3.2}$$

where

$$A_{ij}^{\sigma\rho} = \beta U_{ij} \delta^{\sigma\rho} - \beta^2 V_{ij} \ . \tag{3.3}$$

The matrix U_{ij} is the Hessian of the error function at the optimal weight vector $\mathbf{W}^*$, i.e.,

$$U_{ij} = \int d\mu(\mathbf{S})\, \partial_i \partial_j \epsilon(\mathbf{W}^*, \mathbf{S}) \ . \tag{3.4}$$

The symbol ∂_i denotes $\partial/\partial W_i$. The matrix V_{ij} is

$$V_{ij} = \int d\mu(\mathbf{S})\, \partial_i \epsilon(\mathbf{W}^*, \mathbf{S}) \partial_j \epsilon(\mathbf{W}^*, \mathbf{S}) \ . \tag{3.5}$$

Since $N\alpha\mathcal{G}_r$ defines a Gaussian measure in weight space, it is straightforward to calculate the average deviations of the weights from $\mathbf{W}^*$. They are

$$\langle \delta \mathbf{W}^\sigma \rangle = 0 \ , \tag{3.6}$$

$$\begin{aligned} \langle \delta W_i^\sigma \delta W_j^\rho \rangle &= (N\alpha)^{-1} \, (A^{-1})_{ij}^{\sigma\rho} \\ &= \frac{1}{N\alpha} \left[T (U^{-1})_{ij} \delta^{\sigma\rho} + (U^{-1} V U^{-1})_{ij} \right] \ , \end{aligned} \tag{3.7}$$

where we have already taken the $n \to 0$ limit. Equations (3.6) and (3.7) have a simple meaning in terms of the physical system. Equation (3.6) reads

$$\langle\!\langle \langle \delta \mathbf{W} \rangle_T \rangle\!\rangle = \langle\!\langle \langle \mathbf{W} \rangle_T \rangle\!\rangle - \mathbf{W}^* = 0 \ . \tag{3.8}$$

The diagonal element (in the replica indices) of Eq. (3.7) yields the average correlations

$$C_{ij} = \langle\!\langle \langle \delta W_i \delta W_j \rangle_T \rangle\!\rangle = \frac{1}{N\alpha}[T(U^{-1})_{ij} + (U^{-1}VU^{-1})_{ij}] \ . \tag{3.9}$$

The first term, which is proportional to T, represents the contribution of the thermal fluctuations about $\mathbf{W}^*$. The second term represents the quenched fluctuations due to the random sampling of examples. This interpretation is confirmed by inspecting the off-diagonal element of Eq. (3.7), which is

$$\langle\!\langle \langle \delta W_i \rangle_T \langle \delta W_j \rangle_T \rangle\!\rangle = \frac{1}{N\alpha}(U^{-1}VU^{-1})_{ij} \ . \tag{3.10}$$

To evaluate the corrections to the generalization error, we expand Eq. (2.12) in powers of $\delta\mathbf{W}$, yielding

$$\epsilon_g = \epsilon_{\min} + \frac{1}{2}\,\mathrm{Tr}\,UC \ . \tag{3.11}$$

Substituting Eq. (3.9) one obtains

$$\epsilon_g(T,\alpha) = \epsilon_{\min} + \left(\frac{T}{2} + \frac{\mathrm{Tr}VU^{-1}}{2N}\right)\frac{1}{\alpha} + O(\alpha^{-2}) \ . \tag{3.12}$$

The $1/\alpha$ expansion for the average training error can be evaluated by calculating first the corrections to the average free energy. Substituting Eq. (3.2) in Eq. (2.48) yields, after taking the $n \to 0$ limit

$$\beta F = N\alpha\beta\epsilon_{\min} + \frac{1}{2}[\mathrm{Tr}\,\ln(\beta U) - \beta\,\mathrm{Tr}\,VU^{-1}] + \frac{1}{2}N\ln\frac{N\alpha}{2\pi}. \tag{3.13}$$

Using Eq. (2.16) one obtains

$$\epsilon_t(T,\alpha) = \epsilon_{\min} + \left(\frac{T}{2} - \frac{\mathrm{Tr}VU^{-1}}{2N}\right)\frac{1}{\alpha} + O(\alpha^{-2}) \ . \tag{3.14}$$

The above results predict an important relationship between the expected training and generalization errors at $T = 0$. According to Eqs. (3.12) and (3.14) both errors approach the same limit $\epsilon_{\min}$ with a $1/\alpha$ power law. The coefficients of $1/\alpha$ in the two errors are identical in magnitude but different in sign yielding

$$\frac{\partial\epsilon_t}{\partial\alpha} = -\frac{\partial\epsilon_g}{\partial\alpha}, \quad \alpha \to \infty, \quad T = 0 \ . \tag{3.15}$$

This result can be used to estimate the expected generalization error from the measured training error in smooth networks.

A special case occurs when the rule to be learned by the smooth network is realizable. This means that there exists a weight vector $\mathbf{W}^*$ within the allowed weight space that has zero generalization error, i.e.,

$$\epsilon_{\min} = \epsilon(\mathbf{W}^*) = \int d\mu(\mathbf{S})\,\epsilon(\mathbf{W}^*;\mathbf{S}) = 0 \ . \tag{3.16}$$

This is equivalent to the statement that $\epsilon(\mathbf{W}^*;\mathbf{S}) = 0$ for all input vectors $\mathbf{S}$, because the error function was defined to be non-negative. Since $\mathbf{W}^*$ minimizes $\epsilon(\mathbf{W})$, we can also assume $\partial_i\epsilon(\mathbf{W}^*) = 0$, as long as $\mathbf{W}^*$ lies in the interior of the weight space.

We have $\epsilon(\mathbf{W}^*;\mathbf{S}) \le \epsilon(\mathbf{W};\mathbf{S})$, since the left-hand side is zero, and the right-hand side is non-negative. This implies that

$$\partial_i\epsilon(\mathbf{W}^*;\mathbf{S}) \ge 0 \tag{3.17}$$

for all $\mathbf{S}$, but also

$$\int d\mu(\mathbf{S})\partial_i\epsilon(\mathbf{W}^*;\mathbf{S}) = \partial_i\epsilon(\mathbf{W}^*) = 0 \ . \tag{3.18}$$

Equations (3.17) and (3.18) together imply that $\partial_i\epsilon(\mathbf{W}^*;\mathbf{S}) = 0$ for all $\mathbf{S}$, which in turn implies $V_{ij} = 0$. Finally, we have the result

$$\epsilon_g(T,\alpha) = \frac{T}{2\alpha} + O(\alpha^{-2}) \ . \tag{3.19}$$

The same holds for ϵ_t.

At zero temperature, the coefficient of $1/\alpha$ vanishes for ϵ_g and ϵ_t. Furthermore, for a smooth network learning a realizable rule, the higher-order terms also vanish at $T = 0$. This implies that there is a finite value α_c for which $\epsilon_g = \epsilon_t = 0$ for all $\alpha > \alpha_c$, at $T = 0$. An example of such behavior will be presented in Sec. V A below. Such a state we call a state of *perfect learning*.

It should be noted, however, that a realizable rule with a smooth network is an unrealistic situation. The smoothness requirement, as defined above, implies that the measure of error involves equalities and not inequalities. Therefore to realize a rule would necessitate infinite precision in determining the optimal weights. Unlike the case of discrete problems, learning tasks in smooth networks are generically unrealizable.

IV. LEARNING OF A PERCEPTRON RULE

A. General formulation

The perceptron is a network which sums a single layer of inputs S_j with synaptic weights W_j, and passes the result through a transfer function σ

$$\sigma = g\left(N^{-1/2}\sum_{j=1}^{N} W_j S_j\right) = g\left(N^{-1/2}\mathbf{W}\cdot\mathbf{S}\right) \tag{4.1}$$

where $g(x)$ is a sigmoidal function of x. The normalization $1/\sqrt{N}$ in Eq. (4.1) is included to make the argument of the transfer function be of order unity. Learning is a search through weight space for the perceptron that best approximates a target rule. We assume for simplicity that the network space is restricted to vectors that satisfy the normalization

$$\sum_{i=1}^{N} W_i^2 = N \ . \tag{4.2}$$

The *a priori* distribution on the input space is assumed to be Gaussian,

$$d\mu(\mathbf{S}) = \prod_{i=1}^{N} DS_i \ , \tag{4.3}$$

where Dx denotes the normalized Gaussian measure

$$Dx \equiv \frac{dx}{\sqrt{2\pi}} e^{-x^2/2} \ . \tag{4.4}$$

We consider only the case where the target rule is another perceptron of the form

$$\sigma^0(\mathbf{S}) = g\left(\frac{1}{\sqrt{N}}\mathbf{W}^0 \cdot \mathbf{S}\right) \ , \tag{4.5}$$

and $\mathbf{W}^0$ is a fixed set of N weights W_i^0. We assume that the teacher weights $\mathbf{W}^0$ also satisfy the normalization condition (4.2).

Training is performed by a stochastic dynamics of the form (2.7) with the training energy function (2.1). For each example the error function is taken to be

$$\epsilon(\mathbf{W};\mathbf{S}) = \frac{1}{2}\left[g\left(N^{-1/2}\mathbf{W}\cdot\mathbf{S}\right) - g\left(N^{-1/2}\mathbf{W}^0\cdot\mathbf{S}\right)\right]^2 \ . \tag{4.6}$$

The generalization function is

$$\begin{aligned}\epsilon(\mathbf{W}) &= \int D\mathbf{S}\,\epsilon(\mathbf{W};\mathbf{S}) \\ &= \int Dx \int Dy \,\frac{1}{2}[g(x\sqrt{1-R^2}+yR) - g(y)]^2\end{aligned} \tag{4.7}$$

where R is the overlap of the student network with the teacher network, i.e.,

$$R = \frac{1}{N}\mathbf{W}\cdot\mathbf{W}^0 \tag{4.8}$$

(see Appendix C). The relationship between (4.6) and (4.7) is plain, since in both cases the arguments of g are Gaussian random variables with unit variance and cross correlation R.

It is important to note that in perceptron learning, the generalization function of a network depends only on its overlap with the teacher, i.e., $\epsilon(\mathbf{W}) = \epsilon(R)$. Learning can be visualized very easily since $R = \cos\theta$, where θ is the angle between $\mathbf{W}$ and $\mathbf{W}^0$. The generalization function goes to zero as the angle between the student and the teacher weight vectors vanishes. Perfect learning corresponds to an overlap $R = 1$ or $\theta = 0$.

In the following we discuss perceptrons with either linear or Boolean outputs, and weights that are either binary or obey a spherical constraint.

For the *linear perceptron*, the transfer function is $g(x) = x$. The error function, Eq. (4.6), is in this case a quadratic function in weight space,

$$\epsilon(\mathbf{W};\mathbf{S}) = \frac{1}{2N}[(\mathbf{W}-\mathbf{W}^0)\cdot\mathbf{S}]^2 \ . \tag{4.9}$$

Averaging this function over the whole input space, we find

$$\epsilon(\mathbf{W}) = 1 - R, \tag{4.10}$$

in accord with Eq. (4.7) with a linear g.

A second output function to be considered is the *Boolean* output $g(x) = \mathrm{sgn}(x)$, which corresponds to the original perceptron model studied by Rosenblatt [46]. The *Boolean perceptron* $\sigma = \mathrm{sgn}(\mathbf{W}\cdot\mathbf{S})$ separates the input space in half via a hyperplane perpendicular to the weight vector. The error function, Eq. (4.6), is (up to a factor of 2)

$$\epsilon(\mathbf{W};\mathbf{S}) = \Theta(-(\mathbf{W}\cdot\mathbf{S})(\mathbf{W}^0\cdot\mathbf{S})) \ , \tag{4.11}$$

which is 1 when the student and teacher agree, and 0 otherwise. The generalization error

$$\epsilon(\mathbf{W}) = \frac{1}{\pi}\cos^{-1} R \tag{4.12}$$

is simply proportional to the angle between the student and teacher weight vectors.

B. The annealed approximation for perceptron learning

The annealed free energy of perceptron learning is shown in Appendix C to be

$$-\beta f = G_0(R) - \alpha G_{\mathrm{an}}(R) \ , \tag{4.13}$$

$$G_{\mathrm{an}}(R) = -\ln \int Dx \int Dy \exp\left(-\frac{\beta}{2}\left[g\left(x\sqrt{1-R^2}+yR\right) - g(y)\right]^2\right) \ , \tag{4.14}$$

$$G_0(R) = N^{-1}\ln \int d\mu(\mathbf{W})\,\delta(R - N^{-1}\mathbf{W}\cdot\mathbf{W}^0) \ . \tag{4.15}$$

The function $NG_0(R)$ is the logarithm of the density of networks with overlap R, so we will sometimes refer to it as the "entropy," even though it is not the same as the thermodynamic entropy $s = -\partial f/\partial T$. The properties of the system in the large-N limit are obtained by minimizing the free energy f, which yields

$$\frac{\partial G_0(R)}{\partial R} = \alpha \frac{\partial G_{\rm an}(R)}{\partial R} \ . \tag{4.16}$$

Solving for R one then evaluates the average generalization error via (4.7). Likewise the average training error is evaluated by differentiating $G_{\rm an}$ with respect to β, as in Eq. (2.33).

We will consider the case of a *spherical constraint* and that of an *Ising constraint*. In the perceptron with a spherical constraint, the *a priori* measure $d\mu(\mathbf{W})$ is uniform on the sphere of radius $\sqrt{N}$, given by the normalization condition, Eq. (4.2). We may write the measure more formally as

$$d\mu(\mathbf{W}) \equiv \prod_{i=1}^{N} \frac{dW_i}{\sqrt{2\pi e}} \delta(\mathbf{W}\cdot\mathbf{W} - N) \ , \tag{4.17}$$

which is normalized to $\int d\mu(\mathbf{W}) = 1$. In this case, the fraction of weight space with an overlap R is simply the volume of the $(N-2)$-dimensional sphere with radius $\sqrt{1-R^2}$. Hence the entropy $G_0(R)$, Eq. (4.15), is (in the limit of large N)

$$G_0(R) = \frac{1}{2}\ln(1-R^2) \ , \tag{4.18}$$

a result that is derived in more detail in Appendix C. The entropy diverges as $R \to 1$, as the fraction of weight space with overlap R approaches zero. Such a divergence is typical of a continuous weight space.

The Ising perceptron corresponds to a network with binary valued weights $W_i = \pm 1$, or

$$d\mu(\mathbf{W}) \equiv \prod_{i=1}^{N} dW_i[\delta(W_i - 1) + \delta(W_i + 1)] \ . \tag{4.19}$$

The entropy of Ising networks with an overlap R is given by

$$G_0(R) = -\frac{1-R}{2}\ln\frac{1-R}{2} - \frac{1+R}{2}\ln\frac{1+R}{2} \ , \tag{4.20}$$

a result derived in Appendix C. It approaches zero as $R \to 1$, meaning that there is exactly one state with $R = 1$. This nondivergent behavior is typical of discrete weight spaces.

To conclude, the picture emerging from the AA is extremely simple. The properties of the system can be expressed in terms of a single order parameter, namely, the overlap R. The stochastic fluctuations in the value of R can be neglected in the limit of large N. Hence the system almost always converges to a unique value of R given by the minimum of the free energy $f(R)$. Depending on the details of the problem, $f(R)$ can have more than one local minimum. If this happens, the equilibrium properties of the system are determined by the unique global minimum of f. The local minima represent metastable states. Starting from a value of R near one of these minima the system is likely to converge rapidly to the local minimum. It will remain in this state for a time that scales exponentially with the network size. Hence for large networks the local minima of f can be considered as stable states.

C. The replica theory of perceptron learning

The calculation of $\mathcal{G}_r$, (2.49), for perceptron learning is presented in Appendix D. The dependence of $\mathcal{G}_r$ on the weights is through the order parameters $Q_{\mu\nu}$, Eq. (2.55), and

$$R_\mu = \frac{1}{N}\mathbf{W}^\mu \cdot \mathbf{W}^0 \ , \tag{4.21}$$

which measures the overlap of the networks with the teacher. The values of these order parameters are obtained by extremizing $\mathcal{G}_r$.

In general, evaluating the saddle-point equations for $Q_{\mu\nu}$ and R_μ requires making an ansatz about the symmetry of the order parameters at the saddle point. The simplest ansatz is the replica symmetric (RS) one,

$$Q_{\mu\nu} = \delta_{\mu\nu} + (1-\delta_{\mu\nu})q \ , \tag{4.22}$$

$$R_\mu = R \ . \tag{4.23}$$

In this case the order parameters of the replica theory have simple meanings in terms of thermal and quenched averages. The order parameter q is given by Eq. (2.56), and R is the expected value of the overlap with the teacher,

$$R = \frac{1}{N}\langle\!\langle\, \langle \mathbf{W} \rangle_T \,\rangle\!\rangle \cdot \mathbf{W}^0 \ . \tag{4.24}$$

The RS free energy of perceptron learning is (see Appendix D)

$$-\beta f = \frac{1}{N}\langle\!\langle \ln Z \rangle\!\rangle = G_0(q, R, \hat{q}, \hat{R}) - \alpha G_r(q, R) \ , \tag{4.25}$$

where

$$G_0 = -\frac{1}{2}(1-q)\hat{q} - R\hat{R} + \frac{1}{N}\int Dz \ln \int d\mu(\mathbf{W}) \exp[\mathbf{W}\cdot(\mathbf{z}\sqrt{\hat{q}} + \mathbf{W}^0\hat{R})] \ , \tag{4.26}$$

$$G_r = -\int Dt \int Dy \ln \int Dx \exp\left(-\frac{1}{2}\beta\left[g\left(x\sqrt{1-q} + yR + t\sqrt{q-R^2}\right) - g(y)\right]^2\right) \ , \tag{4.27}$$

and $\mathbf{z}$ is a vector of Gaussian variables $\{z_i\}_{i=1}^N$ with $D\mathbf{z} \equiv \prod_i Dz_i$. The free energy has to be extremized with respect to the order parameters q and R, and their "conjugate" counterparts $\hat{q}$ and $\hat{R}$. Differentiating with respect to $\hat{q}$ and $\hat{R}$ yields the saddle-point equations

$$q = \frac{1}{N}\int D\mathbf{z}\,\langle \mathbf{W}\rangle_z \cdot \langle \mathbf{W}\rangle_z \ , \tag{4.28}$$

$$R = \frac{1}{N}\int D\mathbf{z}\,\langle \mathbf{W}\rangle_z \cdot \mathbf{W}^0 \ . \tag{4.29}$$

The definition of the average $\langle \mathbf{W}\rangle_z$ in these two equations reveals the meaning of the parameters $\hat{q}$ and $\hat{R}$,

$$\langle \mathbf{W}\rangle_z \equiv \frac{\int d\mu(\mathbf{W})\,\mathbf{W}\exp[\mathbf{W}\cdot(\mathbf{z}\sqrt{\hat{q}} + \mathbf{W}^0\hat{R})]}{\int d\mu(\mathbf{W})\exp[\mathbf{W}\cdot(\mathbf{z}\sqrt{\hat{q}} + \mathbf{W}^0\hat{R})]} \ . \tag{4.30}$$

In this equation, the local field $\mathbf{z}\sqrt{\hat{q}} + \mathbf{W}^0\hat{R}$ acting upon $\mathbf{W}$ consists of two parts. The first is a Gaussian random field with variance $\hat{q}$ originating from the random fluctuations of the examples. The second is the bias towards the teacher weights $\mathbf{W}^0$, with an amplitude $\hat{R}$.

In general, we know from Eq. (2.58) that $\mathbf{W}$ must approach the optimal weight vector $\mathbf{W}^*$ as $\alpha \to \infty$. For a realizable perceptron rule ($\mathbf{W}^* = \mathbf{W}^0$), this means that $R \to 1$. If $\mathbf{W}^*$ is unique, the Gibbs distribution in weight space contracts about it as $\alpha \to \infty$, which means that $q \to 1$. The approach to the optimum is reflected in a competition between the two terms of the local field: the strength of the ordering term diverges ($\hat{R} \to \infty$) and the relative strength of the disorder goes to zero ($\sqrt{\hat{q}}/\hat{R} \to 0$).

One criterion for the validity of the RS ansatz is the local stability of the RS saddle point. Often one finds that the RS solution becomes locally unstable at low temperatures, and hence invalid. In the phase diagram, the line at which the instability appears is known as the *Almeida-Thouless line* [47]. To find the true solution beyond this line, one must break replica symmetry.

For systems with discrete-valued degrees of freedom, a simpler diagnostic for RSB is available, based on the fact that such systems must have non-negative entropy. Below the *zero-entropy line*, the entropy is negative and hence the RS ansatz must be incorrect. Hence the zero-entropy line provides a lower bound for the temperature at which RSB first occurs. Since the zero entropy line is easier to calculate than the Almeida-Thouless line, we will rely on it to estimate the location of the RSB region.

Since it is generally extremely difficult to find the correct RSB solution, we will consider only the RS solutions. The only exceptions are the models of Secs. V D and VI C below, for which we analyze the first step of RSB. Otherwise, we expect that the RS solution will still serve as a good approximation in the RSB region.

V. PERCEPTRON LEARNING OF REALIZABLE RULES

A. Linear output with continuous weights

The case of a perceptron with the quadratic error function Eq. (4.9) defined on a continuous weight space is particularly simple. Krogh and Hertz [48] have done a complete analysis of the training dynamics for this model. Here we derive the equilibrium properties using the replica theory.

Applying Eqs. (4.25)–(4.27), yields $-\beta f = G_0 - \alpha G_r$, with

$$G_0 = \frac{1}{2}\lambda + \frac{1}{2}q\hat{q} - R\hat{R} - \frac{1}{2}\ln(\lambda + \hat{q}) + \frac{1}{2}\frac{\hat{R}^2 + \hat{q}}{\lambda + \hat{q}} - \frac{1}{2} \ , \tag{5.1}$$

$$G_r = \frac{1}{2}\ln[1 + \beta(1-q)] + \frac{1}{2}\frac{\beta(q - 2R + 1)}{1 + \beta(1-q)} \ . \tag{5.2}$$

The additional order parameter λ is the Lagrange multiplier associated with the spherical constraint. Extremizing f with respect to the order parameters and eliminating λ yields

$$R = \hat{R}(1-q) \ , \tag{5.3}$$

$$q = (\hat{q} + \hat{R}^2)(1-q)^2 \ , \tag{5.4}$$

$$\hat{R} = \frac{\alpha\beta}{1 + \beta(1-q)} \ , \tag{5.5}$$

$$\hat{q} = \alpha\beta^2\frac{q - 2R + 1}{[1 + \beta(1-q)]^2} \ . \tag{5.6}$$

First we consider the simple case of zero temperature. Only those weight vectors with zero training energy are allowed, i.e., those that satisfy

$$(\mathbf{W} - \mathbf{W}^0)\cdot\mathbf{S}^l = 0 \ , \quad l = 1, \ldots, P \ . \tag{5.7}$$

For $P \le N$ these homogeneous linear equations determine only the projection of $\mathbf{W} - \mathbf{W}^0$ on the subspace spanned by the P random examples $\mathbf{S}^l$. This implies that the subspace of ground states of E has a huge degeneracy; it is $N - P$ dimensional. As $P \to N$ this degeneracy shrinks and for $P \ge N$ there is a unique solution to Eq. (5.7), $\mathbf{W} = \mathbf{W}^0$, for almost every random choice of examples.

At $T = 0$ the saddle-point equations reduce to

$$q = R = \begin{cases} \alpha, & \alpha \le 1 \\ 1, & \alpha \ge 1. \end{cases} \tag{5.8}$$

For $\alpha < 1$, the fact that $q < 1$ reflects the degeneracy of the ground states, according to the definition Eq. (2.56). When α reaches the critical value

$$\alpha_c = 1 \ , \tag{5.9}$$

the degeneracy is finally broken ($q = 1$), and the training energy possesses a unique minimum $\mathbf{W} = \mathbf{W}_0$ ($R = 1$), in agreement with the simple arguments presented above. Thus there is a continuous transition to perfect learning at $\alpha = 1$.

However, this transition does not exist at any finite temperature, because of thermal fluctuations about the global minimum. From the saddle-point equations, one can calculate that the asymptotic generalization curve is given by

$$\epsilon_g = \frac{T}{2\alpha} + O(\alpha^{-2}) \ , \tag{5.10}$$

in accord with the general result (3.19) for smooth networks learning realizable rules. This means that at any finite temperature, perfect generalization is attained only in the limit of infinite α. The above results are in agreement with the dynamic theory of Krogh and Hertz [48].

It is interesting to compare the above exact results with those of the AA. Evaluating Eq. (4.14), with $g(x) = x$ yields

$$G_{\rm an} = \frac{1}{2}\ln\left[1 + 2\beta(1-R)\right] \ . \tag{5.11}$$

Adding this to (4.18) yields the annealed free energy of Eq. (4.13),

$$-\beta f(R) = \frac{1}{2}\ln(1-R^2) - \frac{\alpha}{2}\ln\left[1+2\beta(1-R)\right] \ , \tag{5.12}$$

which is extremized when

$$\frac{R}{1-R^2} = \frac{\alpha\beta}{1+2\beta(1-R)} \ . \tag{5.13}$$

The training error is given by

$$\epsilon_t = \frac{1}{\alpha}\frac{\partial\beta f}{\partial\beta} = \frac{1-R}{1+2\beta(1-R)} \ . \tag{5.14}$$

The asymptotic behavior of $\epsilon_g = 1 - R$ agrees with the correct result Eq. (5.10).

At $T = 0$, the AA predicts

$$\epsilon_g = \frac{2-2\alpha}{2-\alpha} \ , \ \alpha \le 1 \ . \tag{5.15}$$

in contrast to the true quenched result $\epsilon_g = 1-\alpha$, $\alpha \le 1$. Although the value of ϵ_g is incorrect, the second-order transition to perfect learning at $\alpha_c = 1$ is correctly predicted.

B. Boolean output with continuous weights

The Boolean perceptron with continuous weights corresponds to the original perceptron studied in [46, 49]. At zero temperature, weight vectors with zero training energy satisfy the inequalities

$$(\mathbf{W}\cdot\mathbf{S}^l)(\mathbf{W}^0\cdot\mathbf{S}^l) > 0 \ . \tag{5.16}$$

Since these inequalities do not constrain the weight space as much as the equalities (5.7), this model requires more examples for good generalization than did the preceding linear model. The quenched theory of this model has been previously studied in detail [24]. We present below a few of the results for completeness.

Since the *a priori* measure $d\mu(\mathbf{W})$ is the same as in the linear-continuous model of Sec. V C above, G_0 is again given by (5.1). For a Boolean output, G_r equals

$$G_r = -2\int_0^\infty Dy \int_{-\infty}^\infty Dt \ln\left[e^{-\beta} + (1-e^{-\beta})H(u)\right] \ , \tag{5.17}$$

where

$$u \equiv \frac{t\sqrt{q-R^2} - yR}{\sqrt{1-q}} \ , \tag{5.18}$$

and $H(x)$ is defined as in (5.54). The saddle-point equations are

$$R = \hat{R}(1-q) \ , \tag{5.19}$$

$$q = (\hat{q} + \hat{R}^2)(1-q)^2 \ , \tag{5.20}$$

$$\hat{R} = \frac{\alpha}{\pi\sqrt{1-q}}\int_{-\infty}^\infty Dt \frac{e^{-v^2/2}}{(e^\beta-1)^{-1} + H(v)} \ , \tag{5.21}$$

$$\hat{q} = \frac{\alpha}{\pi(1-q)}\int_0^\infty Dy \int_{-\infty}^\infty Dt \frac{e^{-u^2}}{[(e^\beta-1)^{-1} + H(u)]^2} \ , \tag{5.22}$$

where u is defined in Eq. (5.18) above, and

$$v \equiv t\left(\frac{q-R^2}{1-q}\right)^{1/2} \tag{5.23}$$

The solution of these equations leads to a learning curve with a $1/\alpha$ tail for all T. Note that this power law is not a consequence of the general $1/\alpha$ law derived in Sec. III, since a network with a Boolean output is not a smooth network. At $T = 0$, the asymptotic learning curve is

$$\epsilon_g = 2\left(\int Dt\, H^{-1}(t/\sqrt{2})\right)^{-1}\frac{1}{\alpha} + O(\alpha^{-2}) \tag{5.24}$$

$$= \frac{0.625}{\alpha} + O(\alpha^{-2}) \ . \tag{5.25}$$

Unlike the previous models, there is no transition at finite α to perfect learning at $T = 0$. In fact, there is no phase transition at any T and α.

The AA for this model is [Eq. (4.14)]

$$G_{\rm an}(R) = -\ln\left(1 - \frac{1-e^{-\beta}}{\pi}\cos^{-1}R\right) \tag{5.26}$$

yielding for the free energy

$$-\beta f = \frac{1}{2}\ln(1-R^2) + \alpha\ln\left(1 - \frac{1-e^{-\beta}}{\pi}\cos^{-1}R\right) . \tag{5.27}$$

Evaluating R by minimizing the free energy, we find that

$$\epsilon_g = \frac{1}{1-e^{-\beta}}\frac{1}{\alpha} + O(\alpha^{-2}). \tag{5.28}$$

This agrees with the correct power law, but does not predict correctly the prefactor; see Eq. (5.24).

Finally, it has been recently shown using the replica method that the above $1/\alpha$ law can be improved by at most a factor of $\sqrt{2}$, using a Bayes optimal learning algorithm for the perceptron [50, 51].

C. Linear output with discrete weights

Imposing binary constraints on the weights of a perceptron with a linear output modifies its learning performance drastically. Let us first consider the zero-temperature behavior. The weight vector must satisfy the same P homogeneous linear equations as before, Eq. (5.7), but now it is constrained to $W_i = \pm 1$. For almost every continuous input $\mathbf{S}$, the equation $(\mathbf{W} - \mathbf{W}^0) \cdot \mathbf{S} = 0$ cannot be satisfied unless $\mathbf{W} = \mathbf{W}^0$. Hence just one example guarantees perfect learning at zero temperature, i.e.,

$$\alpha_c = 0 \ . \tag{5.29}$$

However, this argument does not exclude the existence of local minima of E. In this case the main effect of increasing the number of examples is to smooth the energy surface, in order to make the unique global minimum dynamically accessible for large networks. In order to investigate these aspects, study of the finite-T version of the problem is extremely useful.

1. Annealed approximation

As the phase diagram of the model at low T is rather complex, it is instructive to first analyze the relatively simple AA, which captures most of the qualitative behavior. Using Eqs. (4.14), (4.15), and (4.20), the annealed free energy is

$$-\beta f = -\frac{1}{2}\alpha \ln[1 + 2\beta(1-R)] - \frac{1-R}{2}\ln\frac{1-R}{2} - \frac{1+R}{2}\ln\frac{1+R}{2} \ , \tag{5.30}$$

which is extremized when

$$\tanh^{-1} R = \frac{\alpha\beta}{1 + 2\beta(1-R)} \ . \tag{5.31}$$

For large α, the learning curve

$$\epsilon_g \equiv 1 - R \approx 2e^{-2\alpha/T} \tag{5.32}$$

has an exponential tail. This decay is much faster than the inverse power law of the linear or continuous model (5.10), reflecting the severe constraints on the weight space. Also, the prefactor of α in the exponent diverges as $T \to 0$, signaling the fact that

$$\epsilon_g = 0 \ , \quad T = 0 \tag{5.33}$$

for all $\alpha \geq 0$, in agreement with Eq. (5.29).

Evaluating Eqs. (5.30) and (5.31) one finds that at low T and small α there are two solutions for R. The full AA phase diagram is drawn in Fig. 1 with solid thin lines. The lines marked a and b bound the region where there are two locally stable states. These lines are called *spinodal lines*. The *thermodynamic transition line* in the middle marks the points where the free energies of the two solutions are the same. All three lines terminate at $T = 0.40$, $\alpha = 0.87$. The appearance of two free-energy minima is demonstrated in Fig. 2 where graphs of the free energy for $T = 0.3$ and various values of α are displayed.

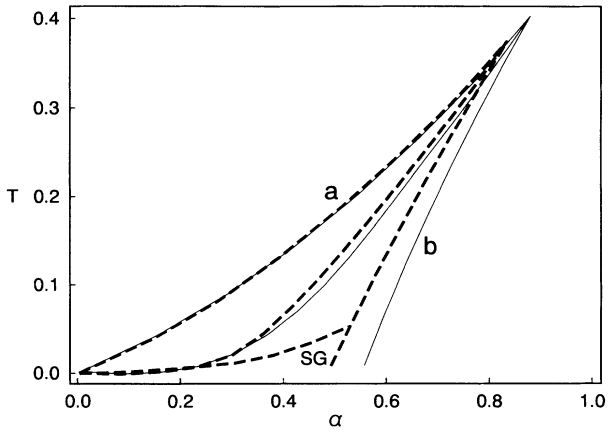

FIG. 1. Phase diagram of the linear-discrete model in the (α, T) plane. The bold dashed lines are from the replica symmetric theory, and the solid thin lines from the annealed approximation. The spinodal lines, marked a and b, demarcate the region where there are two metastable phases. The line running between them is the thermodynamic transition line, at which the two phases have equal free energies. In the RS phase diagram there is a fourth line, running from the origin to spinodal b. This is the RS zero-entropy line of the low-R metastable state.

For $\alpha < 0.730$ there is only a single local minimum. At $\alpha = 0.730$ another local minimum appears at higher R. This is called a ***spinodal point*** of the full spinodal line a in Fig. 1. At $\alpha > 0.756$ the high-R minimum becomes the global minimum of f. However, the local minimum of lower R still exists for $0.756 < \alpha < 0.781$. Only above $\alpha = 0.781$, a second spinodal point, does the low R minimum vanish, leaving only the high-R one.

If the system can be truly equilibrated then the transition to a state with high generalization will occur along the middle line in Fig. 1, which is therefore the thermodynamic transition line. Note that this line starts from

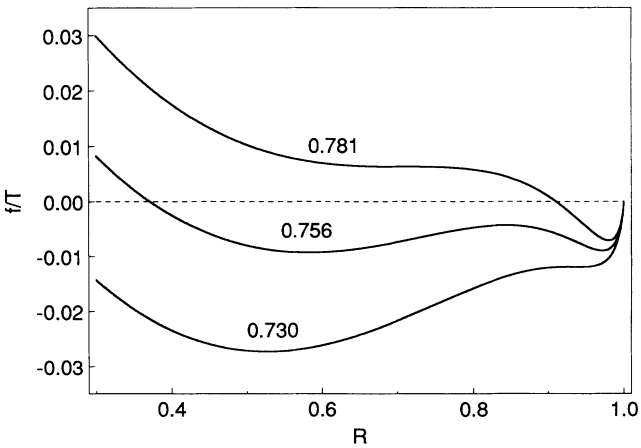

FIG. 2. Annealed free energy of the linear-discrete model as a function of the overlap R at $T = 0.3$ and $\alpha = 0.730$ (spinodal), $\alpha = 0.756$ (thermodynamic transition), and $\alpha = 0.781$ (spinodal).

the origin $\alpha = T = 0$, implying that for any α as $T \to 0$ the equilibrium state is the high-R state. Since in this state $R \to 1$ as $T \to 0$ the equilibrium state at $T \to 0$ is always $R = 1$, in agreement with Eq. (5.29). However, the line approaches the origin as

$$T_c(\alpha) \approx e^{-1/2\alpha} \quad , \quad \alpha \to 0 \ . \tag{5.34}$$

This implies that for a small number of examples even a small noise in the dynamics will generate a transition to the low-R state.

For training in large networks the most important transition is, in general, not the thermodynamic one, but rather the spinodal line b. This is because starting from initially random weights ($R \approx 0$), the system converges quickly to the nearest metastable state, which is the state with low R as long as such a state exists. The time required to cross the free-energy barrier to the thermodynamic high-R phase is prohibitively large, scaling as $t \approx e^{\beta N \Delta f}$ where Δf is the height of the free-energy barrier (per weight) between the two states. It is important to note that, unlike the equilibrium transition line, the spinodal line terminates at $T = 0$ at a finite value of α, $\alpha = 0.556$. This implies that in spite of Eq. (5.29), a finite value of α is required to learn in a finite time. According to the AA the minimal value of α for learning at $T = 0$ in finite time, denoted by α_o, is $\alpha_o = 0.556$.

2. *Replica symmetric theory*

The replica symmetric free energy is given by Eq. (4.25) where G_0, Eq. (4.26), is

$$G_0 = -\frac{1}{2}(1-q)\hat{q} - R\hat{R} + \int Dz \ln 2\cosh(\sqrt{\hat{q}}z + \hat{R}) \ . \tag{5.35}$$

The replicated Hamiltonian G_r, Eq. (4.27), which depends on the error function but not on the weight constraints, remains the same as in Eq. (5.2). The resulting saddle-point equations are

$$R = \int Dz \tanh(\sqrt{\hat{q}}z + \hat{R}) \ , \tag{5.36}$$

$$q = \int Dz \tanh^2(\sqrt{\hat{q}}z + \hat{R}) \ , \tag{5.37}$$

$$\hat{R} = \frac{\alpha\beta}{1 + \beta(1-q)} \ , \tag{5.38}$$

$$\hat{q} = \alpha\beta^2 \frac{q - 2R + 1}{[1 + \beta(1-q)]^2} \ . \tag{5.39}$$

For any fixed temperature, $R \to 1$ and $q \to 1$ as $\alpha \to \infty$. To investigate this approach to the optimum, we note that for large α

$$\hat{q} \sim \alpha\beta^2[2(1-R) - (1-q)] \ , \tag{5.40}$$

$$\hat{R} \sim \alpha\beta \ . \tag{5.41}$$

Clearly there is a divergence of $\hat{R}$, the strength of the ordering term in the local field of Eq. (4.30). At the same time, $\sqrt{\hat{q}}/\hat{R} \to 0$, so that the relative strength of the disordering term is going to zero. This means that the saddle-point equations for q and R, Eqs. (5.36) and (5.37), behave like

$$R \sim \tanh \hat{R} \approx \tanh(\alpha\beta) \ , \tag{5.42}$$

$$q \sim \tanh^2 \hat{R} \approx \tanh^2(\alpha\beta) \ . \tag{5.43}$$

Hence the generalization curve has the same exponential tail $\epsilon_g \approx 2e^{-2\alpha/T}$ as given by the AA in Eq. (5.32).

The RS phase diagram is drawn with bold dashed lines in Fig. 1. The similarity of the RS phase boundaries to those of the AA (thin solid lines) is remarkable. Between the spinodal lines marked a and b, there are two locally stable solutions of the saddle-point equations. The thermodynamic transition line runs between the two spinodals. The line running from the origin to spinodal b is the RS zero entropy line of the low R metastable state.

At $T = 0$ the thermodynamic transition line and the spinodal line b intersect the α axis at

$$\alpha_c = 0 \ , \tag{5.44}$$

$$\alpha_o = 0.48, \tag{5.45}$$

respectively. The result $\alpha_c = 0$ implies that for any $\alpha > 0$, the training energy possesses a unique global minimum $R = 1$. However, the training energy may still possess low-R metastable states. These states vanish above the spinodal point $\alpha_o = 0.48$.

3. *Numerical simulations*

We have used the Metropolis Monte Carlo algorithm to simulate learning in the linear-discrete perceptron. This algorithm is a standard technique for calculating thermal averages over Gibbs distributions [38]. The simulations were performed for multiple samples, i.e., different training sets drawn randomly from a common input distribution. Here the inputs were chosen to be $S_i = \pm 1$ at random, i.e., $\mathbf{S}$ was drawn randomly from the vertices of the N-dimensional hypercube. This discrete input distribution allowed us to take advantage of the speedup offered by integer arithmetic, yet leads to the same learning curves as the Gaussian input distribution (4.3) in the thermodynamic limit (see Appendix C). The quenched average was performed over these samples, and error bars were calculated from the standard error of measurement of the sample distribution. In the figures of this paper, when a Monte Carlo data point lacks an error bar, it means that the error bar would have been smaller than the symbol used to draw that point. In general, fewer samples were required for larger N, because of self-averaging.

In Fig. 3 we present the numerical results as well as the RS theoretical predictions for the training and generalization errors as a function of α. The results of the RS theory are in very good quantitative agreement with Monte Carlo simulations of the model at least for $T \geq 0.2$. At $T = 0.2$ [Fig. 3(a)] the prominent feature is the rapid transition to $R \approx 1$ near $\alpha \approx 0.65$. This is in agreement with the spinodal point $\alpha_o = 0.66$ for this temperature, which can be read from line b in Fig. 1. The location of the thermodynamic transition is shown by a dotted vertical line, and the first spinodal (corresponding to line

a in Fig. 1) is marked by an arrow on the α axis. The roundness of the transition in the simulations is consistent with the expected smearing of the discontinuity in a finite system. At $T = 1.0$ [Fig. 3(b)] the generalization curve decreases smoothly to zero with an exponential tail. The dependence of ϵ_g on N is also shown; the training error did not vary appreciably. It is interesting that the finite-size effects are much more noticeable for the simulations at $T = 1.0$ than at $T = 0.2$.

4. *Spin-glass phase*

The above theoretical results, which were based on the replica symmetric ansatz described by Eqs. (4.22) and (4.23), are not exact, at least for sufficiently low T and α. This is indicated by the fact that the RS entropy of the metastable state with low R becomes negative in the region marked SG in Fig. 1. The zero entropy line is a lower bound for the temperature below which the metastable state must be described by a theory with replica symmetry breaking. The interpretation of the RSB is that for small α the energy surface far away from the optimal overlap $R = 1$ is rough. On the other hand, we expect that the energy surface in the neighborhood of the optimal state is rather smooth. Hence the high-R phase is probably described correctly by the RS theory, and does not exhibit spin-glass properties. This is substantiated by our calculation of the number of local minima in Sec. V C 5. Because of RSB, we expect the true location of the spinodal line at low T and α to differ from the RS results, but this difference, and in particular the value of α_o, may not be large.

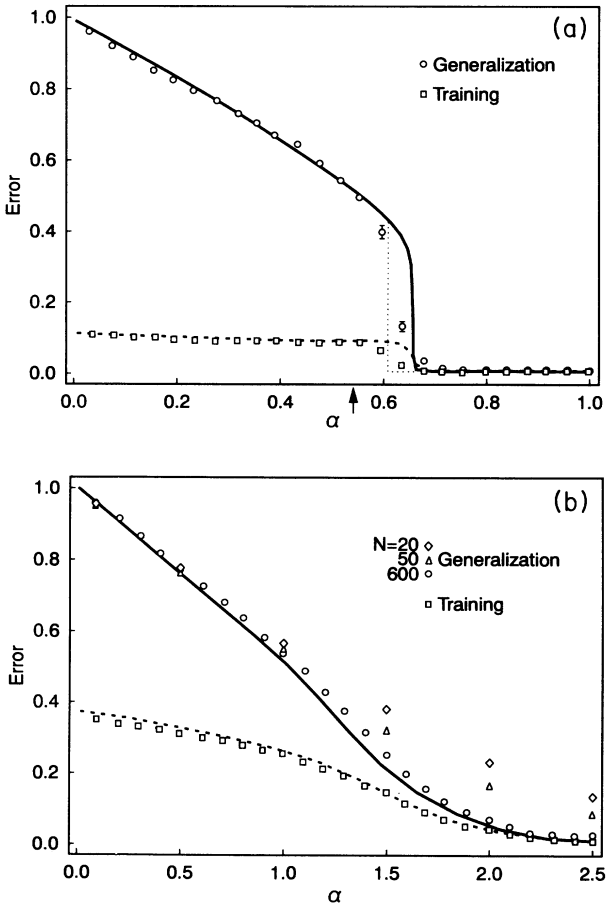

FIG. 3. Learning curves for the linear-discrete model. (a) Monte Carlo simulations at $T = 0.2$ for $N = 100$, averaged over 64 training sets. The generalization and training curves are from the replica symmetric theory, and portray the transition at spinodal b in Fig. 1. The thermodynamic transition is marked by the dotted vertical line, and spinodal a by the arrow on the α axis. (b) Monte Carlo simulations at $T = 1$ for $N = 20$, 50, and 600.

5. *Local minima*

The above finite-T statistical-mechanical results account for the equilibrium state, as well as for metastable states that are separated by barriers that diverge as $N \to \infty$. However, the system may in addition possess states that are local minima of the energy (2.1), but are separated by barriers that remain finite in the $N \to \infty$ limit. Although these states are washed away by thermal fluctuations at any finite temperature, they may dominate the dynamics at $T = 0$. Even at finite low T these barriers may be high enough to prevent equilibration in a reasonable amount of time.

Following Gardner [52], we calculate an upper bound for the number of local minima as a function of the overlap R. Defining the outer product matrix of the examples by

$$T_{ij} = \frac{1}{N}\sum_{l=1}^{P} S_i^l S_j^l \ , \tag{5.46}$$

the energy can be written as

$$E = \frac{1}{2}\sum_{i\neq j=1}^{N} T_{ij} W_i W_j \ - \ \sum_{i,j=1}^{N} T_{ij} W_i W_j^0 \ + \ \frac{1}{2} N\alpha \ , \tag{5.47}$$

with $W_i = \pm 1$. This has the standard form of an Ising Hamiltonian. The condition for a local minimum is

$$h_i W_i > 0 \tag{5.48}$$

for all i, where the local fields h_i are defined by

$$h_i = \sum_{j=1}^{N} T_{ij}(W_j^0 - W_j) + \alpha W_i \ . \tag{5.49}$$

The number of local minima is then

$$\mathcal{N} = \mathrm{Tr}_{\mathbf{W}} \prod_i \Theta(h_i W_i) \ . \tag{5.50}$$

For a typical sample of examples we expect the logarithm of the number of local minima $\ln \mathcal{N}(N,R)$ to be extensive and hence self-averaging. We instead perform the simpler annealed average

$$NF(R,\alpha) \equiv \ln \langle\langle \mathcal{N}(N,R) \rangle\rangle \ , \tag{5.51}$$

which yields an upper bound for the typical number of such states $\langle\langle \ln \mathcal{N}(N,R) \rangle\rangle$. A saddle-point expansion

yields

$$F(R=1,\alpha)=0 \tag{5.52}$$

and (for $R<1$)

$$\begin{aligned}F(R,\alpha)=&\frac{1+R}{2}\ln H(-y)\\&+\frac{1-R}{2}\left(\ln H(x)+\frac{1}{2}(x+y)^2\right)\\&-\frac{\alpha}{2}\left(\frac{x}{y}+\ln\frac{(1-R)y^2}{\alpha}+\ln 2\right)\\&-\frac{1-R}{2}\ln\frac{1-R}{2}-\frac{1+R}{2}\ln\frac{1+R}{2}\ .\end{aligned} \tag{5.53}$$

Here we have defined

$$H(x)\equiv\int_x^\infty Dt\ , \tag{5.54}$$

and F has to be extremized with respect to x and y.

Figure 4 shows graphs of $F(R,\alpha)$ as a function of R for various values of α, obtained by solving the saddle-point equations for x and y numerically. Wherever F is negative, then there are no local minima in the thermodynamic limit, since the annealed average is an upper bound. As seen in the figure, F is negative near $R=1$ for all values of α. In fact, it can be shown that as $R\to 1$, $F\to-\alpha(\frac{1}{2}\ln 2-\frac{1}{4})=-0.0966\alpha$. This implies that there are no local minima in the neighborhood of the optimal $\mathbf{W}$, and the energy surface there is smooth. Note that at $R=1$, $F=0$, i.e., $\mathcal{N}=1$ as expected.

Figure 4 also shows that as α increases the size of the hole in the number of local minima increases, until $\alpha=2.39$, above which F is everywhere negative, so that there are no local minima at all. The implication for learning dynamics is that there exists some $\alpha_c\le 2.39$ above which learning is fast even for $T=0$. This prediction has been confirmed by numerical simulations of the model at $T=0$. We have found that the system converges rapidly to $R=1$ from almost all initial conditions for

$$\alpha\gtrsim 1.0\ . \tag{5.55}$$

D. Boolean output with discrete weights

This Boolean-discrete model, first studied by Gardner and Derrida [23], exhibits a first-order transition from a state of poor learning to a state of perfect learning [37, 53]. Unlike the linear-discrete perceptron discussed above, this model's transition persists at all temperatures. The occurrence of this remarkable transition can be understood using the high-temperature limit.

1. The high-temperature limit

In the high-T limit the energy of the system is given simply as $N\alpha\epsilon(R)$. Hence, using Eq. (4.12) for $\epsilon(R)$, the free energy is simply

$$-\beta f=-\frac{\alpha\beta}{\pi}\cos^{-1}R-\frac{1-R}{2}\ln\frac{1-R}{2}-\frac{1+R}{2}\ln\frac{1+R}{2}\ , \tag{5.56}$$

This free energy is shown in Fig. 5 for various values of α/T. In contrast to previous models, the state $R=1$ is a local minimum of f for all values of T and α. For small values of α/T the state $R=1$ is only a local minimum of f, as can be seen in Fig. 5. The global minimum is given by the solution of the saddle-point equation

$$R=\tanh\left(\frac{\alpha\beta}{\pi\sqrt{1-R^2}}\right)\ . \tag{5.57}$$

This state of poor learning $R<1$ is the equilibrium state for

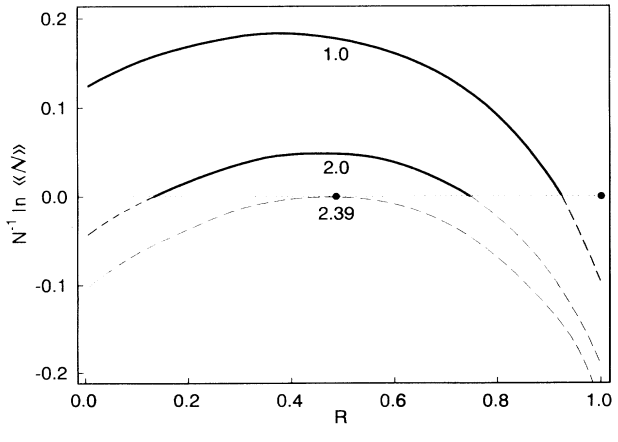

FIG. 4. Annealed upper bound for the logarithm of the density of local minima in the linear-discrete model as a function of R. For $\alpha=1.0$ there is a small gap around $R=1.0$. As α increases, the density of local minima decreases, until above $\alpha=2.39$ there are no local minima at all, except the isolated minimum at $R=1$ (marked with a solid dot).

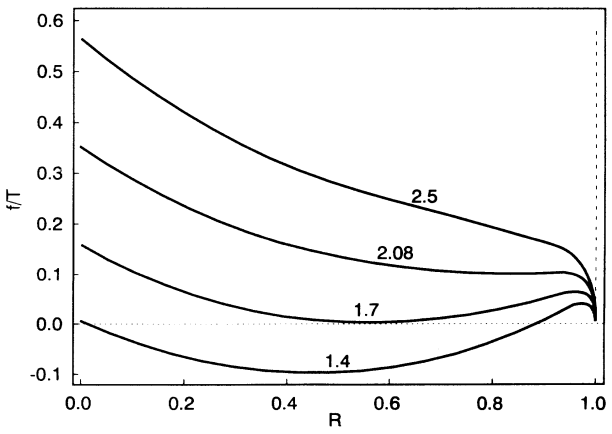

FIG. 5. High temperature limit of the free energy βf of the Boolean-discrete model as a function of the overlap R for $\alpha/T=1.4$, 1.7 (thermodynamic transition), 2.08 (spinodal), and 2.5. The vertical dashed line at $R=1$ marks the upper bound of the allowed range of R.

$$T > 0.59\alpha \ . \tag{5.58}$$

In this regime the optimal state $R = 1$ is only metastable. If the initial network has R which is close to 1 the learning dynamics will converge fast to the state $R = 1$. However starting from a random initial weight vector $R \approx 0$ the system will not converge to the optimal state.

For $T/\alpha < 0.59$ the equilibrium state is $R = 1$, although there is still a local minimum, i.e., a solution of Eq. (5.57) with $R < 1$. Finally for

$$T < 0.48\alpha \ , \tag{5.59}$$

there is no solution with $R < 1$ to Eq. (5.57). In this regime (beyond the spinodal), starting from any initial condition the system converges fast to the optimal state.

The collapse of the system to the energy ground state at finite temperature is an unusual phenomenon. The origin of this behavior is the square-root singularity of $\epsilon(R)$ at $R = 1$. This singularity implies that a state characterized by $\delta R \equiv 1 - R \ll 1$ has an energy which is proportional to

$$E \propto N\sqrt{\delta R} \ . \tag{5.60}$$

This big increase in energy cannot be offset by the gain in entropy, which is proportional to

$$\delta N G_0(R) \propto N(\delta R)\ln(\delta R) \ . \tag{5.61}$$

This effect can be nicely seen using the microcanonical description. According to Eq. (2.27) above, a smooth low-temperature limit exists provided that

$$\lim_{\epsilon \to \epsilon_{\min}} \frac{\partial s(\epsilon)}{\partial \epsilon} = \infty \ . \tag{5.62}$$

On the other hand, Eq. (4.20) implies that in the present case

$$\frac{\partial s(\epsilon)}{\partial \epsilon} \approx -\epsilon \ln \epsilon, \quad \epsilon \to 0 \ . \tag{5.63}$$

Thus the rate of increase in entropy is too small to give rise to thermal fluctuations below some critical temperature.

It is instructive to apply the above argument to the case of states that differ from the ground state by a flip of a single weight. According to Eq. (5.60) the energy of such states is

$$E \propto \sqrt{N} \ , \tag{5.64}$$

whereas the entropy associated with such an excitation is only

$$\delta N G_0(R) \propto \ln N \ . \tag{5.65}$$

It should be emphasized, however, that examining the spectrum of the first excitations is not generally sufficient for determining the thermodynamic behavior at any finite T, where the relevant states are those with energy of order NT.

2. *Replica symmetric theory*

Because of the unusual features of the transition in this model we will analyze the quenched theory in some detail. We first study the replica symmetric theory and then investigate the replica symmetry breaking in this system.

The RS free energy is given by combining G_0 of the perceptron with discrete weights, Eq. (5.35), with G_r for a Boolean output, Eq. (5.17), yielding

$$-\beta f = -\frac{1}{2}(1-q)\hat{q} - R\hat{R} + \int Dz \ln 2\cosh(\sqrt{\hat{q}}z + \hat{R}) + 2\alpha \int_0^\infty Dy \int_{-\infty}^\infty Dt \ln\left[e^{-\beta} + (1 - e^{-\beta})H(u)\right] \ , \tag{5.66}$$

where the function $H(x)$ is as defined in (5.54). The saddle-point equations are

$$R = \int Dz \tanh(\sqrt{\hat{q}}z + \hat{R}) \ , \tag{5.67}$$

$$q = \int Dz \tanh^2(\sqrt{\hat{q}}z + \hat{R}) \ , \tag{5.68}$$

$$\hat{R} = \frac{\alpha}{\pi\sqrt{1-q}} \int_{-\infty}^\infty Dt \frac{e^{-v^2/2}}{(e^\beta - 1)^{-1} + H(v)} \ , \tag{5.69}$$

$$\hat{q} = \frac{\alpha}{\pi(1-q)} \int_0^\infty Dy \int_{-\infty}^\infty Dt \frac{e^{-u^2}}{[(e^\beta - 1)^{-1} + H(u)]^2} \ , \tag{5.70}$$

where u and v are given by Eqs. (5.18) and (5.23) above.

At $T = 0$, the equations simplify somewhat, since $q = R$ and $\hat{q} = \hat{R}$. For α less than

$$\alpha_c = 1.245 \ , \tag{5.71}$$

there are two solutions, one with $R = 1$ (perfect generalization), and one with $R < 1$ (poor generalization). The $R < 1$ saddle point has the lower free energy, and is therefore the equilibrium phase. Upon crossing this critical α, the balance of free energy shifts, and there is a first-order transition to the $R = 1$ state. Hence at $\alpha_c = 1.245$ there is a discontinuity in the generalization curve, a sudden transition to perfect learning. However, the $R < 1$ state still remains as a metastable phase until the spinodal point

$$\alpha_o = 1.492 \ . \tag{5.72}$$

At any fixed T, there is the same sequence of thermodynamic transition followed by spinodal transition with increasing α. The RS phase diagram is shown in Fig. 6. To the left of the dashed thermodynamic transition line, the state of poor generalization $R < 1$ is the equilibrium state, and the state of perfect generalization $R = 1$ is metastable. In the region between the dashed line and the solid spinodal line, the situation reverses, with $R = 1$ becoming the equilibrium state and $R < 1$ the metastable state. To the right of the spinodal line, there is no low R metastable state.

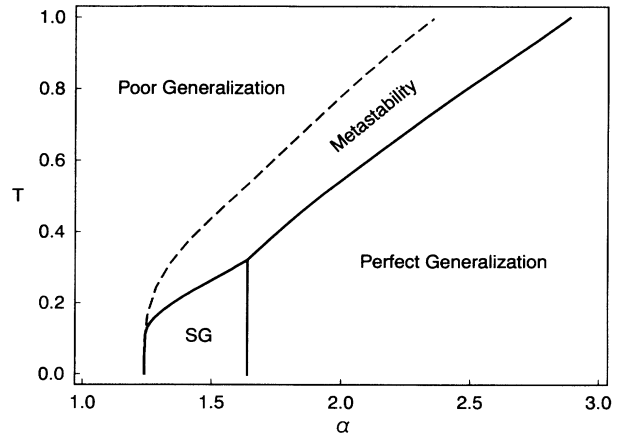

FIG. 6. RS and one-step RSB phase diagram for the Boolean-discrete model. To the left of the dashed thermodynamic transition line, the equilibrium state has $R < 1$ (poor generalization), while the perfect generalization state ($R = 1$) is metastable. Between the dashed line and the solid spinodal line the $R = 1$ state is the equilibrium one, while the $R < 1$ state is metastable. To the right of the spinodal line, there is no $R < 1$ phase. In the region marked "SG," the one-step RSB calculation predicts a metastable spin-glass state.

3. *Replica symmetry breaking and metastable spin-glass phase*

The line $T_g(\alpha)$ in the T-α plane, where the RS "poor generalization" state ($R < 1$) has zero entropy, is the upper border of the region marked SG in Fig. 6. It should be noted that this line $T_g(\alpha)$ is to the right of the thermodynamic transition line and coincides with it only at $T = 0$ (at $\alpha = 1.24$). This implies that the RS theory is invalid for this metastable phase only. To find the correct metastable state (with poor generalization) at low T we must search for saddle-point solutions to the replica theory that break the replica symmetry.

To gain a better understanding of this metastable state, we have studied a solution to the replica mean-field theory with a one-step replica symmetry breaking ansatz. Our study is based on the work of Krauth and Mézard [33] concerning the problem of loading *random* dichotomies on a perceptron with discrete weights. The formal derivation of the mean-field equations is presented in Appendix E. Here we present the main results and their physical meaning.

In the one-step ansatz of replica symmetry breaking, the $n \times n$ order parameter matrix $Q_{\mu\nu}$ acquires two off-diagonal values q_0 and q_1 arranged in a block structure of $m \times m$ submatrices (see Appendix E). This block structure reflects the existence of many, almost degenerate, spin-glass (SG) states [45]. These states are valleys in the free-energy surface that are separated by barriers that diverge with the system size [43]. The parameter q_1 represents the overlap of each state with itself, i.e., it is the order parameter q, Eq. (2.56), measured within a single valley denoted a,

$$q_1 = N^{-1} \langle\!\langle \langle \mathbf{W}_a \rangle_T^2 \rangle\!\rangle \ . \tag{5.73}$$

The parameter q_0 represents the average overlap between a pair of two different states a and b, i.e.,

$$q_0 = N^{-1} \langle\!\langle \langle \mathbf{W}_a \rangle_T \cdot \langle \mathbf{W}_b \rangle_T \rangle\!\rangle \ . \tag{5.74}$$

When the $n \to 0$ limit is taken, the size m of the step in $q(x)$ (see Appendix E) must be determined variationally, like q_0 and q_1. Denoting the Gibbs probabilities of the different states by $P_a = \exp(-\beta F_a)$, it can be shown [45] that

$$m = 1 - \sum_a P_a^2 \ . \tag{5.75}$$

Hence m is the probability of finding two copies of the system in two different states.

In the present model, a one-step RSB solution exists in the regime marked SG in Fig. 6, below the line $T_g(\alpha)$. This SG phase is special in that $q_1 = 1$, independent of both T and α. From Eq. (5.73) it follows that each of the different valleys is completely frozen, i.e., there are no fluctuations in $\mathbf{W}$ within each valley. Indeed, the entropy of this phase is zero (to linear order in N). For fixed α, the order parameter q_0 does not vary with temperature within the SG phase; it is frozen at its value on the phase boundary $T_g(\alpha)$. The same holds true for ϵ_t and R, i.e.,

$$q_0(T, \alpha) = q_0(T_g(\alpha), \alpha) \ , \tag{5.76}$$

$$\epsilon_t(T, \alpha) = \epsilon_t(T_g(\alpha), \alpha) \ , \tag{5.77}$$

$$R(T, \alpha) = R(T_g(\alpha), \alpha) \ , \tag{5.78}$$

everywhere in the SG phase. The values of $\epsilon_t(\alpha)$ and $\epsilon_g(\alpha)$ in this phase are shown in Fig. 7. The parameter m is linear in T, $m = T/T_g(\alpha)$. Near the transition temperature T_g, the degeneracy is very severe so that $m \approx 1$. At $T = 0$ the degeneracy is broken, and the Gibbs weight concentrates in the SG metastable state with minimal energy, resulting in $m = 0$.

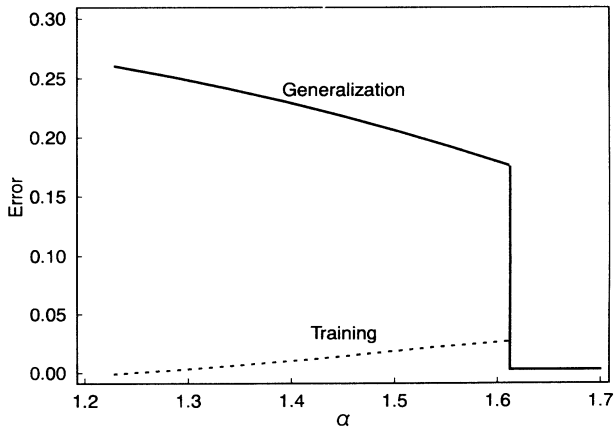

FIG. 7. Training and generalization errors of metastable spin glass phase in the Boolean-discrete model, according to the one-step RSB ansatz. In this phase, the errors are independent of temperature, and are given by their RS values on the zero-entropy line $T_g(\alpha)$ of Fig. 6.

A similar frozen SG phase exists in the perceptron model of Krauth and Mézard [33], and was first discovered by Derrida in the random-energy model [54]. However, unlike these other examples, the SG phase in the present model is only a metastable phase, as evidenced by the nonzero ϵ_t in Fig. 7. Hence the one-step RSB theory does not alter the RS prediction that the zero temperature thermodynamic transition is at $\alpha = 1.245$. However, it raises to

$$\alpha_o = 1.628 \tag{5.79}$$

the spinodal line for the vanishing of the metastable SG phase in Fig. 6.

4. *Numerical simulations*

Figure 8 shows learning curves for Monte Carlo simulations with $N = 75$, averaged over 32 samples. At $T = 1$, there is quite a good fit if the transition is taken to occur at the spinodal line. At lower temperatures, the Metropolis algorithm tends to become trapped in local minima, making equilibration difficult. Hence we cannot use it to check the $T = 0$ predictions of the quenched theory.

The Metropolis algorithm produces a random walk through weight space that samples the Gibbs distribution. For small system sizes, we do not have to sample the weight space; we can explore it exhaustively. Gardner and Derrida [23] applied this idea to compute α_c at $T = 0$. Starting with all 2^N possible student vectors, an example is chosen at random, and all student vectors that "disagree" with the teacher are eliminated. Eventually some number of examples P is reached such that the addition of one more example produces perfect learning. Then $\alpha_c(N) = P/N$ for this sample. The procedure is then repeated with a different set of examples, so that α_c can be sample averaged.

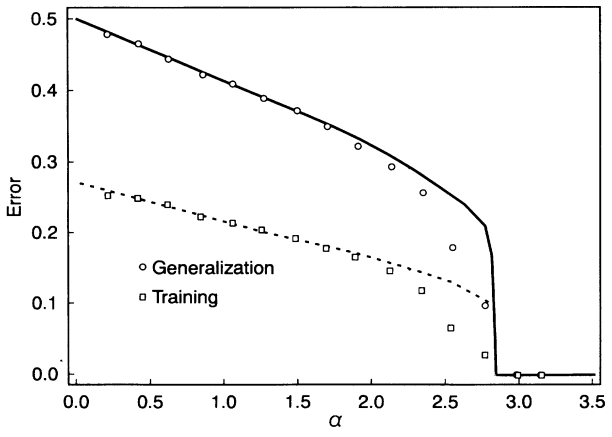

FIG. 8. Learning curves for the Boolean-discrete model. Monte Carlo simulations at $T = 1.0$ and $N = 75$, averaged over 32 samples. The solid line and dashed lines are the RS generalization and training curves for the spinodal transition. The thermodynamic transition is marked by the dotted vertical line.

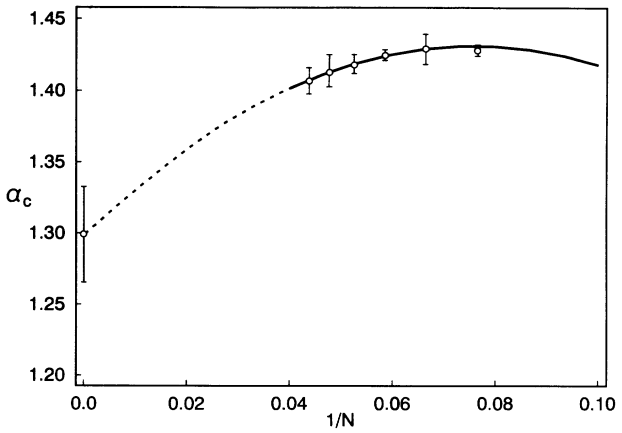

FIG. 9. Graph of α_c vs $1/N$ at $T = 0$ for the Boolean-discrete model from exhaustive search. Each data point is the average of roughly 10^5 or 10^6 random samples of examples. The error bars are the standard error of measurement over the sample distribution. A least-squares fit to a quadratic function yields the extrapolation $\alpha_c(N = \infty) = 1.30 \pm 0.03$. The replica symmetric prediction $\alpha_c = 1.24$ is also marked.

We have extended the results of Gardner and Derrida up to the size $N = 23$. With the discrete inputs $S_i = \pm 1$, only odd N were used, so as to avoid the situation $\mathbf{W} \cdot \mathbf{S} = 0$. Figure 9 exhibits α_c as a function of $1/N$, with each data point an average of from 10^5 (large N) to 10^6 (small N) samples. Fitting to a quadratic function and extrapolating to $N = \infty$ yields the estimate $\alpha_c = 1.30 \pm 0.03$, which is fairly close to the prediction 1.245 of the quenched theory. It is possible that the quadratic function assumed here may be a poor approximation to the true finite-size scaling. This would account for the remaining disagreement between the numerics and the theory. See Ref. [55] for more simulations and an attempt at addressing the question of finite-size scaling.

VI. PERCEPTRON LEARNING OF UNREALIZABLE RULES

So far we have studied examples of perceptron learning of realizable rules. In this section we study three rules that are unrealizable because of architectural mismatch. In the first model, the student and teacher perceptrons are mismatched due to their different transfer functions. In the second and third models, the teacher's weight vector is not included in the weight space of the student.

A. Linear-continuous perceptron with unrealizable threshold

In this model, the perceptrons of the student and teacher both have linear outputs and continuous weights, as in the model of Sec. V A. However, in the present case the transfer function of the teacher has a threshold, i.e., the rule is given by Eq. (4.5) with a transfer function $g(x) = g_0(x)$, where

$$g_0(x) = x + \theta \ . \tag{6.1}$$

The output of the trained network is given by Eq. (4.1) with

$$g(x) = x \ . \tag{6.2}$$

The realizable linear-continuous model of Sec. V A corresponds to the case $\theta = 0$.

The average of the error function

$$\epsilon(\mathbf{W};\mathbf{S}) = \frac{1}{2N}[(\mathbf{W}-\mathbf{W}_0)\cdot\mathbf{S} - \sqrt{N}\theta]^2 \tag{6.3}$$

over the input distribution, Eq. (4.3), yields the generalization error

$$\epsilon(\mathbf{W}) = 1 - R + \frac{1}{2}\theta^2 \ . \tag{6.4}$$

At $R = 1$, the generalization error takes on its minimum value $\epsilon_{\min} = \theta^2/2$. Recall that according to Eq. (2.58), the average training and generalization errors both approach $\epsilon_{\min}$ as $\alpha \to \infty$, at all T.

As in the linear-continuous model, the $T = 0$ behavior can be understood from geometric arguments. Weight vectors with zero training error must satisfy P linear equations

$$(\mathbf{W} - \mathbf{W}^0)\cdot\mathbf{S}^l = \sqrt{N}\theta \ , \quad l = 1 \ldots P \tag{6.5}$$

and the spherical constraint $\mathbf{W}\cdot\mathbf{W} = N$. There exists an $\alpha_c < 1$ such that for $\alpha \le \alpha_c$, these equations have solutions. For $\alpha > \alpha_c$, the equations have no solution, so that the training error rises from zero, and asymptotically approaches $\epsilon_{\min}$ as $\alpha \to \infty$.

The RS free energy is

$$-\beta f = -\beta f_{\theta=0} - \frac{1}{2}\alpha\frac{\beta\theta^2}{1+\beta(1-q)} \ , \tag{6.6}$$

where $f_{\theta=0}$ is the RS free energy of the realizable case; see Eqs. (5.1) and (5.2). The saddle-point equations are the same as (5.3)–(5.6), except that Eq. (5.6) for $\hat{q}$ must be modified to

$$\hat{q} = \alpha\beta^2\frac{q - 2R + 1 + \theta^2}{[1+\beta(1-q)]^2} \ . \tag{6.7}$$

For fixed temperature and $\alpha \to \infty$, we find the asymptotic learning curves

$$\epsilon_g = \epsilon_{\min} + \frac{T+\theta^2}{2\alpha} + O(\alpha^{-2}), \tag{6.8}$$

$$\epsilon_t = \epsilon_{\min} + \frac{T-\theta^2}{2\alpha} + O(\alpha^{-2}). \tag{6.9}$$

To compare with the general results Eqs. (3.12) and (3.14) for smooth networks, we note that in the present case $\partial_i\epsilon(\mathbf{W}^0,\mathbf{S}) = -N^{-1/2}\theta S_i$ and $\partial_i\partial_j\epsilon(\mathbf{W}^0,\mathbf{S}) = N^{-1}S_iS_j$, so that the matrices defined in Eqs. (3.4) and (3.5) are

$$U_{ij} = N^{-1}\delta_{ij} \ , \tag{6.10}$$

$$V_{ij} = N^{-1}\theta^2\delta_{ij} \ . \tag{6.11}$$

Hence the coefficient $N^{-1}\,\mathrm{Tr}\,VU^{-1}$ in Eqs. (3.12) and (3.14) equals θ^2, in agreement with Eq. (6.8) above.

In the $T \to 0$ limit there is a critical α below which the examples can be loaded with zero training error. This value is

$$\alpha_c = \frac{2+\theta^2-\theta\sqrt{2+\theta^2}}{2} \ . \tag{6.12}$$

For $\alpha < \alpha_c$, q is less than 1 and the values of the order parameters are given by the saddle-point equations

$$R = \alpha \ , \tag{6.13}$$

$$q = \alpha + \frac{\alpha\theta^2}{1-\alpha} \ . \tag{6.14}$$

In this regime the minimum of ϵ_t is highly degenerate with an extensive zero-temperature entropy. Note that for any threshold θ, the critical α satisfies the bounds $1/2 < \alpha_c < 1$. For $\alpha > \alpha_c$, $q \to 1$ as $T \to 0$, with $\beta(1-q)$ approaching a finite value. Thus at $T = 0$ the order parameters are given above α_c by

$$2R^3 = (2+\alpha+\theta^2)R^2 - \alpha \ , \tag{6.15}$$

$$\beta(1-q) = \frac{R}{\alpha - R} \ . \tag{6.16}$$

The zero temperature α_c of this and other unrealizable models is similar to the α_c defined for the problem of loading random patterns onto a network. In both cases α_c is the limit of storage capacity above which no weight vectors can achieve zero training error. This is in contrast to the α_c that we defined for realizable models, above which there is exactly one weight vector which can achieve zero training error.

As in the zero-threshold linear-continuous model, we expect that in this linear model the RS theory is exact for all T and α. Furthermore, the learning dynamics should be quick, since there are no spurious local minima in the training energy.

In Sec. II D we noted that the AA may yield the wrong $\alpha \to \infty$ limit, for unrealizable, non-Boolean rules. The present model is a simple example of this phenomenon. In the present case, G_{an}, Eq. (4.14), is

$$G_{\mathrm{an}}(R) = \frac{1}{2}\ln[1+2\beta(1-R)] + \frac{1}{2}\frac{\beta\theta^2}{1+2\beta(1-R)} \ . \tag{6.17}$$

The resulting annealed free energy, Eq. (4.13), is

$$-\beta f(R) = \frac{1}{2}\ln(1-R^2) - \frac{\alpha}{2}\ln[1+2\beta(1-R)] - \frac{\alpha}{2}\frac{\beta\theta^2}{1+2\beta(1-R)} \ . \tag{6.18}$$

In the $\alpha \to \infty$ limit R is determined by minimizing G_{an}, yielding

$$R = \begin{cases} 1, & \theta^2 \le T \\ 1 - \frac{1}{2}(\theta^2 - T), & T < \theta^2 < T+4 \\ -1, & T+4 \le \theta^2. \end{cases} \tag{6.19}$$

According to these results, when $\theta^2 > 4$ and $T = 0$, the weight vector $\mathbf{W}$ approaches $-\mathbf{W}^0$ as $\alpha \to \infty$.

To understand the origin of this gross failure of the annealed approximation we recall from Sec. II D that the annealed approximation can be viewed as the exact theory for a stochastic training dynamics in both weight and input space, leading to the Gibbs distribution (2.43) in both $\mathbf{W}$ and $\mathbf{S}$ space. The magnitude of the resulting distortion in the posterior distribution of the inputs relative to their *a priori* one determines the quality of the approximation. In the present case we can obtain a measure of this distortion by calculating the average of an input vector, e.g., $\langle \mathbf{S}^1 \rangle_{\rm an}$. In the *a priori* Gaussian distribution of inputs (4.3), the average value of an input is of course zero. On the other hand, evaluating this average using the posterior Gibbs distribution (2.43), implied by the AA, we find

$$\langle \mathbf{S} \rangle = -\frac{1}{\sqrt{N}} \frac{\beta\theta(1-R)}{1+2\beta(1-R)} \mathbf{W}^0 , \tag{6.20}$$

implying that the inputs are biased towards $-\theta\mathbf{W}^0$. It is important to note that the magnitude of the bias is small—down by $\sqrt{N}$ from the magnitude of $\mathbf{S}$. However, it is enough to push $\mathbf{W}$ towards $-\mathbf{W}^0$, leading to overlaps as low as $R = -1$. Finally, note that in the realizable case $\theta = 0$, the average of $\mathbf{S}$ is zero. However, even in this case the second moments of $\mathbf{S}$ are in general distorted, except when $R = 1$. This is consistent with the general expectation that the distortions of the input distribution, implied by the AA, are relatively small in realizable rules. But they vanish only in the limit of $\alpha \to \infty$.

B. Linear output with weight mismatch

In this model, the unrealizability is due to a mismatch in weight space between teacher and student. We assume that the weights of the teacher network $\mathbf{W}^0$ are real valued whereas the trained network is restricted to $W_i = \pm 1$. For simplicity we consider here the case where the individual teacher weights W_i^0 are drawn from a continuous Gaussian distribution $P(W_i^0)$,

$$P(W_i^0) = (2\pi)^{-1/2} e^{-(1/2)(W_i^0)^2} , \tag{6.21}$$

As in the previous linear perceptron models the error function is given by Eq. (4.9), and the generalization error for a given network by Eq. (4.10). In the present model the optimal weights for the restricted architecture of the trained network are

$$W_i^* = \mathrm{sgn}(W_i^0) , \tag{6.22}$$

which corresponds to the maximal overlap

$$R_\infty = \frac{1}{N} \sum_i |W_i^0| = \int dW^0 \, P(W^0) \, |W^0| = \sqrt{2/\pi} . \tag{6.23}$$

The second equality holds in the thermodynamic limit. The minimal generalization error, achieved in the limit $\alpha \to \infty$, is

$$\epsilon_{\rm min} = 1 - R_\infty = 0.202 . \tag{6.24}$$

Before we present the replica solution we note that both the high-T and the annealed approximations predict for this model $\epsilon_g(\alpha) - \epsilon_{\rm min} \propto \alpha^{-2}$, $\alpha \to \infty$. As we shall see below, this is not the correct asymptotic behavior.

1. Replica symmetric theory

Here the replicated Hamiltonian G_r is that of the previous linear models, Eq. (5.2). From Eq. (4.26) G_0 may be calculated for the case of mismatched weight spaces. Making the change of variables $\hat{q} + \hat{R}^2 \to \hat{q}$, and then eliminating $\hat{R}$ altogether, one obtains the free energy

$$-\beta f = -\frac{1}{2}(1-q)\hat{q} - \frac{1}{2}\frac{R^2}{1-q} + \int Dz \ln 2\cosh(\sqrt{\hat{q}}z) - \frac{\alpha}{2}\ln[1+\beta(1-q)] - \frac{\alpha}{2}\frac{\beta(q-2R+1)}{1+\beta(1-q)} . \tag{6.25}$$

The saddle-point equations are

$$\hat{q} = \frac{R^2}{(1-q)^2} + \alpha\beta^2 \frac{q-2R+1}{[1+\beta(1-q)]^2} , \tag{6.26}$$

$$R = \frac{\alpha\beta(1-q)}{1+\beta(1-q)} , \tag{6.27}$$

$$q = \int Dz \tanh^2(\sqrt{\hat{q}}z) . \tag{6.28}$$

To obtain the asymptotic form of the learning curves we have expanded the solutions in powers of α^{-1} keeping T fixed. We find

$$\epsilon_g = \epsilon_{\rm min} + \frac{\epsilon_{\rm min} R_\infty}{\alpha} + R_\infty \left[-\frac{3}{2} - \frac{3}{\pi^3} - \frac{5}{\pi} + 4R_\infty + \frac{\pi^2}{24}\left(T - \frac{6R_\infty}{\pi^2}\right)^2 \right] \frac{1}{\alpha^2} + O(\alpha^{-3}) , \tag{6.29}$$

$$\epsilon_t = \epsilon_{\rm min} - \frac{\epsilon_{\rm min} R_\infty}{\alpha} + R_\infty \left(\frac{1}{2} + \frac{1}{\pi} - R_\infty + \frac{\pi^2}{24} T^2 \right) \frac{1}{\alpha^2} + O(\alpha^{-3}) . \tag{6.30}$$

Thus for any fixed T the leading behavior of the errors is an α^{-1} power law. This power law is not a consequence of the general results (3.12) and (3.14), since the present network is not smooth. Note that in the high-T limit, only the term $T^2\alpha^{-2}$ survives. Thus for this unrealizable rule the behavior predicted by the high-T limit (as well as by the annealed approximation) does not reflect the correct behavior at large α for any fixed T.

In the $T \to 0$ limit with $1 - q$ finite the saddle-point equations reduce to

$$R = \alpha , \tag{6.31}$$

$$\hat{q}(1-q)^2 = -\alpha^2 + \alpha(1+q) , \tag{6.32}$$

$$1-q=\int Dz\,\mathrm{sech}^2(\sqrt{\hat{q}}z)\ . \tag{6.33}$$

Criticality is reached when $\hat{q}\to\infty$, and $1-q\to 0$, which happens at

$$\alpha_c=1-\left(1-\frac{2}{\pi}\right)^{1/2}=0.397\ . \tag{6.34}$$

Above α_c, the limit $\beta\to\infty$ must be taken with $\beta(1-q)$ constant, yielding

$$0=2R^3-(2+\alpha)R^2+\frac{2\alpha}{\pi}\ , \tag{6.35}$$

$$\beta(1-q)=\frac{R}{\alpha-R}\ . \tag{6.36}$$

At α_c, $\beta(1-q)$ diverges, indicating that the finite $1-q$ solution takes over.

2. Optimal temperature

We define the *optimal temperature* $T_{\rm opt}(\alpha)$ as the temperature that minimizes $\epsilon_g(T,\alpha)$. For our models of realizable rules, the optimal generalization error was at $T=0$ for all values of α. But in general, $T_{\rm opt}$ may be greater than zero, although the convexity of the free energy guarantees that the *training* error is a nondecreasing function of T. This is the case in the present model, as can be seen from Eq. (6.29), which implies that

$$T_{\rm opt}(\alpha)\to\frac{6R_\infty}{\pi^2}=0.485 \tag{6.37}$$

as $\alpha\to\infty$. Note, however, that the leading term of Eq. (6.29) is independent of T, implying that in the present model the effect on ϵ_g of optimizing with respect to T is relatively small.

3. Replica symmetry breaking

The above RS theory is not exact at low T. First, the prediction of a finite value of α_c is probably incorrect. As in the corresponding realizable discrete model of Sec. V B, we expect that

$$\alpha_c=0\ , \tag{6.38}$$

i.e., there is no vector of discrete weights that satisfies order N real, random linear equations. Second, the entropy of the RS solution becomes negative at low T, as shown in Fig. 10. The asymptotic form of the zero-entropy line can be calculated by expanding the entropy in powers of α^{-1},

$$s=\frac{\pi}{6R_\infty}\left(T-\frac{3R_\infty}{\pi^2}\right)\frac{1}{\alpha}+O(\alpha^{-2})\ . \tag{6.39}$$

The $s=0$ temperature approaches the finite limit

$$T_{s=0}(\alpha)\to\frac{3R_\infty}{\pi^2}=0.242 \tag{6.40}$$

as $\alpha\to\infty$. The full line is drawn in Fig. 10 for all α. This

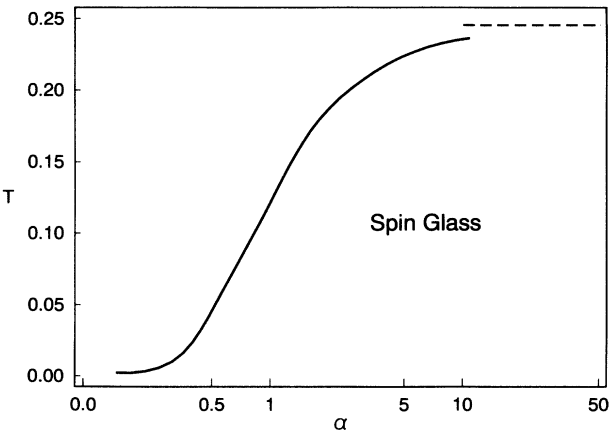

FIG. 10. The RS zero-entropy line for the linear-mismatched model, which is a lower bound for the temperature at which RSB occurs. The line approaches the limit $T=0.242$ (dashed line) as $\alpha\to\infty$.

temperature gives a lower bound on the temperature at which replica symmetry is broken. Below this temperature, a spin-glass phase with replica symmetry breaking occurs. As $T\to 0$, the zero-entropy line approaches the origin, i.e., the RS entropy is negative for all α at $T=0$. This means that the RS solution is incorrect at $T=0$ for any α, and in particular that the RS prediction (6.34) for α_c is incorrect, which is consistent with our prediction, Eq. (6.38).

Below $T=0.242$, the system never escapes from the spin-glass phase even as $\alpha\to\infty$ (see Fig. 10). Note, however, that the entropy (6.39) approaches zero from below. This suggests that the effects of RSB become less severe as $\alpha\to\infty$.

4. Numerical simulations

Figure 11 exhibits Monte Carlo simulations at $T=0.5$ and 0.1 with $N=100$ and 64 samples. Even at $T=0.1$ [Fig. 11(b)] the fit to the RS theory is very good, even though for $\alpha>0.76$ the curves are in the RSB region. This suggests that the effects of RSB in this system are weak. The RS $T=0$ learning curves are also plotted for comparison. Note that above $\alpha>0.6$, the generalization error for $T=0.1$ is less than that at $T=0$. At least in this range, we may thus conclude that the optimal generalization temperature $T_{\rm opt}>0$, assuming that the RS theory is a good approximation for the true behavior in the RSB region.

C. Boolean output with weight mismatch

As in the previous model, the teacher weights are again drawn from a continuous Gaussian distribution, whereas the student weights are constrained to ± 1. However, here we consider the case where the perceptron transfer functions of both the teacher and the student are Boolean. The maximal overlap is still $R_\infty=\sqrt{2/\pi}$, but the optimal generalization error is now

$$\epsilon_{\min} = \frac{1}{\pi}\cos^{-1} R_\infty = 0.206 \ . \tag{6.41}$$

As in the previous model, both the high-T and the annealed approximations predict $\epsilon_g(\alpha) - \epsilon_{\min} \propto \alpha^{-2}$, $\alpha \to \infty$. We shall see below that the true asymptotics are quite different.

1. Replica symmetric theory

For this model, G_0 is the same as that of the previous section, and G_r is that of the previous Boolean models Eq. (5.17). Hence the RS free energy is given by

$$-\beta f = -\frac{1}{2}(1-q)\hat{q} - \frac{1}{2}\frac{R^2}{1-q} + \int_{-\infty}^{\infty} Dz \ln 2\cosh(\sqrt{\hat{q}}z) + 2\alpha \int_0^\infty Dy \int_{-\infty}^{\infty} Dt \ln\left[e^{-\beta} + (1-e^{-\beta})H(u)\right] \ , \tag{6.42}$$

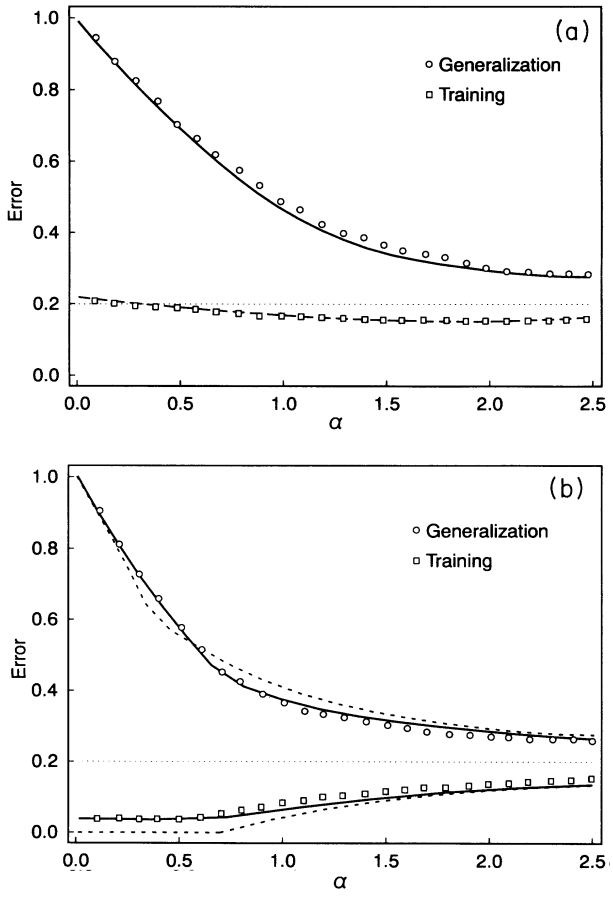

FIG. 11. Learning curves for the linear-mismatched model. (a) Monte Carlo simulations at $T = 0.5$ with $N = 100$ and 64 samples, with lines from RS theory. The dotted horizontal line is the asymptotic error $\epsilon_{\min}$. (b) Simulations at $T = 0.1$. Note that the RS curves (solid lines) fit the data even beyond the zero-entropy point $\alpha = 0.76$, where there is RSB. The dashed line is the $T = 0$ RS curve, shown for comparison.

where the function $H(x)$ is as defined in (5.54) above. The saddle-point equations are

$$\frac{R}{\sqrt{1-q}} = \frac{\alpha}{\pi}\int_{-\infty}^{\infty} Dt \frac{e^{-v^2/2}}{(e^\beta - 1)^{-1} + H(v)} \ , \tag{6.43}$$

$$\hat{q} = \frac{R^2}{(1-q)^2} + \int_0^\infty Dy \int_{-\infty}^{\infty} Dt \frac{e^{-u^2}}{[(e^\beta - 1)^{-1} + H(u)]^2} \ , \tag{6.44}$$

$$q = \int Dz \tanh^2(\sqrt{\hat{q}}z) \ . \tag{6.45}$$

The quantities u and v are given by Eqs. (5.18) and (5.23) above.

The large-α limit of the RS theory is derived by expansion in $1/\sqrt{\alpha}$ keeping β fixed,

$$\epsilon_g = \epsilon_{\min} + \frac{1}{2\sqrt{\pi}}\left(\frac{\pi}{2} - 1\right)^{1/4} \frac{I(\beta)}{\beta^{3/2}} \frac{1}{\sqrt{\alpha}} + O(\alpha^{-1}) \ , \tag{6.46}$$

where we have defined the integral

$$I(\beta) = \int Dx \frac{x}{(e^\beta - 1)^{-1} + H(x)} \ . \tag{6.47}$$

As $\beta \to \infty$, $I(\beta) \approx (2\beta)^{3/2}/3$, so that

$$\epsilon_g(T=0,\alpha) = \epsilon_{\min} + \frac{1}{3}\sqrt{2/\pi}\left(\frac{\pi}{2} - 1\right)^{1/4} \frac{1}{\sqrt{\alpha}} + O(\alpha^{-1}) \ . \tag{6.48}$$

Thus for any fixed temperature, the RS generalization curve possesses a $1/\sqrt{\alpha}$ tail.

At $T = 0$, one solution of the RS equations is obtained by taking the limit $\beta \to \infty$ with $1-q$ finite. This solution ceases to exist when $q \to 1$, which happens at the point

$$\frac{R_c}{\sqrt{1-R_c^2}} = \frac{\alpha_c}{\pi} = \cot\frac{2}{\alpha_c} \tag{6.49}$$

or

$$\alpha_c = 1.99, \quad R_c = 0.534 \ . \tag{6.50}$$

Above α_c, the limit $T \to 0$ must be taken with $\beta(1-q)$ finite, yielding

$$\frac{R}{\sqrt{1-R^2}} = \frac{\alpha}{\pi}\left(1 - e^{-\beta(1-q)/(1-R^2)}\right) \ , \tag{6.51}$$

$$\frac{2}{\pi} - R^2 = 2\alpha \int_0^{\sqrt{2\beta(1-q)}} Du\, u^2 H\left(\frac{Ru}{\sqrt{1-R^2}}\right) \ . \tag{6.52}$$

The RS theory predicts that the optimal temperature for generalization $T_{\rm opt}(\alpha)$ is nonzero for α above $\alpha_{\rm th} = 1.27$. The shape of $T_{\rm opt}(\alpha)$ for large α can be derived using the high-temperature expansion of Appendix

F. The dominant terms are

$$\epsilon_g = \epsilon_{\min} + \frac{1}{4\pi}\left(\frac{\pi}{2}-1\right)^{1/4}\frac{1}{T}\sqrt{T/\alpha} + \frac{\pi^2}{12}\left(\frac{\pi}{2}-1\right)^{1/2}\frac{T^2}{\alpha^2} . \tag{6.53}$$

The $1/\sqrt{\alpha}$ term comes from expanding the coefficient of the quenched result (6.46) in β. Thus it represents the result of first taking the $T/\alpha \to 0$ limit and then the $\beta \to 0$ limit. The $1/\alpha^2$ term is just the leading term from the high-T limit, which amounts to taking first the $\beta \to 0$ limit and then the $T/\alpha \to 0$ limit. The optimal T at large α is obtained by minimizing with respect to T, yielding

$$T_{\rm opt} \sim 0.239\,\alpha^{3/5} . \tag{6.54}$$

However, as we will see below, this RS optimal temperature lies below the zero temperature line, i.e., in the regime where the RS solution is unstable.

2. Spin-glass phase

The RS entropy vanishes on the line shown in Fig. 12. Here again the $s = 0$ line provides a lower bound for the spin-glass transition temperature. Comparing with Fig. 10 two differences should be noted. First, the line intersects the α axis at a finite value, at

$$\alpha_c = 1.106 . \tag{6.55}$$

Second, $T_{s=0}(\alpha)$ diverges with α. The asymptotic behavior can be calculated using a double expansion in the variables β and T/α. In Appendix F, we discuss the expansion, and how to locate the dominant terms through power counting. They are

$$s = -\frac{1}{4\pi}\left(\frac{\pi}{2}-1\right)^{1/4}\frac{1}{T}\sqrt{\alpha/T} + \frac{\pi^2}{6}\left(\frac{\pi}{2}-1\right)^{1/2}\frac{T}{\alpha} , \tag{6.56}$$

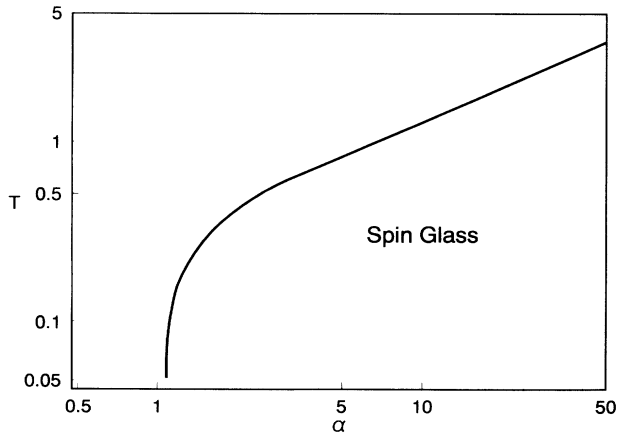

FIG. 12. RS zero-entropy line for the Boolean-mismatched model. The line intersects the α axis at the finite value $\alpha = 1.106$.

which leads to the power law

$$T_{s=0} \sim 0.315\,\alpha^{3/5}, \quad \alpha \to \infty . \tag{6.57}$$

Equation (6.56) also reveals that at fixed T, the RS entropy goes to $-\infty$ as $\alpha \to \infty$, indicating that replica symmetry is violated more and more severely.

Like the realizable Boolean model with discrete weights, this model possesses a frozen one-step RSB solution (see Appendix E). However, in contrast to the Boolean or discrete model, this one-step solution is the equilibrium, not the metastable phase, and exists for arbitrarily high α. According to this RSB solution there is spin-glass phase everywhere below the $s = 0$ line. The training and generalization errors in the SG phase are given by their values on the phase boundary $T_g = T_{s=0}$ at the same α [see Eq. (5.76)]. This is shown in the plot of ϵ_g in Fig. 13. Since $T_g(\alpha)$ is above the RS $T_{\rm opt}(\alpha)$, there is no minimum in $\epsilon_g(T)$ for any α. Instead there is a whole regime of temperatures where ϵ_g does not change with T, implying that optimal generalization can be obtained anywhere on or below the zero entropy line.

For any fixed T, the $\alpha \to \infty$ limit enters the RSB regime. Hence the large α limit of ϵ_g for any fixed T is given by substituting Eq. (6.57) in Eq. (6.53), yielding

$$\epsilon_g(T,\alpha) - \epsilon_{\min} \sim 0.185\alpha^{-4/5} . \tag{6.58}$$

which is independent of T. This power-law decrease is faster than the RS prediction $1/\sqrt{\alpha}$, Eq. (6.48).

3. Numerical simulations

Figure 14 shows Monte Carlo simulations at $T = 0.5$ and 1.0 with $N = 75$, averaged over 64 samples. Below $T = 1$ the training curve is significantly different from the RS result. These deviations for $T = 0.5$ are significantly larger than those observed for the linear-

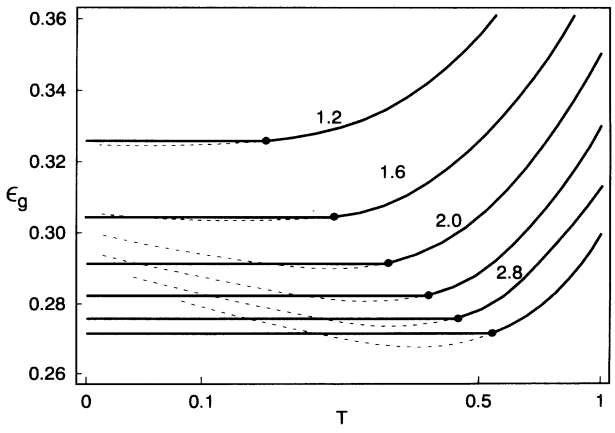

FIG. 13. Generalization error for fixed $\alpha = 1.2$, 1.6, 2.0, 2.4, 2.8, and 3.2 as a function of T for the Boolean-mismatched model. Each curve is a combination of the one-step RSB solution below $T_g(\alpha)$ (marked with a dot on each curve) and the RS solution above. The light dotted lines are the continuation of the RS curves below $T_g(\alpha)$. Note that the minimum of each of the RS curves is below $T_g(\alpha)$.

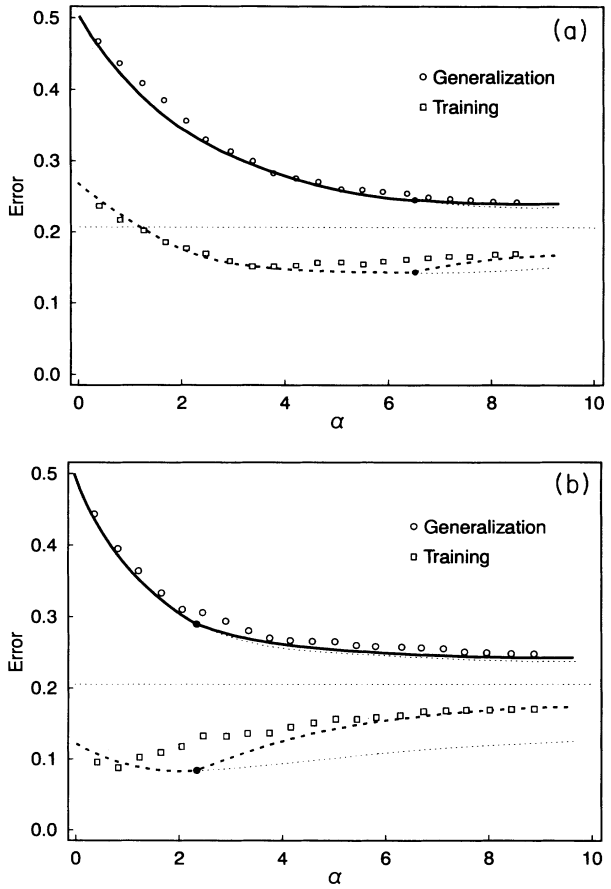

FIG. 14. Learning curves for the Boolean-mismatched model. (a) Simulations at $T = 1.0$ with $N = 100$ and 32 samples. Each learning curve is a combination of the RS solution below the zero-entropy point $\alpha = 6.4$ and the one-step RSB solution above. The light dotted lines are the continuation of the RS curves above $\alpha = 6.4$. (b) Simulations at $T = 0.5$.

mismatched model, indicating the importance of RSB for this model. The RSB solution fits fairly well at large α, but in the vicinity of the transition, there are still some deviations between the theory and simulations. These deviations may be the result of difficulties in equilibrating the system near the transition.

VII. DISCUSSION

A. Learning at finite temperature

In this paper we have studied the process of learning from examples with a stochastic training dynamics. The level of noise in the dynamics is denoted by the temperature T. One of the most important results of our analysis is that learning at finite temperature is possible, and sometimes advantageous. For any finite T, as the number of examples increases, the network weights approach their optimal values, namely the values that minimize the generalization error. Thus even when the generalization error increases with T it may be profitable in certain circumstances to train the system at finite T because convergence times may be prohibitively long at $T = 0$. This is particularly true for highly nonlinear models, such as the Boolean perceptron with discrete weights. Although the critical number of examples per weight $\alpha_c(T)$ increases with T in this model, we have found in our simulations that the time it takes to converge to the optimal state in this model, increases dramatically as T is lowered. It should be stressed however that we have used only a simple Monte Carlo algorithm. Recently several heuristic training algorithms for perceptrons with binary weights have been proposed [34,56]. It would be interesting to study their dynamic and generalization properties.

B. The high-temperature and annealed approximations

We have presented two approximate theoretical approaches to the problem of learning from examples in layered networks. The first approximation replaces the training energy by the number of examples times the generalization error, and becomes exact in the limit of learning with high thermal noise level. The dependence on T and α is only through the effective temperature T/α. Even in this simple approximation, perceptron models exhibit a rich spectrum of generalization behaviors.

The second approximation, the annealed approximation, reduces to the proper high-T limit, but deviates from it significantly at finite α and T, where the behavior is no longer a function of only the ratio of the two parameters.

In all four realizable perceptron rules studied here, these approximations have predicted correctly the shapes of the learning curves at finite T and large α. Furthermore, the AA has yielded interesting results at finite T and α that are qualitatively correct. For instance, the first-order transitions predicted by the AA for the perceptrons with discrete weights are clearly observed in the simulations, Figs. 3 and 8, and are in agreement with the full quenched theory.

On the basis of our general arguments in Secs. II and III, we expect that these approximations will also hold for realizable rules in the more complex cases of multilayer networks. This is borne out in recent studies of two-layer networks of local feature detectors [35], and other multilayer systems [36]. Thus these approximations provide powerful theoretical tools for the study of learning from examples, at least for realizable rules.

Our treatment should be contrasted with the difficult problems of the capacity of single- and multilayer networks [57–59]. The capacity problems usually deal with loading random sets of data. In this case the system is highly frustrated and one has to employ the complex methods of spin-glass theory, such as replica symmetry breaking [33,60]. In the learning problems of the present work, the training set consists of very structured data, generated by a well-defined rule. The underlying structure is represented in our formulation by the generalization function $\epsilon(\mathbf{W})$, Eq. (2.4), which dominates the

behavior of the system at least when the number of examples is large. This function would be completely flat, and therefore meaningless, in the case of random training sets.

In fact, the high-T limit ignores completely the fluctuations due to finite sampling of the rule. It is thus useful in studying how learning is affected by the nature of the target rule [e.g., through the function $\epsilon(\mathbf{W})$] as well as by the network architecture [e.g., the entropy $s(\epsilon)$, Eq. (2.26)]. As has been explained in Sec. III the AA does take into account the effect of randomness induced by the examples, though only approximately. The reasonable results generated by the AA even at finite α and T imply that this randomness may not have major effects (e.g., frustration and other spin-glass phenomena), at least in models where the chosen architecture is compatible with the rule to be learned. This may explain why, in many applications of supervised learning, simple local gradient algorithms seem to yield good results.

It should be emphasized that the above approximations may be useful not only for studying specific "toy" models, but also in generating general approximate predictions that could perhaps be useful in applied research on neural networks. An example is Eq. (2.40), which provides a way of estimating the generalization error from the observed training error for a multilayer network with a Boolean output.

C. Inverse power law for smooth networks

The most important general result is the inverse power law for the asymptotic behavior of both the generalization and training errors, Eqs. (3.12) and (3.14) of Sec. III. These results also provide a simple relationship between the two errors, namely Eq. (3.15). This power law is consistent with the general bound obtained within PAC learning theory. However, our theory is not distribution-free and holds only for smooth networks. On the other hand, it holds for general unrealizable rules, whereas the PAC bounds are essentially for realizable rules.

The results regarding smooth networks were derived using a perturbative approach, i.e., assuming that essentially all the components of $\mathbf{W}$ deviate only slightly from $\mathbf{W}^*$ for sufficiently large α. Of course in reality there can also be contributions coming from $\mathbf{W}$ far away from $\mathbf{W}^*$, leading to nonperturbative terms to ϵ_g. However, we expect that for sufficiently large α the nonperturbative terms will be negligible (e.g., exponentially dependent on α) relative to the power-law contribution of the smooth fluctuations. On the other hand, these localized nonperturbative errors may be important for the dynamics of the training, since their relaxation may be extremely slow compared with the continuous fluctuations.

In nonsmooth networks the generalization performance depends on the nature of the task as well as the network architecture, as the results of Secs. V and VI demonstrate. These results also indicate that there is a qualitative difference between the learning of realizable and unrealizable rules, as discussed below.

D. Learning realizable rules

The results of the specific models studied here indicate that the shapes of learning curves may be very different from the PAC learning bound of an inverse power law. The shape of the generalization curve depends strongly on the degree of constraint imposed on the network space. In the case of a linear output, the imposition of binary constraints changes a T/α tail [Eq. (5.10)] into an exponential $\exp(-2\alpha/T)$ [Eq. (5.32)]. In addition, at low T there is a discontinuous transition at finite α from poor to good generalization; see Figs. 1 and 3. The superior generalization ability of the constrained network is not surprising. Since the target rule itself was assumed to be realizable in the constrained architecture, imposing the constraints is essentially using prior knowledge (or assumptions) of the nature of the rule to restrict the space of possible networks.

The most dramatic effect of constraining the network is found for the conventional perceptron rule, i.e., one with a Boolean output. In the unconstrained network the generalization error again behaves as an inverse power law [this time even at $T = 0$, Eq. (5.28)]. On the other hand, in the case of binary weights there is a discontinuous transition at a critical α from poor to perfect learning (see Fig. 6). This transition is unique in that it exists even in learning at high temperatures. The collapse of a thermodynamic system at finite T to its ground state above some α_c stems from the singular spectrum of excitations above this state, as discussed in Sec. V D.

In all the realizable models studied in this work, the quenched RS behavior at finite T is qualitatively similar to that given by the AA. Furthermore, the asymptotic shapes of the RS learning curves in the different models agree with those of the AA. This suggests that for realizable rules the effect of disorder is minor for large α.

The main qualitatively different result of the quenched theory is the appearance of spin-glass phases, as shown in Figs. 1 and 6. These phases result from the randomness and frustration associated with optimizing for a particular realization of examples, and cannot be predicted by the high-T or the annealed approximations. Two special features distinguish the SG phases of realizable models. First, it exists only at low T and in restricted regime of α, typically $\alpha \lesssim 1$. Second, it is only metastable, i.e., its free energy (as well as training energy) is higher than that of the optimal state, which is therefore the true equilibrium state in that regime of α.

The absence of SG phases at large α indicates that as α increases, the relative scale of fluctuations in the training energy becomes smaller, i.e., the training energy surface gets smoother as the number of examples increases. This is indeed demonstrated by the analytical bounds of Sec. V B on the number of local minima for the linear perceptron with discrete weights (Fig. 4). These bounds exhibit two important, and possibly general, features. First, the energy surface in the neighborhood of the optimal state $R = 1$ is smooth. Second, the number of local minima is exponentially large for small α, but decreases monotonically with α. In the linear-discrete model there are no local minima above a critical value of

α. Fontanari and Köberle [61] have studied numerically the local minima of the Boolean-discrete model for small system sizes. The number of local minima was found to scale exponentially with N and to decrease monotonically with α. No critical value of α for the disappearance of local minima in the Boolean model was found. However, a definitive conclusion would require the investigation of larger sizes.

Finally, in all our realizable models the generalization error decreases monotonically with T, so that the optimal temperature for learning is

$$T_{\rm opt} = 0 \ . \tag{7.1}$$

Of course this refers to the equilibrium properties, whereas from dynamic point of view the optimal temperatures may be finite even in realizable rules, as has been pointed out above.

E. Classification of learning curves for realizable rules

The discontinuous behavior of the Boolean perceptron with discrete weights calls for an understanding of the general conditions which determine whether a given learning task will be achieved gradually or not. Insight into this question is provided by the high-T limit of Sec. II C. According to Eqs. (2.25)–(2.27), a system can be classified according to the behavior of its entropy $s(\epsilon)$. Recall that $e^{Ns(\epsilon)}$ is defined as the number of networks that yield a given generalization error ϵ. The general form of $s(\epsilon)$ for small ϵ is expected to be

$$s(\epsilon) \propto \epsilon^x \ln \epsilon, \quad \epsilon \to 0 \ . \tag{7.2}$$

For a smooth weight space we expect s to diverge logarithmically to $-\infty$, i.e., $x = 0$. This naturally leads, via Eq. (2.27), to the inverse power law. For a discrete weight space, we expect the entropy to approach a finite value, i.e., $x > 0$. When $0 < x < 1$ the generalization curve obeys a nontrivial power law

$$\epsilon_g \approx (\alpha\beta)^{-1/(1-x)}, \quad \alpha\beta \to \infty, \quad 0 < x < 1 \ . \tag{7.3}$$

up to logarithmic corrections. When $x = 1$ this law turns into an exponential,

$$\ln \epsilon_g \propto -\alpha\beta, \quad \alpha\beta \to \infty, \quad x = 1 \ . \tag{7.4}$$

An example of this case is the perceptron with discrete weights and linear output; see Eq. (5.32). When $x > 1$, $\partial s/\partial \epsilon$ remains finite at small ϵ, so there must be a discontinuous transition at some $(\alpha\beta)_c$ to $\epsilon_g = 0$,

$$\epsilon_g = 0, \quad \alpha\beta > (\alpha\beta)_c, \quad x > 1 \ . \tag{7.5}$$

The Boolean perceptron with discrete weights is an example of an entropy with $x = 2$; see Eqs. (5.62) and (5.63).

In our perceptron models, nontrivial shapes of learning curves resulted only when the weights were constrained to discrete values. In contrast, for multilayer networks we expect that the exponent x in Eq. (7.2) may be nontrivial even when the weights are allowed to vary continuously, possibly leading to Eq. (7.4) or (7.5). This will occur, for example, if the optimal solution involves *discrete* internal representations of the hidden neurons, as demonstrated in the two-layer model network for the contiguity problem [35].

The above classification requires knowledge of the behavior of $s(\epsilon)$ near $\epsilon = 0$. In general, evaluating s may be difficult even in relatively simple models, since the primary interest is its behavior in the limit $N \to \infty$. For this reason, the simple existence of a gap in the spectrum of ϵ is an insufficient basis for classification. At the very least, the scaling of the gap with N must be determined (see Sec. IV F). Finally, it should be stressed that the classification of the asymptotic shapes of the learning curves according to the properties of $s(\epsilon)$ has been justified so far only in the high-T limit. It may be, however, that in many classes of realizable rules these results apply also to finite T or even $T = 0$. If this is true then our results may provide useful hints for understanding or even predicting the behavior of some real-world learning problems.

F. Unrealizable rules

For unrealizable rules the optimal weight vector $\mathbf{W}^*$ is reached only in the limit $\alpha \to \infty$. This is true at all temperatures. As $T \to 0$ the generic shape of the learning curves is roughly that shown in Fig. 11(b). There is a critical value α_c which marks the loading capacity, i.e., the point below which the examples are memorized perfectly. Above α_c the training error increases, and approaches the same limit as the generalization error, i.e., approaches $\epsilon_{\rm min}$ from below as $\alpha \to \infty$. Also, in some unrealizable models (not encountered in this work, but see Refs. [22, 24]) there is a local maximum in ϵ_g near α_c.

One of the important results of our study is that for unrealizable rules the learning curves are substantially different from the high-T and AA predictions at any fixed T. For example, for the models with weight mismatch, these approximations predict α^{-2} tails for fixed T, whereas the RS theory yields [see Eqs. (6.29)]

$$\epsilon_g(T, \alpha) - \epsilon_{\rm min} \sim \alpha^{-1} \tag{7.6}$$

for the case of a linear output. In the Boolean output case the full spin-glass theory predicts

$$\epsilon_g(T, \alpha) - \epsilon_{\rm min} \sim \alpha^{-4/5}. \tag{7.7}$$

An $\alpha^{-1/2}$ tail in ϵ_g has also been found in Boolean perceptrons with continuous weights where the rule is unrealizable due to corruption of the examples by noise [24] or due to the random nature of the rule itself [22].

The RS theory of perceptrons with unrealizable rules predicts that, for sufficiently large α, ϵ_g is nonmonotonic with T, but possesses a nonzero optimal temperature. Similar results have been obtained in [22, 24]. In the case of a Boolean perceptron with weight mismatch, where the corrections to RS theory could be calculated, it has been found that rather than having a minimum at finite T, ϵ_g is independent of T at low T; see Fig. 13. Whether correction to the RS theory will modify substantially the conclusions regarding $T_{\rm opt}$ of the RS theory in the other models remains to be studied.

Another feature of unrealizable rules is the prevalence

of spin-glass phases even at large α. In the present work we have evaluated the T at which the entropy vanishes. This provides a lower bound for the onset of the spin-glass phase. Here again the linear and Boolean perceptrons differ. In the linear case this temperature levels off at a finite value as α increases, as shown in Fig. 10. In the Boolean case it grows with the same power law as $T_{\rm opt}$; see Eq. (6.57) and Fig. 12. These results suggest that spin-glass effects are strong in unrealizable cases.

The fact that the spin-glass phase exists also for large α suggests that the fluctuations in the training energy do not necessarily shrink as α increases. Nevertheless, we believe that their *relative* scale does indeed vanish. This is because the energy itself grows linearly with α. This conjecture is supported by the observation that the temperature that marks the onset of spin-glass phenomena grows only sublinearly with α [see Eq. (6.57)]. The effect of increasing the number of examples on the roughness of the training energy surface (particularly in unrealizable rules) is an important issue which deserves further study.

G. Uniqueness of the optimal solution

Throughout this paper we have assumed that the optimal network weights, i.e., the components of the weight vector $\mathbf{W}^*$ that globally minimizes the generalization function $\epsilon(\mathbf{W})$, are unique. If they are nonunique, we assume that the degenerate global minima are at least widely separated in network space, forming a discrete set. This is related to our assumption that the training dynamics searches for the values of the network weights within a well-defined architecture. Under these conditions it is reasonable to expect that the optimal solution is generally unique up to obvious symmetries. For instance, in multilayer networks the solution may be unique only up to permutation of the hidden neurons [62]. Other symmetries may result from the nature of the rule itself. An example is the up-down degeneracy of the optimal solution of the two layer network of edge detectors in Ref. [35]. This uniqueness does not hold if the architecture of the network, e.g., the number of neurons or the number of layers, is allowed to vary significantly. The important issue of learning under these circumstances is planned to be discussed elsewhere.

ACKNOWLEDGMENTS

We would like to thank E. Baum, D. Hansel, D. Haussler, and D. Huse for helpful discussions. We are grateful to G. Györgyi for his comments on an earlier version of the manuscript and for drawing our attention to the RSB phase in the Boolean model with mismatched weights. We also thank the Aspen Center for Physics for hosting a Workshop on Neural Networks where part of this research was conducted. H. S. S. acknowledges support from Harvard University.

APPENDIX A: TRAINING-GENERALIZATION INEQUALITY

Let E_0 denote the average value of the training energy $P\epsilon(\mathbf{W})$, as in Eq. (2.18) and δE the difference $E - E_0$. Define the function

$$\Delta(\beta,\gamma) \equiv \frac{\int d\mu(\mathbf{W}) e^{-\gamma E_0 - \beta\delta E}\delta E}{\int d\mu(\mathbf{W}) e^{-\gamma E_0 - \beta\delta E}} . \tag{A1}$$

This can be interpreted as the energy of a system at temperature $1/\beta$ with Hamiltonian δE and measure $d\mu(\mathbf{W})e^{-\gamma E_0}$. By the convexity of the free energy, the energy is a decreasing function of β. Hence Δ is bounded above,

$$\Delta(\beta,\gamma) \le \Delta(\beta=0,\gamma) = \frac{\int d\mu(\mathbf{W}) e^{-\gamma E_0}\delta E}{\int d\mu(\mathbf{W}) e^{-\gamma E_0}} . \tag{A2}$$

We now take the quenched average of both sides. In the integrals, the average can be applied directly to δE, since E_0 is independent of the quenched disorder. But $\langle\!\langle\, \delta E \,\rangle\!\rangle = 0$, so the right-hand side vanishes, and we have

$$\langle\!\langle\, \Delta(\beta,\gamma) \,\rangle\!\rangle \le \langle\!\langle\, \Delta(\beta=0,\gamma) \,\rangle\!\rangle = 0 . \tag{A3}$$

In particular, for $\gamma = \beta$, $\langle\!\langle\, \Delta \,\rangle\!\rangle = P(\epsilon_t - \epsilon_g)$, so that we finally obtain

$$\epsilon_t(T,\alpha) \le \epsilon_g(T,\alpha) , \tag{A4}$$

which was stated without proof in Eq. (2.20).

APPENDIX B: HIGH-TEMPERATURE EXPANSION

In this appendix, we show using a cumulant expansion that the free energy (2.13) can be written as a power series in β, with coefficients that are functions of $\alpha\beta$. The zeroth-order term of this series is the high-temperature limit discussed in Sec. II C.

The first step is to separate the energy into random and nonrandom parts. The nonrandom part is

$$E_0 \equiv \langle\!\langle\, E \,\rangle\!\rangle = P\epsilon(\mathbf{W}) = N\alpha\epsilon(\mathbf{W}) , \tag{B1}$$

and the random part is

$$\delta E \equiv E - E_0 = \sum_{\mu=1}^{P} \delta\epsilon_\mu , \tag{B2}$$

where

$$\delta\epsilon_\mu \equiv \epsilon(\mathbf{W};\mathbf{S}^\mu) - \epsilon(\mathbf{W}). \tag{B3}$$

We now treat δE as a perturbation to the "free Hamiltonian" E_0. The partition function takes the form

$$Z = Z_0\langle e^{-\beta\delta E}\rangle_0 . \tag{B4}$$

Here $\langle\,\rangle_0$ denotes the average with respect to the distri-

bution $Z_0^{-1}e^{-\beta E_0}$. The factor

$$Z_0(\beta\alpha) = \int d\mu(\mathbf{W}) e^{-\beta E_0} = \int d\mu(\mathbf{W}) e^{-N\beta\alpha\epsilon(\mathbf{W})} \tag{B5}$$

is the high-T partition function introduced in Sec. II C.

Taking the logarithm of both sides and performing the quenched average, we obtain

$$-\beta F = \ln Z_0(\alpha\beta) + \sum_{j=1}^{\infty} \frac{(-\beta)^j}{j!} \langle\langle C_j \rangle\rangle, \tag{B6}$$

where the C_j are from the cumulant expansion of $\ln\langle e^{-\beta\delta E}\rangle_0$

$$C_1 = \langle \delta E \rangle_0, \tag{B7}$$

$$C_2 = \langle (\delta E)^2 \rangle_0 - \langle \delta E \rangle_0^2, \tag{B8}$$

$$C_3 = \cdots . \tag{B9}$$

If δE were a quantity of order unity, the cumulants would be functions only of $\alpha\beta$, the only parameter governing the distribution $e^{-\beta E_0}$. Hence Eq. (B6) would be the desired power series in β. In fact, the situation is somewhat more complicated because δE scales like α.

To investigate the scaling of the cumulants with α, we write them as sums over connected correlation functions

$$C_j = \sum_{\mu_1,\ldots,\mu_j=1}^{P} \langle \delta\epsilon_{\mu_1} \cdots \delta\epsilon_{\mu_j} \rangle_c . \tag{B10}$$

Counting the P^j terms in this sum, the naive estimate would be that $C_j \sim P^j$. However, the quenched average makes any term containing an unrepeated index vanish, since $\langle\langle \delta\epsilon_\mu \rangle\rangle = 0$. In other words, if an index appears, it must appear at least twice for the term to be nonvanishing. This means that the cumulants can scale no faster than $C_j \sim P^{[j/2]}$. At the same time, there are also factors of N which make the cumulant extensive, so that it behaves like

$$C_j \sim N\alpha^{[j/2]} c_j(\alpha\beta) , \tag{B11}$$

to leading order in α.

In the expansion for the free energy, the term containing C_j only contributes to terms of order $\beta^{[j/2]}$ or higher in the ultimate high-temperature expansion. Hence the free energy can be written in the form

$$-\beta F = \ln Z_0 + \sum_{j=1}^{\infty} \beta^j F_j(\alpha\beta), \tag{B12}$$

where F_j contains contributions from the finite number of cumulants $C_1, \ldots, C_{2j}$.

In general, any quantity A that is finite in the high-T limit possesses a high-temperature expansion of the form

$$A(\beta, T/\alpha) = \sum_{i=0}^{\infty} \beta^i A_i(T/\alpha). \tag{B13}$$

In the high-temperature limit ($\beta \to 0$, $T/\alpha = \text{const}$),

$$A(\beta, T/\alpha) \to A_0(T/\alpha) , \tag{B14}$$

which depends only on the effective temperature T/α. In general, the functions A_i can be nonanalytic functions of T/α. In particular, the high-T limit term A_0 can be nontrivial and lead to such behavior as the first-order transition in the Boolean-discrete model.

APPENDIX C: ANNEALED APPROXIMATION FOR PERCEPTRON LEARNING

In this appendix, we give a fuller account of the results that were outlined in Secs. IV A and IV B. We begin with a derivation of Eq. (4.7) for the average generalization error, which illustrates many of the calculational techniques of this paper. Integrating the error function (4.6) over the *a priori* input measure (4.3), we obtain

$$\epsilon(\mathbf{W}) = \int D\mathbf{S}\, \epsilon(\mathbf{W};\mathbf{S}) = \int dx \int dy \frac{1}{2}[g(x) - g(y)]^2 \int D\mathbf{S}\, \delta\left(x - N^{-1/2}\mathbf{W}\cdot\mathbf{S}\right) \delta\left(y - N^{-1/2}\mathbf{W}^0\cdot\mathbf{S}\right), \tag{C1}$$

$$\epsilon(\mathbf{W}) = \int \frac{dx\, d\hat{x}}{2\pi} \int \frac{dy\, d\hat{y}}{2\pi} e^{ix\hat{x}+iy\hat{y}} \frac{1}{2}[g(x) - g(y)]^2 \int D\mathbf{S} \exp\left[-iN^{-1/2}(\mathbf{W}\hat{x} + \mathbf{W}^0\hat{y})\cdot\mathbf{S}\right] . \tag{C2}$$

The two auxiliary variables x and y are introduced to remove $\mathbf{S}$ from the argument of the g functions, and $\hat{x}$ and $\hat{y}$ are introduced to transform the δ functions into exponentials using the identity

$$\delta(x) = \int \frac{d\hat{x}}{2\pi} e^{ix\hat{x}} . \tag{C3}$$

When the Gaussian average over $\mathbf{S}$ is now performed, a simple Gaussian integral in $\hat{x}$ and $\hat{y}$ is left. These variables can in turn be integrated out, leaving

$$\epsilon(\mathbf{W}) = \int \frac{dx\, dy}{2\pi\sqrt{1-R^2}} \exp\left(-\frac{x^2+y^2-2xyR}{2(1-R^2)}\right) \times \frac{1}{2}[g(x) - g(y)]^2 . \tag{C4}$$

This result has a simple interpretation. It is the average

of $\frac{1}{2}[g(x)-g(y)]^2$, where x and y, like $\mathbf{W}\cdot\mathbf{S}/\sqrt{N}$ and $\mathbf{W}^0\cdot\mathbf{S}/\sqrt{N}$ in (4.6), are Gaussian variables with unit variance and cross correlation R. A simple change of variables in Eq. (C4), a shift followed by a rescaling, yields the form (4.7).

To derive the annealed approximation for perceptron learning, we begin by evaluating Eq. (2.31) for $G_{\rm an}(\mathbf{W})$. The calculation is essentially the same as the previous one for $\epsilon(\mathbf{W})$, and results in a similar formula,

$$G_{\rm an}(\mathbf{W}) = -\ln \int \frac{dx\,dy}{2\pi\sqrt{1-R^2}} \exp\left(-\frac{x^2+y^2-2xyR}{2(1-R^2)}\right) \times \exp\{-\beta[g(x)-g(y)]^2/2\} \ . \tag{C5}$$

Again, the answer depends on $\mathbf{W}$ only through the overlap R. A change of variables in the integral (C5) yields the form (4.14). Evaluating the integral for the cases of $g(x) = x$ (linear perceptron) and $g(x) = \mathrm{sgn}(x)/\sqrt{2}$ (Boolean perceptron), we obtain

$$G_{\rm an} = \begin{cases} \frac{1}{2}\ln\left[1+2\beta(1-R)\right] & \text{(linear)} \\ -\ln\left[1-(1-e^{-\beta})\pi^{-1}\cos^{-1}R\right] & \text{(Boolean)}. \end{cases} \tag{C6}$$

Although derived using the Gaussian *a priori* input distribution (4.3), the above results for $\epsilon(\mathbf{W})$ and $G_{\rm an}(\mathbf{W})$ apply also to the case of the discrete inputs $S_i = \pm 1$ in the thermodynamic limit. This insensitivity to input distribution is explained by the central limit theorem, which guarantees that $\mathbf{W}\cdot\mathbf{S}/\sqrt{N}$ and $\mathbf{W}^0\cdot\mathbf{S}/\sqrt{N}$ are Gaussian variables (in the $N\to\infty$ limit) with very weak assumptions about the distribution of $\mathbf{S}$. This assertion may be verified by a straightforward calculation for discrete $\mathbf{S}$, which yields Eqs. (C6) as the leading terms in a saddle-point expansion in $1/N$.

Since $G_{\rm an}(\mathbf{W})$, Eq. (2.31), depends only on the overlap R, the annealed partition function (2.30) can now be rewritten as an integral over R,

$$\langle\!\langle\, Z \,\rangle\!\rangle = \int dR \exp N[G_0(R) - \alpha G_{\rm an}(R)] \ , \tag{C7}$$

where

$$NG_0(R) = \ln \int d\mu(\mathbf{W})\, \delta(R - N^{-1}\mathbf{W}\cdot\mathbf{W}^0) \tag{C8}$$

is the logarithm of the density of networks with overlap R. In the thermodynamic limit ($N\to\infty$), the integral can be evaluated as

$$-\beta f(T,\alpha) \equiv \frac{1}{N}\ln\langle\!\langle\, Z \,\rangle\!\rangle = \mathrm{extr}_R[G_0(R) - \alpha G_{\rm an}(R)] \ . \tag{C9}$$

Hence the thermodynamic free energy $f(T,\alpha)$ is determined by extremizing the free-energy function $-\beta f(R) \equiv G_0(R) - \alpha G_{\rm an}(R)$. Differentiating $f(R)$ with respect to R, we obtain the stationarity condition Eq. (4.16).

We can write the stationarity equations in a more revealing form by proceeding further in the evaluation of G_0. The δ function in (C8) can be expanded by introducing another order parameter $\hat{R}$,

$$G_0 = \frac{1}{N}\ln \int_{-i\infty}^{i\infty} \frac{d\hat{R}}{2\pi i} \exp\Big(-NR\hat{R} + \ln \int d\mu(\mathbf{W})\, e^{\hat{R}\mathbf{W}\cdot\mathbf{W}^0} \Big) \ . \tag{C10}$$

In the thermodynamic limit, this reduces to

$$G_0 = -R\hat{R} + \frac{1}{N}\ln \int d\mu(\mathbf{W})\, e^{\hat{R}\mathbf{W}\cdot\mathbf{W}^0} \ , \tag{C11}$$

where the right-hand side must be stationary with respect to the saddle-point parameter $\hat{R}$. The free energy can now be written as a saddle point over two order parameters

$$-\beta f = \mathrm{extr}_{R,\hat{R}}\Big(-R\hat{R} + \frac{1}{N}\ln \int d\mu(\mathbf{W})\, e^{\hat{R}\mathbf{W}\cdot\mathbf{W}^0} - \alpha G_{\rm an}(R) \Big) \tag{C12}$$

The saddle-point equations are

$$\hat{R} = -\alpha \frac{\partial G_{\rm an}}{\partial R} \ , \tag{C13}$$

$$R = \frac{1}{N}\langle\mathbf{W}\rangle_{\hat{R}}\cdot\mathbf{W}^0 \ , \tag{C14}$$

where

$$\langle\mathbf{W}\rangle_{\hat{R}} \equiv \frac{\int d\mu(\mathbf{W})\,\mathbf{W}\exp(\hat{R}\mathbf{W}\cdot\mathbf{W}^0)}{\int d\mu(\mathbf{W})\exp(\hat{R}\mathbf{W}\cdot\mathbf{W}^0)} . \tag{C15}$$

The order parameter $\hat{R}$ has a natural interpretation: it is the strength of a local field pushing $\mathbf{W}$ in the direction of $\mathbf{W}^0$. Since it increases with α, it forces $\mathbf{W}$ toward $\mathbf{W}_0$ as $\alpha\to\infty$. Upon eliminating $\hat{R}$, these equations reduce to Eq. (4.16).

Equation (C11) can be evaluated quite easily for the case of the Ising constraint $W_i = \pm 1$. Then we can make the replacement

$$\int d\mu(\mathbf{W}) \to \sum_{W_i=\pm 1} \ , \tag{C16}$$

which leads finally to

$$G_0(R) = -R\hat{R} + \ln 2\cosh\hat{R} \ , \tag{C17}$$

assuming that the teacher weights also satisfy the Ising constraint. Extremizing with respect to $\hat{R}$, we obtain the equation

$$R = \tanh\hat{R} \tag{C18}$$

Eliminating $\hat{R}$, one can finally derive the result (4.20), which is purely a function of R, and is the familiar result for the entropy of the Ising model as a function of the magnetization R.

Calculating G_0 for the spherical distribution is somewhat more complicated. We rewrite the *a priori* spherical

distribution (4.17) as

$$d\mu(\mathbf{W}) = \prod_{i=1}^{N} \frac{dW_i}{\sqrt{2\pi e}} \int_{-i\infty}^{i\infty} \frac{d\lambda}{2\pi i} e^{\lambda(\mathbf{W}\cdot\mathbf{W}-N)} \tag{C19}$$

and then G_0 is determined as a saddle point over $\hat{R}$ and λ,

$$\begin{aligned} G_0 &= -\frac{1}{2}\ln(2\pi e) + R\hat{R} + \lambda + \ln \int dW e^{-\lambda W^2 - \hat{R}W} \\ &= -\frac{1}{2} + R\hat{R} + \lambda - \frac{1}{2}\ln(2\lambda) + \frac{\hat{R}^2}{4\lambda}. \end{aligned} \tag{C20}$$

Eliminating $\hat{R}$ and λ finally yields the result (4.18), which was justified previously by geometric arguments.

APPENDIX D: REPLICA THEORY OF PERCEPTRON LEARNING

The starting point of the replica calculations is the derivation of the replicated Hamiltonian (2.49), which resembles the derivation of the annealed G_{an} described in Appendix C. We introduce auxiliary variables x_σ, $\hat{x}_\sigma$, y, and $\hat{y}$ in order to simplify the average over $\mathbf{S}$,

$$\begin{aligned} e^{-\mathcal{G}_r[\mathbf{W}^\sigma]} &= \int D\mathbf{S} \exp\left(-\beta \sum_{\sigma=1}^{n} \epsilon(\mathbf{W}^\sigma; \mathbf{S})\right) \\ &= \int \prod_\sigma dx_\sigma \int dy \exp\left(-\frac{1}{2}\beta \sum_\sigma [g(x_\sigma) - g(y)]^2\right) \int D\mathbf{S} \prod_{\sigma=1}^{n} \delta\left(x_\sigma - N^{-1/2}\mathbf{W}^\sigma \cdot \mathbf{S}\right) \delta\left(y - N^{-1/2}\mathbf{W}^0 \cdot \mathbf{S}\right) \\ &= \int \prod_\sigma \frac{dx_\sigma\, d\hat{x}_\sigma}{2\pi} \int \frac{dy\, d\hat{y}}{2\pi} \exp\left(-\frac{1}{2}\beta \sum_\sigma [g(x_\sigma) - g(y)]^2 + i\sum_\sigma x_\sigma \hat{x}_\sigma + iy\hat{y}\right) \\ &\quad \times \int D\mathbf{S} \exp\left[-iN^{-\frac{1}{2}}\left(\sum_\sigma \mathbf{W}^\sigma \hat{x}_\sigma + \mathbf{W}^0 \hat{y}\right)\cdot \mathbf{S}\right]. \end{aligned} \tag{D1}$$

The average over $\mathbf{S}$ is now a simple Gaussian integral and yields

$$\begin{aligned} e^{-\mathcal{G}_r[Q_{\sigma\rho}, R_\sigma]} &= \int \prod_\sigma \frac{dx_\sigma\, d\hat{x}_\sigma}{2\pi} \int \frac{dy\, d\hat{y}}{2\pi} \exp\left(-\frac{1}{2}\beta \sum_\sigma [g(x_\sigma) - g(y)]^2 + i\sum_\sigma x_\sigma \hat{x}_\sigma + iy\hat{y}\right) \\ &\quad \times \exp\left(-\frac{1}{2}\sum_{\sigma,\rho} \hat{x}_\sigma \hat{x}_\rho Q_{\sigma\rho} - \hat{y}\sum_\sigma \hat{x}_\sigma R_\sigma - \frac{1}{2}\hat{y}^2\right). \end{aligned} \tag{D2}$$

Since $\mathcal{G}_r$ depends on the weights only through the order parameters $Q_{\sigma\rho}$ and R_σ that were defined in Eqs. (2.55) and (4.21), the replicated partition function can be written as an integral over these order parameters

$$\begin{aligned} \langle\!\langle Z^n \rangle\!\rangle &= \int \prod_{\sigma<\rho} dQ_{\sigma\rho} \int \prod_\sigma dR_\sigma \exp(-N\alpha \mathcal{G}_r[Q_{\sigma\rho}, R_\sigma]) \\ &\quad \times \int \prod_\sigma d\mu(\mathbf{W}^\sigma) \prod_{\sigma<\rho} \delta\left(Q_{\sigma\rho} - N^{-1}\mathbf{W}^\sigma \cdot \mathbf{W}^\rho\right) \prod_\sigma \delta\left(R_\sigma - N^{-1}\mathbf{W}^\sigma \cdot \mathbf{W}^0\right), \end{aligned} \tag{D3}$$

$$\langle\!\langle Z^n \rangle\!\rangle = \int \prod_{\sigma<\rho} \frac{dQ_{\sigma\rho} d\hat{Q}_{\sigma\rho}}{2\pi i} \int \prod_\sigma \frac{dR_\sigma d\hat{R}_\sigma}{2\pi i} \exp N(\mathcal{G}_0[Q_{\sigma\rho}, R_\sigma, \hat{Q}_{\sigma\rho}, \hat{R}_\sigma] - \alpha \mathcal{G}_r[Q_{\sigma\rho}, R_\sigma]), \tag{D4}$$

where

$$\mathcal{G}_0 = -\sum_\sigma \hat{R}_\sigma R_\sigma - \sum_{\sigma<\rho} \hat{Q}_{\sigma\rho} Q_{\sigma\rho} + \frac{1}{N} \ln \int \prod_\sigma d\mu(\mathbf{W}^\sigma) \exp\left(\sum_\sigma \hat{R}_\sigma \mathbf{W}^\sigma \cdot \mathbf{W}^0 + \sum_{\sigma<\rho} \hat{Q}_{\sigma\rho} \mathbf{W}^\sigma \cdot \mathbf{W}^\rho\right) \tag{D5}$$

is the logarithm of the density of replicated networks with the overlaps $Q_{\sigma\rho}$ and R_σ.

In the thermodynamic limit, the integral (D3) over the order parameters is dominated by the saddle point in $Q_{\sigma\rho}$ and R_σ. The free energy is obtained by analytically continuing this saddle point to $n = 0$,

$$-\beta f = \lim_{n\to 0} \frac{1}{nN} \ln \langle\langle Z^n \rangle\rangle = \operatorname*{extr}_{Q_{\sigma\rho}, R_\sigma} \left\{ \mathcal{G}_0[Q_{\sigma\rho}, R_\sigma, \hat{Q}_{\sigma\rho}, \hat{R}_\sigma] - \alpha \mathcal{G}_r[Q_{\sigma\rho}, R_\sigma] \right\} . \tag{D6}$$

According to the RS ansatz, the saddle point takes the form

$$Q_{\sigma\rho} = \delta_{\sigma\rho} + (1 - \delta_{\sigma\rho}) q , \tag{D7}$$

$$R_\sigma = R , \tag{D8}$$

$$\hat{Q}_{\sigma\rho} = \delta_{\sigma\rho} + (1 - \delta_{\sigma\rho}) \hat{q} , \tag{D9}$$

$$\hat{R}_\sigma = \hat{R} . \tag{D10}$$

With this substitutions, the free energy takes the form

$$-\beta f = \operatorname*{extr}_{q,R} [G_0(q, R, \hat{q}, \hat{R}) - \alpha G_r(q, R)] , \tag{D11}$$

where we have defined

$$G_r \equiv \lim_{n\to 0} \frac{\mathcal{G}_r}{n} , \tag{D12}$$

$$G_0 \equiv \lim_{n\to 0} \frac{\mathcal{G}_0}{n} . \tag{D13}$$

To calculate G_r, we substitute the RS ansatz into Eq. (D2) and perform the integral over $\hat{y}$, leaving

$$e^{-\mathcal{G}_r} = \int Dy \int \prod_\sigma \frac{dx_\sigma d\hat{x}_\sigma}{2\pi} \exp\left(-\frac{1}{2}\beta \sum_\sigma [g(x_\sigma) - g(y)]^2 \right) \times \exp\left(-\frac{1}{2}(1-q) \sum_\sigma \hat{x}_\sigma^2 + i \sum_\sigma \hat{x}_\sigma (x_\sigma - Ry) \right) \int Dt \exp\left(-it\sqrt{q - R^2} \sum_\sigma x_\sigma \right) . \tag{D14}$$

The auxiliary variable t has been introduced via the identity $\int Dt\, e^{bt} = \exp(\frac{1}{2}b^2)$. Performing the integrals over the $\hat{x}_\sigma$, and shifting and rescaling the x_σ integrals, we finally obtain

$$e^{-\mathcal{G}_r} = \int Dt \int Dy \left[\int Dx \exp\left(-\frac{1}{2}\beta \left[g\left(x\sqrt{1-q} + yR + t\sqrt{q - R^2} \right) - g(y) \right]^2 \right) \right]^n . \tag{D15}$$

Taking the limit $n \to 0$ gives Eq. (4.27) for G_r.

The RS result for $\mathcal{G}_0$ is

$$\mathcal{G}_0 = n \left(-R\hat{R} - \frac{1}{2} n(n-1) q\hat{q} - \frac{1}{2}\hat{q} \right) + \frac{1}{N} \ln \int D\mathbf{z} \left(\int d\mu(\mathbf{W}) \exp[\mathbf{W} \cdot (\mathbf{z}\sqrt{\hat{q}} + \mathbf{W}^0 \hat{R})] \right)^n . \tag{D16}$$

The $n \to 0$ limit of this expression yields Eq. (4.26) for G_0.

APPENDIX E: ONE-STEP RSB FOR PERCEPTRONS

Following the Parisi theory of RSB [45] we make the following "one-step" ansatz for the form of the order parameter matrix $Q_{\mu\nu}$:

$$Q_{\mu\nu} = \begin{pmatrix} 1 & q_1 & \cdots & q_0 & q_0 & \cdots & \\ q_1 & 1 & \cdots & q_0 & q_0 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \\ q_0 & q_0 & \cdots & 1 & q_1 & \cdots & \\ q_0 & q_0 & \cdots & q_1 & 1 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \\ & \vdots & & & \vdots & & \ddots \end{pmatrix} . \tag{E1}$$

Each block in this structure is an $m \times m$ matrix. All off-diagonal blocks consist of $Q_{\mu\nu} = q_0$. The diagonal block has $Q_{11} = 1$ and $Q_{\mu\nu} = q$ for $\mu \neq \nu$. The conjugate matrix $\hat{Q}_{\mu\nu}$ has a similar block structure. The order parameters R_μ

and $\hat{R}_\mu$ are symmetric at the saddle point, i.e., $R_\mu = R$ and $\hat{R}_\mu = \hat{R}$, as before.

Upon replacing Eqs. (D7)–(D10) by the RSB ansatz and taking the appropriate $n \to 0$ limit, the free energy (D6) is given by

$$-\beta f = G_0(q_0, q_1, \hat{q}_0, \hat{q}_1, R, \hat{R}, m) - \alpha G_r(q_0, q_1, R, m) \ , \tag{E2}$$

where

$$G_0 = \frac{1}{2}\{mq_0\hat{q}_0 + [(1-m)q_1 - 1]\hat{q}_1\} - R\hat{R} + \frac{1}{Nm}\int D\mathbf{z}_0 \ln \int D\mathbf{z}_1 \left(\int d\mu(\mathbf{W}) \exp(\mathbf{W}\cdot\mathbf{Z})\right)^m , \tag{E3}$$

$$\mathbf{Z} \equiv \mathbf{z}_0\sqrt{\hat{q}_0} + \mathbf{z}_1\sqrt{\hat{q}_1 - \hat{q}_0} + \mathbf{W}\hat{R}, \tag{E4}$$

$$G_r = -\frac{1}{m}\int Dy \int Dt_0 \ln \int Dt_1 \left[\int Dx \exp\left(-\frac{1}{2}\beta[g(t) - g(y)]^2\right)\right]^m , \tag{E5}$$

$$t \equiv t_0\sqrt{q_0 - R^2} + t_1\sqrt{q_1 - q_0} - yR \ . \tag{E6}$$

The free energy has to be minimized with respect to q_0, q_1, $\hat{q}_0$, $\hat{q}_1$, R, $\hat{R}$, and m. Note that after the $n \to 0$ limit has been taken the allowed range of m is $0 \le m \le 1$. Also, the physical meaning of q_0, q_1, and m is explained in Sec. V D 2 [see Eqs. (5.73)–(5.75)]. Upon substituting $m = 0$, the free energy f reduces to the RS result, Eq. (5.66).

Specializing to the case of Ising constrained weights in G_0 and Boolean output in G_1, we obtain

$$G_0 = \frac{1}{2}(mq_0\hat{q}_0 + [(1-m)q_1 - 1]\,\hat{q}_1) - R\hat{R} + \frac{1}{Nm}\sum_i \int Dz_0 \ln \int Dz_1 \left[2\cosh\left(z_0\sqrt{\hat{q}_0} + z_1\sqrt{\hat{q}_1 - \hat{q}_0} + W_i^0\hat{R}\right)\right]^m , \tag{E7}$$

$$G_r = -\frac{2}{m}\int_0^\infty Dy \int Dt_0 \ln \int Dt_1 \left[e^{-\beta} + (1 - e^{-\beta})H(\tau)\right]^m , \tag{E8}$$

$$\tau \equiv t_0\sqrt{\frac{q_0 - R^2}{1 - q_1}} + t_1\sqrt{\frac{q_1 - q_0}{1 - q_1}} - \frac{yR}{\sqrt{1 - q_1}} \ . \tag{E9}$$

We now search for a solution to the saddle-point equations at finite temperature with the property

$$q_1 = 1, \qquad \hat{q}_1 = \infty \ . \tag{E10}$$

According to Eq. (5.73), for such a solution each pure phase consists of a single (or a few almost identical) microscopic state. Thus the system collapses at finite temperatures to phases with zero entropy. This is not unlike the collapse of the system to the ground state at higher values of α (i.e., the perfect generalization state). However, it occurs in the metastable, spin-glass regime.

To find a solution with the property (E10) we take the limit $q_1 \to 1$, $\hat{q}_1 \to \infty$ of Eqs. (E7) and (E9) while keeping β finite. We obtain

$$G_0 = \frac{1}{m}[m^2\hat{q}_0(q_0 - 1) - mR\hat{R}] + \frac{1}{Nm}\sum_i \int Dz \ln 2\cosh\left(zm\sqrt{\hat{q}_0} + W_i^0 m\hat{R}\right) , \tag{E11}$$

$$G_r = -\frac{2}{m}\int_0^\infty Dy \int Dt \ln\left[e^{-\beta m} + (1 - e^{-\beta m})H(u)\right] , \tag{E12}$$

$$u \equiv t\sqrt{\frac{q_0 - R^2}{1 - q_0}} - \frac{yR}{\sqrt{1 - q_0}} \ . \tag{E13}$$

Comparing Eqs. (E11)–(E13) with Eq. (5.66) one obtains

$$f_{\mathrm{RSB}}(q_0, \hat{q}_0, R, \hat{R}, m, \beta) = \frac{1}{m} f_{\mathrm{RS}}(q_0, m^2\hat{q}_0, R, m\hat{R}, \beta m) \ . \tag{E14}$$

This structure is similar to the result of Krauth and Mézard [33] for the case of training a perceptron with random input-output mappings.

Stationarity with respect to q_0, $\hat{q}_0$, R and $\hat{R}$ results in

$$q_0(T, m, \alpha) = q_{\mathrm{RS}}(T/m, \alpha) \ , \tag{E15}$$

$$R(T, \alpha, m) = R_{\mathrm{RS}}(T/m, \alpha) \ . \tag{E16}$$

Finally, stationarity with respect to m yields

$$s_{\rm RS}(T/m,\alpha)=0\ , \tag{E17}$$

which implies

$$m(T,\alpha)=T/T_g(\alpha)\ , \tag{E18}$$

where $T_g(\alpha)$ is the $s=0$ line of the RS solution.

We have not attempted to search for other RSB solutions or to check the stability of this solution. However, this solution is probably exact both here and in the random perceptron problem of Krauth and Mézard. Similar completely frozen SG phases are known to exist in the REM, the "simplest spin glass" [54], and the large-P Potts glass. Recently, they have also been found in learning of random mappings by two-layer networks [60,62,63].

APPENDIX F: POWER COUNTING IN THE HIGH-TEMPERATURE EXPANSION

For the Boolean-mismatched model, the optimal generalization and zero entropy lines both follow power laws of the form $T\sim\alpha^r$, with $0<r<1$. The determination of this exponent requires the balancing of the two dominant terms in the asymptotic expansion. This is somewhat tricky, because determining which are the two dominant terms in turn depends on the exponent.

Since both T and α are diverging, it would seem natural to perform a double expansion around $(1/T,1/\alpha)=(0,0)$. Such an expansion is in fact ill defined, since the existence of a nontrivial high-temperature limit shows that the $T,\alpha\to\infty$ limit depends on the ratio T/α.

As discussed in Appendix B, the proper high-temperature expansion for a quantity A that is finite in the high-T limit is

$$A(\beta,T/\alpha)=\sum_{i=0}^{\infty}\beta^i A_i(T/\alpha)\ . \tag{F1}$$

If one proceeds to expand the A_i, one obtains a double expansion about $(1/T,T/\alpha)=(0,0)$

$$A(1/T,T/\alpha)=\sum_{a,b}A_{ab}\left(\frac{1}{T}\right)^a\left(\frac{T}{\alpha}\right)^b\ . \tag{F2}$$

This high-T expansion is the proper tool for investigating power laws of the form $T\sim\alpha^r$. We must only assume that $0<r<1$, so that both $1/T$ and T/α approach zero as $\alpha\to\infty$.

Assuming now that we have an expression for ϵ_g of the form (F2), let us find the power law $T_{\rm opt}\sim\alpha^r$ such that ϵ_g decreases at the fastest rate. In the series (F2) each term a,b scales like $1/\alpha^{ra+(1-r)b}$, so that the exponent is some weighted average of a and b. The power of the dominant term in the sum is $\min_i[ra_i+(1-r)b_i]$. This exponent must be maximized, to ensure the fastest decrease of ϵ_g.

The problem thus reduces to linear programming. Given a set of pairs (a_i,b_i), find the r that maximizes $\min_i[ra_i+(1-r)b_i]$. The problem looks difficult because there are an infinite number of pairs (a_i,b_i) to consider, but in fact most of the pairs can be eliminated from consideration. We define a partial ordering of the set of pairs: $(a_i,b_i)\le(a_j,b_j)$ means $a_i\le a_j$ and $b_i\le b_j$. A term (a,b) is *minimal* if $(a,b)\ge(c,d)$ implies $(a,b)=(c,d)$ for all (c,d). Only the subset of minimal pairs (which cannot be ordered) are relevant, because $(a_i,b_i)\le(a_j,b_j)$ implies $ra_i+(1-r)b_i\le ra_j+(1-r)b_j$. The linear programming problem thus only includes the finite minimal subset of pairs (a_i,b_i).

For the Boolean-mismatched model, we calculated the asymptotics of the generalization error as a series in $\alpha^{-1/2}$ at fixed T,

$$\epsilon_g(T,\alpha)-\epsilon_{\rm min}=\sum_{j=1}^{\infty}e_{j/2}(\beta)\alpha^{-j/2}\ . \tag{F3}$$

To convert this to the form (F2) of the high-T expansion, we expand the e_j in powers of β. Only the leading term from each e_j need be retained, since the higher-order terms are irrelevant. From the terms of order $\alpha^{-1/2}$ through α^{-2} we obtain the $(1,1/2)$, $(1,1)$, $(1,3/2)$, and $(0,2)$ terms in the high-T expansion (F2). From these four terms, we finally extract the minimal subset $(1,1/2)$ and $(0,2)$. Terms in (F3) of order $\alpha^{-5/2}$ and higher need not be considered, since they are all bounded below by $(0,2)$ and are thus irrelevant.

The maximin problem, which involves only $(1,1/2)$ and $(0,2)$, has $r=3/5$ as its solution, justifying the result $T_{\rm opt}\sim\alpha^{3/5}$ quoted in Eq. (6.54). The calculation of the zero-entropy line is similar, except that the minimal terms must add up to zero.

* Present address: Department of Computer Science, Hebrew University, Jerusalem 91904, Israel.

[1] R. P. Lippman, Neural Comput. **1** , 1 (1989).
[2] T. J. Sejnowski and C. Rosenberg, Complex Syst. **1**, 145 (1987).
[3] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, Neural Comput. **1**, 541 (1989).
[4] N. Qian and T. J. Sejnowski, J. Mol. Biol. **202**, 865 (1988).
[5] D. Zipser and R. A. Andersen, Nature **331**, 679 (1988).
[6] S. R. Lockery, G. Wittenberg, Jr., W. B. Kristan, and W. G. Cottrell, Nature **340**, 468 (1989).
[7] *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, edited by D. E. Rumelhart and J. L. McClelland (MIT Press, Cambridge, MA, 1986).
[8] T. Poggio and F. Girosi, Proc. IEEE **78**, 1481 (1990).
[9] L.G. Valiant, Commun. ACM **27**, 1134 (1984).
[10] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, J. ACM **36**, 929 (1989).
[11] E. Baum and D. Haussler, Neural Comput. **1**, 151 (1989).
[12] D. Haussler, Inf. Computation (to be published).
[13] S. Judd, J. Complexity **1**, 177 (1988).
[14] V. N. Vapnik and A. Y. Chervonenkis, Theory Probab.

Appl. **16**, 264 (1971).
[15] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data* (Springer-Verlag, New York, 1982).
[16] Y. S. Abu-Mostafa, Neural Comput. **1**, 312 (1989).
[17] A. R. Barron, in *Proceedings of the 28th Conference on Decision and Control* (IEEE Control Systems Society, New York, 1989), Vol. 1, pp. 280–285.
[18] L. Devroye, IEEE Trans. Pattern Anal. Mach. Intell. **10**, 530 (1988).
[19] P. Carnevali and S. Patarnello, Europhys. Lett. **4**, 1199 (1987).
[20] N. Tishby, E. Levin, and S. Solla, in *Proceedings of the International Joint Conference on Neural Networks* (IEEE, New York, 1989), Vol. 2, pp. 403–409.
[21] E. Levin, N. Tishby, and S. A. Solla, Proc. IEEE **78**, 1568 (1990).
[22] P. del Giudice, S. Franz, and M. A. Virasoro, J. Phys. (Paris) **50**, 121 (1989); D. Hansel and H. Sompolinsky, Europhys. Lett. **11**, 687 (1990).
[23] E. Gardner and B. Derrida, J. Phys. A **22**, 1983 (1989).
[24] G. Györgyi and N. Tishby, *Neural Networks and Spin Glasses*, edited by W. K. Theumann and R. Köberle (World Scientific, Singapore, 1990), pp. 3–36.
[25] G. Györgyi, Phys. Rev. Lett. **64**, 2957 (1990).
[26] W. Krauth, M. Mézard, and J.-P. Nadal, Complex Syst. **2**, 387 (1988).
[27] J. A. Hertz, in *Statistical Mechanics of Neural Networks: Proceedings of the Eleventh Sitges Conference*, edited by L. Garrido (Springer, Berlin, 1990), pp. 137–153.
[28] D. B. Schwartz, V. K. Samalam, J. S. Denker, and S. A. Solla, Neural Comput. **2**, 374 (1990).
[29] J. Shrager, T. Hogg, and B. A. Huberman, Science **242**, 414 (1988).
[30] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackel, and J. Hopfield, Complex Syst. **1**, 877 (1987).
[31] E.L. Thorndike, *Fundamentals of Learning* (Columbia University Press, New York, 1932).
[32] W. Kohler, *Gestalt Psychology* (Liveright, New York, 1929).
[33] W. Krauth and M. Mézard, J. Phys. (Paris) **50**, 3057 (1989).
[34] S. S. Venkatesh, in *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, edited by M.K. Warmuth and L.G. Valiant (Kaufmann, San Mateo, CA, 1991), pp. 257–266.
[35] H. Sompolinsky and N. Tishby, Europhys. Lett. **13**, 567 (1990).
[36] I. Kocher and R. Monasson (unpublished).
[37] H. Sompolinsky, N. Tishby, and H. S. Seung, Phys. Rev. Lett. **65**, 1683 (1990).
[38] K. Binder and D. W. Heerman, *Monte Carlo Simulation in Statistical Mechanics* (Springer-Verlag, Berlin, 1988).
[39] S. Kirkpatrick, Jr., C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).
[40] S. F. Edwards and P. W. Anderson, J. Phys. F **5**, 965 (1975).
[41] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).
[42] S. Kirkpatrick and D. Sherrington, Phys. Rev. B **17**, 4384 (1978).
[43] H. Sompolinsky, Phys. Rev. Lett. **47**, 935 (1981).
[44] K. Binder and A. P. Young, Rev. Mod. Phys. **58**, 801 (1986).
[45] M. Mézard, G. Parisi, and M.A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[46] F. Rosenblatt, *Principles of Neurodynamics* (Spartan, New York, 1962).
[47] J. R. L. de Almeida and D. J. Thouless, J. Phys. A **11**, 129 (1978).
[48] A. Krogh and J. A. Hertz, in *Advances in Neural Information Processing Systems*, edited by R. P. Lippman, J. E. Moody, and D.S. Touretzky (Kaufmann, San Mateo, CA, 1991), pp. 897-903; J. Phys. A (to be published).
[49] M. L. Minsky and S. Papert, *Perceptrons* (MIT Press, Cambridge, MA, 1988).
[50] M. Opper and D. Haussler, Phys. Rev. Lett. **66**, 2677 (1991).
[51] M. Opper and D. Haussler, in *Proceedings of the Fourth Annual Workshop on Computational Learning Theory* (Ref. [34]), pp. 75–87.
[52] E. Gardner, J. Phys. A **19**, L1047 (1986).
[53] G. Györgyi, Phys. Rev. A **41**, 7097 (1990).
[54] B. Derrida, Phys. Rev. B **24**, 2613 (1981).
[55] B. Derrida, R. B. Griffiths, and A. Prügel-Bennett, J. Phys. A **24**, 4907 (1991).
[56] J. F. Fontanari and R. Meir, Network **2**, 353 (1991).
[57] T. M. Cover, IEEE Trans. Electron. Comput. **14**, 326 (1965).
[58] E. Gardner, J. Phys. A **21**, 257 (1988).
[59] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[60] E. Barkai, D. Hansel, and I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990).
[61] J. F. Fontanari and R. Köberle, J. Phys. (Paris) **51**, 1403 (1990).
[62] E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992).
[63] E. Barkai and I. Kanter, Europhys. Lett. **14**, 107 (1991).