Neural networks optimally trained with noisy data

K. Y. Michael Wong* and David Sherrington†
Department of Physics, University of Oxford, 1 Keble Road, Oxford OX1 3NP, United Kingdom
(Received 2 December 1991; revised manuscript received 13 January 1993)

We study the retrieval behaviors of neural networks which are trained to optimize their performance for an ensemble of noisy example patterns. In particular, we consider (1) the performance overlap, which reflects the performance of the network in an operating condition identical to the training condition; (2) the storage overlap, which reflects the ability of the network to merely memorize the stored information; (3) the attractor overlap, which reflects the precision of retrieval for dilute feedback networks; and (4) the boundary overlap, which defines the boundary of the basin of attraction, and hence the associative ability for dilute feedback networks. We find that for sufficiently low training noise, the network optimizes its overall perforance by sacrificing the individual performance of a minority of patterns, resulting in a two-band distribution of the aligning fields. For a narrow range of storage level, the network loses and then regains its retrieval capability when the training noise level increases, and we interpret that this reentrant retrieval behavior is related to competing tendencies in structuring the basins of attraction for the stored patterns. Reentrant behavior is also observed in the space of synaptic interactions, in which the replica symmetric solution of the optimal network destabilizes and then restabilizes when the training noise level increases. We summarize these observations by picturing training noises as an instrument for widening the basins of attractions of the stored patterns at the expense of reducing the precision of retrieval.

I. INTRODUCTION

The intimate relation between the training and retrieving stages has been a subject of central interest in the statistical-mechanical study of neural networks. To build a network with optimal performance during retrieval, it is important to train it using the appropriate example patterns, performance criteria, and algorithms. Since the introduction of the Hopfield model using a Hebbian learning rule [1] and its statistical-mechanical analysis by Amit, Gutfreund, and Sompolinsky [2], various other learning rules and algorithms have been studied extensively, including the pseudoinverse [3,4], Adaline [5], perceptron [6], AdaTron [7], and "training with noise" [8] algorithms. These can be considered as attempts to devise efficient learning which optimizes certain aspects of retrieval performance, such as retrieval accuracy or robustness.

A unified view of these attempts can be provided by associating the dynamical processes of learning with the optimization of performance functions in the space of synaptic interactions [9–11]. For each kind of optimization issue with respect to a set of patterns to be stored, one may define an appropriate performance function of the synaptic weights. Learning can then proceed stepwise by a gradient-ascent process in the interaction space, which can be considered as a search for a configuration which optimizes the performance function [12]. This procedure is equivalent to the search for ground states of relevant Hamiltonians in many-body systems, and extends analogies used to minimize cost functions in some classic hard combinatorial optimization problems [13].

We have studied the performance optimization for noisy training and retrieving conditions [9,14]. Noises in neural networks may be present in the training data, the data to be retrieved, or the retrieval dynamics. In this paper, *training* noise refers to the random distortions of the examples, presented to the network during the training stage, with respect to the perfect data to be stored; *retrieving* noise refers to the disorder present during the retrieving stage in the input data for which the network is supposed to retrieve; and *thermal* noise refers to the stochastic noise present in the retrieving dynamics during the operation of the network. Retrieving noise may be present in data available to the network, but in feedback networks, it may also be caused, or amplified, by the thermal noise which distorts the output signals before they are fed back to the input ends of the neurons.

The study of noises in neural networks is interesting for the following reasons. First, there are many situations in the application of neural networks where only noisy data are available in the training stage. Algorithms which optimize the network performance for a set of perfect training patterns may not be suitable for noisy training patterns. For example, the perceptron learning rule [6] was devised to train a network using perfect training examples. It has been empirically adapted to process noisy training examples in the "training with noise" algorithm [8]. A convergence theorem can be proved for both cases, which states that the learning rule will converge to a network storing the examples, *provided* that it exists. However, for an arbitrary set of examples corrupted by noise, it is not clear whether the network corresponding to the convergence limit of the "training with noise" algorithm *does* exist. In fact, we have demonstrated that any network will inevitably retrieve the patterns

with errors if the noise present in a training ensemble is excessive [14]. It is therefore interesting to study the effects of the deliberate introduction of training noise on networks trained with algorithms which assume an otherwise perfect training set.

Noises in training are not necessarily destructive in all aspects. In the context of feedforward networks, it has been shown that training noises improve the generalization ability of a network in the so-called "teacher network" problem [15–17] and the "proximity" problem [18]. In the context of attractor networks, we have shown that training noises improve the associativity of the networks, i.e., they widen the basins of attraction of the stored patterns [14]. Furthermore, it has also been shown that training noises enhance the robustness of retrieval against thermal noise in the retrieving dynamics [9,19] and network damage [20]. To summarize, these performance measures (generalization, associativity, robustness against thermal noise and dilution) all involve the processing of noisy data during retrieval, and it is not surprising that training with noisy data can improve these performances. On the other hand, other performance measures such as the precision of the recall of the trained examples, or attractor overlaps, deteriorate with training noises, when these performances measures require the precision processing of clean input data. This trade-off in the various performance measures has led us to formulate the *principle of specialization,* namely, that networks which are optimal for one performance measure do not necessarily optimize others [9]. In fact, this principle is merely a manifestation of the fact that the optimal network depends on the choice of the performance function to be optimized. The improvement with the use of training noises in some aspects of the performance (despite the deterioration of the others) constitutes another motivation for a more detailed study of their effects.

Finally, the above observations have assisted us to devise strategies for optimal performance in the presence of retrieving noise, thermal noise, and indeed other factors affecting the environment of operation of the network [9]. This involves tuning the form of the performance function to be optimized and the noise level of the example patterns during the training stage of the network. The optimal strategy can be summarized by the *principle of adaptation,* which states that networks which perform optimally in any particular retrieving environment are those which optimize the same performance criteria in the same environment in training. Thus, for example, when thermal noises are present, the feedback network with optimal attractor overlap can be attained by setting the algebraic form of the training performance criterion corresponding to the same level of thermal noise in the updating dynamics, and setting the training noise level the same as the error level at the attractor. Similarly, the network with optimal associativity can be attained by setting the training noise the same as the error level at the boundary of the basin of attraction. Networks constructed by this strategy are called optimally adapted networks, for they can be trained, at least in principle, by the process of *self-adaptation.* In the self-adaptation of the feedback network with optimal attractor overlap, for example, training and retrieving are not separate processes. Instead one allows the synaptic weights of the network to be adiabatically adjusted to optimize the attractor overlap during the retrieval stage, using the retrieved signals of the network as the training ensemble. The characteristic of this process is that the "on-line" training is influenced by, and in turn influences, the retrieval performance of the network; in this sense it is similar to the adaptation of an organism in response to environmental conditions. The principle of adaptation can also be applied to optimize the robustness against random dilution of synapses [21]. Hence the study of training noise effects is an important problem with potential applications to a variety of retrieving conditions.

Indeed, the preliminary results presented in Ref. [14] demonstrated the significance of training noises, and the rich behavior on their variations. The introduction of an infinitesimal training noise level is sufficient to convert a marginally stable perceptron to a maximally stable one; we refer to a network of maximally stable perceptrons as the maximally stable network (MSN). Further increase in training noise level introduces errors in both the feedforward and attractor modes, but enhances the associativity of the network. In the case of dilute attractor neural networks storing random uncorrelated patterns, as we have already reported in Ref. [14], for $\alpha=0.5$ (i.e., 0.5 patterns per synapse) this enhancement in associativity is indicated by a transition of the overlap m_B at the basin boundary of attraction. For low training noise, m_B is nonzero and the network is said to be in the narrow retrieval regime; but when training noise increases, m_B vanishes and the wide retrieval regime sets in. The behaviors at other values of storage level have not been discussed explicitly, although it is clear that there are interesting transitions as the storage level and training noise are varied. For example, no narrow retrieval regime is present at any training noise levels for α below 0.42. On the other hand, the network is in the nonretrieval regime in the high-training-noise limit for α above 0.64; this can be seen from the fact that in the high-training-noise limit the simple Hebbian covariance rule performs optimally, and its storage capacity is 0.64. It is therefore interesting to map out the phase diagram in the space of storage level and training noise, and the extent of validity of the picture presented in Ref. [14]. Remarkably, there exists a narrow range of storage levels for which the network exhibits reentrant retrieval, which means that when the training noise level increases, the system loses, and then regains, its ability to retrieve the stored patterns.

Equally surprising are the regions of stability of the replica symmetric solution to the performance optimization in the interaction space [22]. For all storage levels below the optimal storage capacity (i.e., two patterns per synapse) there exist reentrant de Almeida–Thouless transitions [23], in which the replica symmetric solution destabilizes and then restabilizes when the training noise level increases. In other words, there are two separate regions of replica stability separated by a region of replica instability for intermediate-training-noise levels. Furthermore, the aligning field distributions in the two re-

gions are very different. In the low-training-noise region the distribution has two bands, whereas in the high-training-noise region it has one continuous band. Since it is generally believed that replica instability is a manifestation of the degree of frustration in the system, it appears that the network finds itself less frustrated when it settles in either a high- or a low-training-noise configuration in the interaction space. For intermediate-training-noise levels, strong competition is apparently present, resulting in the multiplicity of optima associated with replica symmetry breaking.

These observations are consistent with the qualitative picture that the network optimal to a high-training-noise level exhibits wide but imperfect basins of attraction associated with the stored patterns, as represented by the Hebbian network [24]. (In fact, the Hebbian covariance rule in its present mathematical form was first adopted in the Hopfield model [1], which subsequently stimulated the development of a whole class of related models, using the Hebbian rule or otherwise [2]; here we refer to the network trained with the Hebbian rule as the Hebbian, rather than the Hopfield, network for specificity.) Here the width of a basin of attraction refers to the range of neuronal states which will converge to the attractor, and its imperfectness refers to the quality of the retrieval. On the other hand, low-training-noise levels result in narrow but perfect basins as represented by the MSN. This accounts for the behavioral dependence on the training noise level. High generalization, associativity, low attractor overlaps, and low storage capacities are the features of networks with wide, imperfect basins, whereas low generalization, associativity, high attractor overlaps, and high storage capacities are associated with narrow, perfect basins. In terms of information retrieval, wide, imperfect basins correspond to retrieving using broad associations, whereas narrow, perfect basins correspond to retrieving using specific initial data.

This paper is organized as follows. In Sec. II we formulate the appropriate performance function and its optimization. In Sec. III we consider the aligning field distribution; all the performance measures considered in this paper are determined by this distribution. In Sec. IV we consider two performance measures which are relevant to retrieval in feedforward networks. They are the overlaps with the stored outputs when the input pattern, respectively, has the same noise level as the training noise, or is entirely free of noise. They will be called performance overlaps and storage overlaps, respectively. In Sec. V we discuss two other performance measures which are relevant to attractor networks. They are the overlaps with the stored patterns at the attractors and at the basin boundaries of attraction. They will be called attractor overlaps and boundary overlaps, respectively. We can then map out a phase diagram for the different dynamical behaviors in the space of storage level and training noise level, and describe the reentrant retrieval behavior. Preliminary results have already been presented in Ref. [25]. Since the dynamical equations for extensively connected networks are too complicated, we shall restrict ourselves to dilute attractor networks whose dynamics is solvable using iterative maps. We believe, however, that the picture is still qualitatively valid for networks with more general connectivity. In Sec. VI we consider the regions of replica symmetry breaking, and report on the reentrant de Almeida–Thouless transition. In Sec. VII we summarize and discuss the implications of our observations. The condition for replica symmetry breaking is derived in the Appendix.

II. FORMULATION

Learning in neural networks can be considered as an optimization process in the space of synaptic interactions. This may be done by defining an energy function $E(\{J_j\})$ in the space of interactions equal to the negative of an appropriate performance function. Learning can then be achieved by a dynamical process described by a Langevin equation

$$\frac{\partial J_j}{\partial t}=-\frac{\partial}{\partial J_j}E(\{J_j\})+\eta_j(t)\ , \tag{2.1}$$

where $\eta_j(t)$ is a noise term of zero mean and satisfies $\langle \eta_j(t)\eta_k(t')\rangle = 2T_{\rm an}\delta_{jk}\delta(t-t')$. The parameter $T_{\rm an}$ is called the annealing temperature, and is introduced into the dynamics of learning to prevent the network from being trapped in local minima. If one is interested in attaining the ground state of the system, one may tune the annealing temperature of this gradient-descent process to zero at a sufficiently slow rate. In terms of the performance function, this can be described as a gradient-ascent process searching for the optimal state. In fact, this process is equivalent to the computational method of simulated annealing applied to complex optimization problems.

Of course, learning can always be achieved in principle by simulated annealing, as in any problem of optimization. It may also be achievable by other methods, such as an extension of the method of Thouless, Anderson, and Palmer [26], introduced for the spin-glass problem, as further developed to the cavity method [22].

In this paper we shall concentrate on the asymptotic state of the learning process, assuming that the learning dynamics is carried out sufficiently carefully that the system does not get trapped in local minima, and consequently the equilibrium state can be described by a canonical ensemble corresponding to the temperature $T_{\rm an}$. By introducing a free energy corresponding to a thermodynamic average at temperature $T_{\rm an}$ and then taking the limit $T_{\rm an}\to 0$, one obtains the ground state, which is equivalent to the state of the maximum performance as defined by the energy function. In the statistical-mechanical approach, we are mainly interested in the so-called thermodynamic limit, in which the network consists of many neurons storing many patterns. In this limit it is sufficient to consider the averaged behavior for a set of random patterns obeying the same distribution, and the replica method [27] can be used to produce relevant results by facilitating the quenched averaging over the random patterns.

This approach has been adopted by Gardner and Derrida [12] to give the storage capacity of the maximally

stable network. There the performance function to be optimized is virtually the output overlap with the stored patterns corresponding to an input of the uncorrupted patterns. They found that not only can the patterns be *marginally* stabilized (i.e., the aligning fields of the stored patterns are merely bounded below by zero), but that they can be *maximally* stabilized (i.e., the aligning fields are bounded below by a positive stability parameter K, which depends on the storage level). The MSN ensures the strongest possible memory associativity without sacrificing the retrieval accuracy.

An alternative approach has been adopted by Gardner [28], where the volume, or the entropy, in the space of interactions stabilizing the patterns is calculated. The storage capacity is attained when this volume shrinks to zero. However, in the study of training noise, we are mainly interested in cases where patterns are only highly probable to be stabilized, for generic (i.e., not necessarily small) training noise levels prevent them from being completely stabilized. The volume in the interaction space stabilizing the patterns is therefore nonexistent, and the energetic (or canonical) approach turns out to be more suitable than the entropic (or microcanonical) approach.

In contrast to the above case of clean inputs, here we consider optimizing the output overlap when noisy versions of the patterns to be retrieved are presented. This optimization criterion tells us how best the network can perform when it is trained by noisy data. It is convenient to consider these inputs being chosen from an ensemble of noisy versions of the stored patterns, and we are interested in ensembles of very large sizes. The maximization of this noise-optimal performance function can be compared to the stepwise "training with noise" algorithm studied by Gardner, Stroud, and Wallace [8]. The algorithm is an adaptation of the perceptron learning rule, in which the clean patterns are presented to the network one at a time repeatedly, and the synaptic interactions are updated whenever the aligning field of a pattern is weaker than the stability parameter K. In the training with noise algorithm, the synaptic interactions are again updated according to the perceptron learning rule, but with the input patterns very slightly distorted by random noise.

Provided that a network configuration stabilizing the set of examples exists, it can be proved that these procedures ensure the network configuration converges to the solution. In this case, the result of the noise-optimal performance function is identical to that of the training with noise algorithm and this holds for infinitesimal training noise. However, for the generic training noise levels in which we are interested in this paper, the ensemble of noisy examples cannot be completely stabilized, and the network resultant from a corresponding training with noise algorithm with the same noise level may not be identical to that resultant from the training noise optimization. In fact, the performance function corresponding to the training with noise algorithm is not the output overlap of the noisy inputs, but is instead a function linear in the stability violation [10]. The noise-optimal network studied here can only act as giving an upper bound on the performance of the network retrieval quality of the training with noise procedure. Nevertheless, we believe that the effects of training noises are qualitatively the same.

To obtain meaningful results in the limit of infinite ensemble size, two mathematical approaches can now be adopted. In the first approach, the performance function is first averaged over the infinite training ensemble, and the network configuration optimizing this averaged performance function is sought. If this procedure is modeled by simulated annealing techniques, it involves evaluating the drop in the network energy (or the increase in the network performance), averaged over noisy examples according to their probability of occurrence in the training ensemble, before making a Monte Carlo move. This approach is adopted in our previous letter [14], and can be called an annealed optimization approach.

The second approach treats the examples in a training ensemble of fixed size as individual patterns. The network configuration optimizing the performance function of this training ensemble of fixed size is then sought. The network performance corresponding to a training ensemble of infinite size is then obtained by extrapolation. In terms of practical procedures, this involves optimizing the performance of a training ensemble of fixed size, and then performing a finite-size analysis to find the asymptotic behavior in the thermodynamic limit. This approach, which can be called quenched optimization, is similar to the work of Hansel and Sompolinsky [18], in which a training ensemble of large but finite size is considered. However, as we shall find, the results obtained by annealed or quenched optimization should be identical in the limit of an ensemble of infinite size. More specifically, the two results converge when the number of examples per pattern exceeds the inverse of the annealing temperature.

We now proceed to the mathematical formulation. Consider, storing in a network of N McCulloch-Pitts binary neurons, a set of p random patterns in the Ising representation, i.e., $\{\xi_j^\mu = \pm 1\}$ with $1 \le j \le N$ and $1 \le \mu \le p$. The network topology is fixed and only the synaptic strengths are modified, but our calculation applies to both a feedforward network with a single layer architecture, or an attractor network. The training ensemble consists of Q noisy examples for each pattern. This means that, for $1 \le \nu \le Q$, the probability distribution of the presented examples $\{R_j^{\mu\nu}\}$ is

$$P(R_j^{\mu\nu}) = \tfrac{1}{2}(1+m_t)\delta(R_j^{\mu\nu}-\xi_j^\mu) + \tfrac{1}{2}(1-m_t)\delta(R_j^{\mu\nu}+\xi_j^\mu) , \tag{2.2}$$

where we call m_t the training overlap, since it is the average of $R_j^{\mu\nu}\xi_j^\mu$ over the training ensemble. It is related to the training noise d_t by $m_t = 1-2d_t$. The retrieving stage of the network is deterministically described by outputting, at neutron i, the Ising bit S_i' with the same sign as the local field of the input state, i.e.,

$$S_i' = \mathrm{sgn}\left[\frac{1}{\sqrt{C}}\mathbf{J}_i\cdot\mathbf{S}_i\right] , \tag{2.3}$$

where C is the input connectivity of a neuron, and $\mathbf{S}_i$ represents the C-component input state feeding neuron i.

The noise-optimal performance function in the space of interactions J_{ij} is given, for the output at neuron i, by the overlaps of the output states S_i' with the stored outputs ξ_i^μ for all input states in the training ensemble

$$\sum_\mu g_{i\mu}=\frac{1}{Q}\sum_{\mu,\nu}\xi_i^\mu \mathrm{sgn}\left[\frac{1}{\sqrt{C}}\mathbf{J}_i\cdot\mathbf{R}_i^{\mu\nu}\right] . \tag{2.4}$$

Since the optimization on any one neuron is independent of the others, it is sufficient to consider the performance function defined on a single neuron, and subscripts i are hereafter implicit.

In the annealed optimization approach, the performance function is averaged in the limit Q approaching infinity. The averaged performance function is now the output overlap for noisy inputs drawn with the probability of occurrence according to (2.2). It can be written as

$$\sum_\mu g_\mu=\sum_{\mathbf{R}^\mu} P(\mathbf{R}^\mu)\mathrm{sgn}\left[\frac{1}{\sqrt{C}}\xi^\mu\mathbf{J}\cdot\mathbf{R}^\mu\right] . \tag{2.5}$$

This performance function can equivalently be interpreted as the average output overlap of neurons whose average input overlap with a nominated pattern is m_t. Its optimization thus corresponds to the update from overlap m_t in one time step. In the spherical model of the synaptic interactions, $\sum_j J_j^2=C$, and the argument in the sign function is a Gaussian variable of mean $m_t\Lambda^\mu$ and variance $1-m_t^2$, where $\Lambda^\mu\equiv\xi^\mu\mathbf{J}\cdot\xi^\mu/\sqrt{C}$ is the aligning field of the clean version of pattern μ [27–29]. The performance function is now reduced to $\sum_\mu g_{m_t}(\Lambda^\mu)$, where

$$g_{m_t}(\Lambda)=\mathrm{erf}\left[\frac{m_t\Lambda}{[2(1-m_t^2)]^{1/2}}\right] . \tag{2.6}$$

Using the replica method we have derived, in the replica symmetric ansatz, an optimization procedure for an arbitrary performance function g which is dependent on the aligning fields Λ^μ. This procedure of optimization has provided a unified perspective for various learning rules [9–11], and is explicitly derived in Appendix 1 of Ref. [9]. We summarize it as follows.

Consider optimizing the performance $\sum_\mu g(\Lambda_\mu)$. The averaged maximum performance per pattern is $\int d\Lambda\,\rho(\Lambda)g(\Lambda)$, where $\rho(\Lambda)$ is the aligning field distribution given by

$$\rho(\Lambda)=\int Dt\,\delta(\Lambda-\lambda(t)) , \tag{2.7}$$

where $Dt\equiv dt\,\exp(-t^2/2)/\sqrt{2\pi}$, and $\lambda(t)$ is the inverse function of $t(\lambda)$ defined by

$$t(\lambda)=\lambda-\gamma g'(\lambda) , \tag{2.8}$$

where γ is the interaction susceptibility given by $\gamma\equiv(\langle J_j^2\rangle_{T_{\mathrm{an}}}-\langle J_j\rangle_{T_{\mathrm{an}}}^2)/T_{\mathrm{an}}$, with $\langle\ \rangle_{T_{\mathrm{an}}}$ representing the thermodynamic average over the canonical ensemble in the interaction space at the annealing temperature T_{an}. It is determined by the condition

$$\int Dt[\lambda(t)-t]^2=\alpha^{-1} , \tag{2.9}$$

with $\alpha=p/C$ being the storage level. When the function $\lambda(t)$ is multivalued, we choose the argument which gives the largest value of $g(\lambda)-(\lambda-t)^2/2\gamma$. This is equivalent to discarding the range of argument $[\lambda_<,\lambda_>]$ given by the Maxwell construction

$$\int_{\lambda_<}^{\lambda_>} d\lambda\, t(\lambda)=t_0(\lambda_>-\lambda_<) , \tag{2.10}$$

where $t_0=t(\lambda_<)=t(\lambda_>)$. See Fig. 1.

To study the retrieval stage of this optimized network, we consider the output overlaps $f(m)$ with a stored pattern for an arbitrary input overlap m. Again, this mapping is determined by the aligning field distribution $\rho(\Lambda)$. Following the argument used in deriving (2.6), namely, that the local field for an input overlap m with pattern μ is a Gaussian variable with mean $m\Lambda\xi^\mu$ and variance $1-m^2$, we have

$$f_{m_t}(m)=\int d\Lambda\,\rho_{m_t}(\Lambda)g_m(\Lambda) , \tag{2.11}$$

where $g_m(\Lambda)$ is given by (2.6) with m_t replaced by m. Here we have explicitly included the subscript m_t to emphasize the dependence on both the training overlap m_t and the retrieval overlap m. Note that the output overlap reduces to the maximum performance function when m becomes m_t.

This completes the formulation of the annealed optimi-

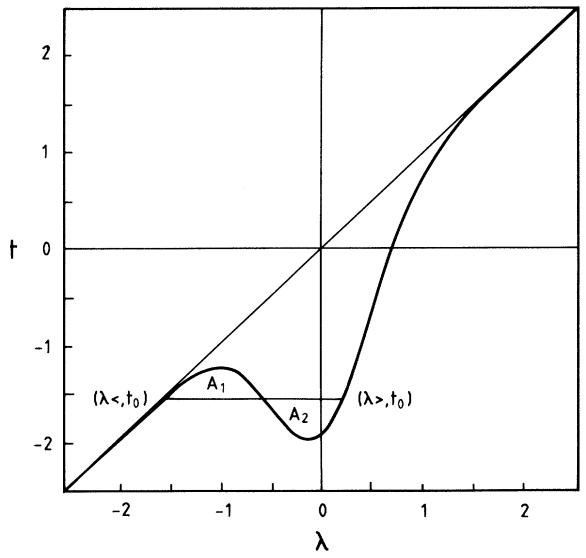

FIG. 1. The Maxwell construction for $\lambda(t)$. The points $(\lambda_>,t_0)$ and $(\lambda_<,t_0)$ are chosen such that the areas A_1 and A_2 are equal. The continuous full curve $t(\lambda)$ is given by (2.7), but the physical curve for $\lambda(t)$ has the discontinuity indicated. Here $\alpha=1.5$ and $m_t=0.9$.

zation procedure.

In the quenched optimization approach, we treat the Q examples of each pattern as individual patterns. Each particular choice of patterns and noisy examples gives, in general, a different performance function and therefore, at annealing temperature $T_{\rm an}$, a different free energy. The free energy is then quench averaged over the example patterns and the stored patterns in turn. The derivation proceeds as in the case of annealed optimization. As in Appendix 1 of Ref. [9], the pattern-averaged free energy is given by

$$\langle\langle \ln Z \rangle\rangle = \lim_{n\to 0} \frac{1}{n} \left\{ \mathrm{extr} \exp \left[C \left[-\sum_{\substack{\alpha,\beta \\ \alpha<\beta}} q_{\alpha\beta} F_{\alpha\beta} + G_J(\{E_\alpha\},\{F_{\alpha\beta}\}) + \alpha G_\Lambda(\{q_{\alpha\beta}\}) \right] \right] - 1 \right\}, \tag{2.12}$$

where G_J and G_Λ are given by

$$\exp G_J(\{E_\alpha\},\{F_{\alpha\beta}\}) = \prod_\alpha \int dJ_\alpha \exp\left[\sum_\alpha E_\alpha (1-J_\alpha^2) + \sum_{\substack{\alpha,\beta \\ \alpha<\beta}} F_{\alpha\beta} J_\alpha J_\beta \right], \tag{2.13}$$

$$\exp G_\Lambda(\{q_{\alpha\beta}\}) = \prod_\alpha \int \frac{d\lambda_\alpha dx_\alpha}{2\pi} \exp\left[\sum_\alpha i\lambda_\alpha x_\alpha - \sum_{\substack{\alpha,\beta \\ \alpha<\beta}} q_{\alpha\beta} x_\alpha x_\beta \right] \times \left\{ \prod_\sigma \int \frac{d\Delta_\sigma dy_\sigma}{2\pi} \exp\left[\sum_\sigma \frac{\beta_{\rm an}}{Q} g(\Delta_\sigma) + i\Delta_\sigma y_\sigma - i m_t y_\sigma \lambda_\sigma - (1-m_t^2) \sum_{\substack{\sigma,\tau \\ \sigma<\tau}} q_{\sigma\tau} y_\sigma y_\tau \right] \right\}^Q, \tag{2.14}$$

where $\beta_{\rm an} \equiv T_{\rm an}^{-1}$ and, in general, $g(\Delta)$ is an arbitrary performance function and in our case $g(\Delta)=\mathrm{sgn}(\Delta)$. In the replica symmetric ansatz, in which $q_{\alpha\beta}=q$ for all $\alpha\neq\beta$, the expression raised to power Q can be rewritten, after integrating over y_σ and taking the limit $\beta_{\rm an} \ll Q$, as

$$\exp\left\{ \sum_\alpha \int Dz \int \frac{d\Delta}{[2\pi(1-m_t^2)(1-q)]^{1/2}} \beta_{\rm an} g(\Delta) \times \exp\left[-\frac{\{\Delta - m_t\lambda_\alpha - [(1-m_t^2)q]^{1/2} z\}^2}{2(1-m_t^2)(1-q)} \right] \right\}, \tag{2.15}$$

which is further reduced in the low-temperature limit ($\beta_{\rm an}\to\infty, q\to 1$) to $\exp\{\sum_\alpha \beta_{\rm an} g_{\rm eff}(\lambda_\alpha)\}$ where

$$g_{\rm eff}(\Lambda) = \int Dz\, g(m_t\Lambda + (1-m_t^2)^{1/2} z) . \tag{2.16}$$

Substituting this result into (2.14), and comparing with (A1.5) of Ref. [9], we see that $g_{\rm eff}(\Lambda)$ is indeed the effective performance function in the present noisy training situation. Since the function g in the integrand of (2.16) is the sign function, the effective performance becomes identical to the performance function in (2.6). The subsequent procedure is the same as (2.6)–(2.11). Thus we have verified the equivalence of the annealed and quenched optimization approaches *provided* that the size of the training ensemble satisfies $Q \gg \beta_{\rm an}$. Physically, this means that the two approaches are indistinguishable, provided that $T_{\rm an} \gg Q^{-1}$, i.e., the optimization procedure cannot resolve the fluctuation caused by one further or lesser example per pattern in the training ensemble.

III. THE ALIGNING FIELD DISTRIBUTION

Having derived the basic results in (2.6)–(2.11), we now proceed to consider various cases. The performance measures considered in this paper depend on the aligning field distribution $\rho(\Lambda)$, which in turn depends on the inverted function $\lambda(t)$. Below we consider the effects of progressively decreasing the training overlap m_t. Figure 2 shows the field distribution $\rho(\Lambda)$ for different training overlaps.

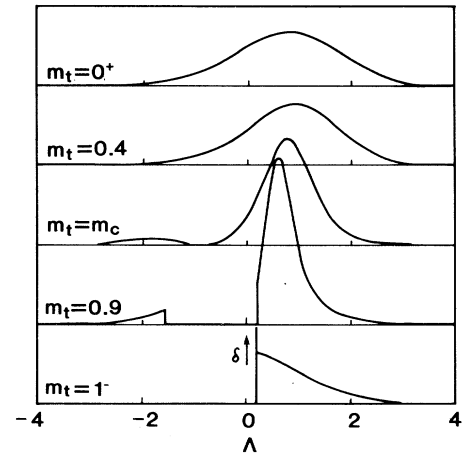

FIG. 2. The aligning field distribution $\rho(\Lambda)$ for different training overlaps from the MSN limit to the Hebbian network limit, at $\alpha=1.5$. Starting from the bottom, the curves correspond to $m_t=1^-$, 0.9, m_c, 0.4, and 0^+. (1^- denotes a number infinitesimally less than 1, and 0^+ denotes a number infinitesimally greater than 0.) Here $m_c=0.78$. The vertical scale for the individual curves is separated by 0.6 units. The curve for $m_t=1^-$ has a δ-function peak at $K(\alpha)=0.19$.

A. $m_t=1$

First consider the case with no training noise, i.e., $m_t=1$, when $g(\Lambda)$ in (2.6) becomes $\mathrm{sgn}(\Lambda)$. The inverted function $\lambda(t)$ then becomes

$$\lambda(t)=\begin{cases} t\ , & t\geq 0 \text{ or } t<-\sqrt{4\gamma} \\ 0\ , & 0>t\geq -\sqrt{4\gamma}\ . \end{cases} \tag{3.1}$$

γ is determined by the condition (2.9), which yields

$$\alpha^{-1}=\int_{-\sqrt{4\gamma}}^{0} Dt\, t^2\ . \tag{3.2}$$

The aligning field distribution is given by a normalized Gaussian, with the truncated region between 0 and $-\sqrt{4\gamma}$ replaced by a δ function of the same weight located at $\Lambda=0$, i.e.,

$$\rho_{m_t}(\Lambda)=[\Theta(\Lambda)+\Theta(-\sqrt{4\gamma}-\Lambda)]\frac{\exp(-\Lambda^2/2)}{\sqrt{2\pi}} + \left[\int_{-\sqrt{4\gamma}}^{0} Dt\right]\delta(\Lambda)\ , \tag{3.3}$$

where $\Theta(\Lambda)$ is the step function of its argument Λ. This shows that when γ is finite, there is a nonvanishing probability that a pattern cannot be stabilized, i.e., its aligning field is negative. In this case, the network optimizes its performance by having two bands in its aligning field distribution, so that those patterns which violate their prescribed outputs do so by the fullest extent of violation. Apparently, this is the consequence of the nature of the particular performance function, which is a sign function, for the network pays the same penalty so long as the aligning field of a pattern is negative, irrespective of whether the aligning field is weakly or strongly negative. It seems that by keeping the aligning field violations as *strong* as possible, the network manages to keep the violations as few as possible, so that the *overall* performance is maximized by this *sacrificial* effect.

Since the right-hand side of (3.2) is bounded above by $\frac{1}{2}$, the optimization calculation is valid only for α greater than 2. When α approaches 2, γ approaches infinity and the lower band of the aligning field distribution vanishes. The network becomes *marginally* stable because the aligning fields are merely bounded below by zero. This yields a storage capacity for errorless output at $\alpha_c=2$, agreeing with the results of Gardner and Derrida [12].

Below the storage capacity, no solution for (3.3) exists, and the optimization calculation is no longer valid. This is because we have assumed, in the derivation of (2.5)–(2.9), that the ground-state configuration is nondegenerate, so that the overlap order parameter q approaches 1 in the low-temperature limit. Below the storage capacity, the volume of the ground-state configuration remains finite and q remains below 1 [28].

B. $m_t=1^-$

Next we consider the case with an infinitesimal training noise, i.e., $m_t=1^-$. It is now convenient to introduce λ_1, defined by the relation

$$2\gamma\frac{\exp\{-[m_t^2/2(1-m_t^2)]\lambda_1^2\}}{[2\pi(1-m_t^2)/m_t^2]^{1/2}}=1\ , \tag{3.4}$$

so that the function $t(\lambda)$ can be written, in the limit that the training noise $d_t\equiv(1-m_t)/2\to 0$, as

$$t(\lambda)=\lambda-\exp\left[\frac{1}{8d_t}(\lambda_1^2-\lambda^2)\right]\ . \tag{3.5}$$

This function has the property that $t(\lambda)\approx\lambda$ for $\lambda^2>\lambda_1^2$, and $t(\lambda)$ grows abruptly to large, negative values at $\lambda=\pm\lambda_1$. By the Maxwell construction (2.10), the area A_1 in Fig. 1 now approaches a triangle with vertices $(\lambda_<,t_0)$, $(-\lambda_1,t_0)$, and $(-\lambda_1,-\lambda_1)$, and the area A_2 becomes a strip enclosed between $\lambda=\pm\lambda_1$. Furthermore, we shall justify *a posteriori* that $\lambda_1\sim O(1)$, so that the area of $A_2\sim\exp[\lambda_1^2/(8d_t)]\gg 1$, and we can assume that t_0 and $\lambda_<$ are large and negative. Thus (3.5) can be approximated to give

$$t_0\approx\lambda_<\ , \tag{3.6}$$

$$t_0=\lambda_>-\exp[(\lambda_1^2-\lambda_>^2)/(8d_t)]\ , \tag{3.7}$$

and the Maxwell construction (2.10) becomes

$$\tfrac{1}{2}(\lambda_>^2-\lambda_<^2)-2\gamma\approx t_0(\lambda_>-\lambda_<)\ . \tag{3.8}$$

From (3.6) and (3.8),

$$t_0=\lambda_<=\lambda_>-\sqrt{4\gamma}\ . \tag{3.9}$$

In the limit d_t approaching 0, (3.4) implies that $\gamma\sim\exp[\lambda_1^2/(8d_t)]$. Thus combining (3.4), (3.7), and (3.9), we can show that to within logarithmic corrections which are negligible in the limit of small d_t,

$$\lambda_>=\frac{\lambda_1}{\sqrt{2}}+O(d_t\ln d_t) \quad\text{and}\quad t_0=\lambda_<=\frac{\lambda_1}{\sqrt{2}}-\sqrt{4\gamma}\ . \tag{3.10}$$

It remains to determine the value of λ_1. Equation (3.7) entails that

$$\lambda(t)=\begin{cases} t\ , & t>\lambda_1 \text{ or } t<\lambda_< \\ \lambda_1\ , & \lambda_1>t\sim O(1)\ . \end{cases} \tag{3.11}$$

[For $\lambda_<<t<\lambda_1$, $\lambda(t)$ varies slowly from λ_1 to $\lambda_>$. However, as t moves away from λ_1 in this range, its contribution to the integral in (2.9) is rapidly fading, owing to the Gaussian prefactor. Thus to the lowest-order approximation in d_t, assigning $\lambda(t)$ to be λ_1 in this range is already sufficient.] Thus λ_1, and hence γ, can be derived from the condition (2.9), giving

$$\int_{-\infty}^{\lambda_1} Dt(\lambda_1-t)^2=\alpha^{-1}\ . \tag{3.12}$$

This justifies our previous approximation that $\lambda_1\sim O(1)$. Strikingly, this expression for λ_1 is identical to that for the maximal stability parameter $K(\alpha)$ introduced by Gardner [28]. She found that network configurations stabilizing the prescribed patterns exist, even when the lower bound of the aligning fields is as high as $K(\alpha)$. Furthermore, the stepwise perceptron learning rule can be adapted, so that synaptic updating proceeds whenever the aligning field is weaker than $K(\alpha)$, and a convergence theorem guarantees its converging to the target network

configuration.

In the case of the training with noise algorithm, convergence to the target network configuration is also guaranteed, if such a configuration storing the examples in the training ensemble exists. (We shall return to this condition of existence below.) In this case, by introducing an infinitesimally small training noise, the learning rule *automatically* results in a network with maximal stability after sufficiently long training. What is remarkable is that the stability requirement of $K(\alpha)$ need not be imposed at each training step; it simply "picks up" the stability $K(\alpha)$ by scanning through the ensemble of slightly noisy examples.

Thus the network changes from marginally stable to maximally stable when an infinitesimal training noise is introduced. This discontinuity of network behavior can be traced to a discontinuity of the training ensemble in the two cases. In the $m_t=1$ training ensemble, all the examples of a pattern are identical, and the network is merely adapted to the retrieval of clean patterns. On the other hand, the $m_t=1^-$ ensemble contains a full range of distinct noisy example patterns, each of whose probability of occurrence decreases with its Hamming distance from the clean patterns. The network is therefore adapted to the retrieval of noisy patterns, resulting in the maximal stability. It should be stressed, however, that this discontinuity in network behavior is dependent on the infinite size of the training ensemble, and an ensemble of finite size will smooth out the discontinuity.

The aligning field distribution now has two bands for all storage levels (not only for $\alpha>2$). The upper band is bounded below by $K(\alpha)/\sqrt{2}$, has negligible weight up to $K(\alpha)$, a very sharp peak at $K(\alpha)$, and essentially a Gaussian of mean 0 and width 1 above $K(\alpha)$. The lower band is a normalized Gaussian truncated at the upper bound $K(\alpha)-\sqrt{4\gamma}$ with $\gamma\to\infty$. Thus the distribution becomes indistinguishable with that for the MSN [29–31].

C. $m_t<1$

When m_t falls further below 1, the band gap in the aligning field distribution starts to narrow. The lower band has its weight increasing with training noise, and its upper bound shifts upwards.

In the upper band, the sharp peak in the neighborhood of $\Lambda=K$ degenerates into a broader, but still sharp peak. Despite the narrowing of the band gap, the lower bound of the upper band first shifts slightly upward before it eventually shifts downward with increasing training noise. In fact, if we expand (3.6)–(3.10) to the next higher order in d_t, we see that while $\lambda_<\sim -O(\exp(K^2/(16d_t))$ accounts for the narrowing of the band gap, $\lambda_>$ is given by

$$\lambda_>=\frac{K}{\sqrt{2}}\left[1+\frac{2}{K^2}d_t|\ln d_t|\right], \tag{3.13}$$

where K is still given by λ_1 in (3.12). However, the initial upward shift of $\lambda_>$ is only very small. For example, when $\alpha=0.7$, $\lambda_>$ increases from 0.53 to a maximum of 0.54 when m_t drops from 1 to 0.98, and then decreases on dropping m_t thereafter.

The change in the shape of the distribution function when the training noise changes can be considered as a manifestation of the sacrificial effect. When the sigmoidal performance function (2.6) smoothens on increasing training noise, the network manages to maximize the performance by increasing the weight of the lower band at the far negative end. By making this sacrifice, the sharp peak in the upper band manages to broaden itself in a region where the performance function has a greater slope. In the neighborhood of the peak position λ_1,

$$\rho_{m_t}(\Lambda)=t'(\Lambda)\frac{\exp[-t(\Lambda)^2/2]}{\sqrt{2\pi}} \sim\exp\left\{-\frac{\lambda_1}{4d_t}(\Lambda-\lambda_1) -\frac{1}{2}\left[\lambda_1-\exp\left[-\frac{\lambda_1}{4d_t}(\Lambda-\lambda_1)\right]\right]^2\right\} \tag{3.14}$$

by virtue of (2.8) and (3.5). This shows that the distribution drops much more drastically on the side $\Lambda<\lambda_1$ than $\Lambda>\lambda_1$. This asymmetric broadening allows the weight of the distribution function to shift more positive and perform better than the field distributions of lower training noises, which resemble more the δ-truncated Gaussian of the MSN.

This sacrificial mechanism enables us to improve the performance for noisy retrieval inputs. The same kind of observation has prompted a number of authors to look for performance improvement above saturation [19,10] although their performance functions do not have an optimal form. In fact, we have formulated a principle of adaptation [9], stating that the network performs best for a retrieval noise identical to the training noise.

A further reduction in m_t results in a narrowing of the band gap and a smoothening of the peak. At $m_t=m_c$ the two bands merge. This corresponds to the coalescence of $\lambda_>$ and $\lambda_<$ in the Maxwell construction (2.10). Thus we have

$$t'(\lambda_c)=t''(\lambda_c)=0\ , \tag{3.15}$$

where $\lambda_>=\lambda_<=\lambda_c$. This reduces to

$$\lambda_c=-\left[\frac{(1-m_c^2)^{1/2}}{m_c}\right]^{1/2}, \qquad \gamma=\left[\frac{\pi e}{2}\right]^{1/2}\left[\frac{1-m_c^2}{m_c^2}\right]. \tag{3.16}$$

Substituting into the condition (2.9), we obtain a relation between m_c and α,

$$\int\frac{d\lambda}{\sqrt{2\pi}}e^{-t_c(\lambda)^2/2}t_c'(\lambda)[\lambda-t_c(\lambda)]^2=\alpha^{-1}\ , \tag{3.17}$$

where $t_c(\lambda)=\lambda+\lambda_c\exp[(1-\lambda^2/\lambda_c^2)/2]$.

In the extremely noisy limit, $m_t\to 0$, the performance

function becomes linear in the aligning field Λ. The inverted function $\lambda(t)$ then becomes

$$\lambda(t)=t+\frac{1}{\sqrt{\alpha}}\ , \tag{3.18}$$

and the aligning field distribution becomes

$$\rho_{m_t}(\Lambda)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left[\Lambda-\frac{1}{\sqrt{\alpha}}\right]^2\right]\ , \tag{3.19}$$

which is a normalized Gaussian of mean $1/\sqrt{\alpha}$. This distribution coincides with that of the Hebbian rule with

$$J_{ij}=\frac{1}{\sqrt{\alpha C}}\sum_{\mu}\xi_i^{\mu}\xi_j^{\mu}\ . \tag{3.20}$$

The network optimal in the high-training-noise limit is necessarily the Hebbian network. Our studies in Boolean networks [32] showed that the Hebbian network minimizes the output error among all Boolean networks in the high-training-noise limit. Since the set of synaptic networks is merely a subset of Boolean networks, the Hebbian network must also minimize the output error among all synaptic networks in that limit, which is indeed the present result.

IV. PERFORMANCE AND STORAGE OVERLAPS

The first performance measure of interest is $f_{m_t}(m_t)$, i.e., the output overlap with a stored pattern when, during retrieval, the input state obeys the same statistics as the training ensemble. We call this the *performance overlap* since this is precisely the function which is maximized in the optimization procedure. In feedforward networks, this performance measure is related to the generalization ability of the network. It is also the training measure if the training ensemble is of infinite size. An example of this application is given in the proximity problem considered by Hansel and Sompolinsky [18], who also used a noisy training stage.

In attractor networks, the performance overlap merely describes the behavior in one iteration, and is therefore not directly relevant to the attractor performance. However, the formulation of the principle of adaptation [9] has brought a new dimension to the performance overlap in dilute attractor networks. Since the optimal performance is attained at identical training and retrieving overlaps, the performance overlap $f_{m_t}(m_t)$ as a function of m_t becomes an envelope of the retrieval functions of all network configurations. Consequently, the stable fixed points of this retrieval envelope become, in the dilute attractor network, the best attractor overlaps attainable by any network configuration, and the unstable fixed points delimit the widest basin boundary of attraction. Furthermore, retrieval envelopes describe the dynamics of some *self-adaptive* processes, in the same way that individual retrieval functions describe the dynamics of retrieval processes. Self-adaptation is a process in which the output states of the network are fed back to the input ends of the neurons, to act as input states for further retrieval, as well as training examples for further adiabatic modification of the synaptic weights, so that the network optimizes its performance in the environment generated by its own output at the retrieval attractor. Our study of retrieval envelopes [9] points to the possibility of such processes, although practical adaptive algorithms have not been studied in detail.

In contrast to the learning of perfect patterns, in which the resultant performance can be errorless below the storage capacity, training with a noisy ensemble usually leads to imperfect outputs. To estimate the amount of training noise required to cause disruption in the performance overlap, we note that in the low-training-noise limit, the performance overlap $f_{m_t}(m_t)$ is given by (2.11), where the function $\lambda(t)$ is given by (3.11), and $\lambda_1=K(\alpha)$, so that

$$f_{m_t}(m_t)\sim 1-O\left[\exp\left[-\frac{K(\alpha)^2}{8d_t}\right]\right]\ . \tag{4.1}$$

For attractor neural networks, errorless retrieval of a pattern in one iteration at all the N output nodes is possible only up to a training noise which causes the performance overlap to drop from 1 by an amount of the order N^{-1}. This implies that the maximum training noise for errorless performance is given by

$$d_t=\frac{K(\alpha)^2}{8\ln N}\ . \tag{4.2}$$

Any training noise higher than this, and in particular any generic training noise of the order N^0, will inevitably cause error in the performance measure. Similarly, for feedforward neural networks with only one output node, errorless retrieval of all the p patterns is possible only up to a training noise given by $d_t=K(\alpha)^2/(8\ln p)$. Figure 3 shows the dependence of the performance overlap on training noise of the order N^0. It decreases from 1 to 0 when the training overlap decreases from 1 to 0.

Another performance measure of interest is $f_{m_t}(1)$, i.e.,

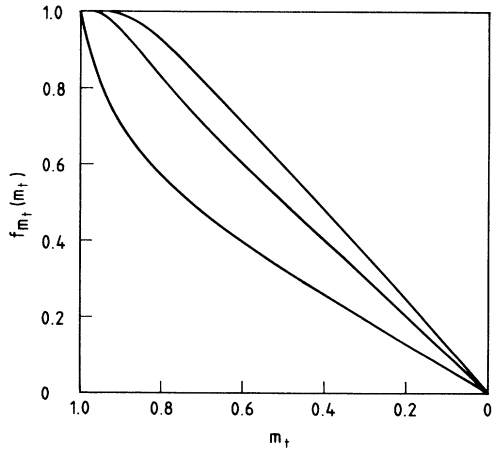

FIG. 3. The dependence of the performance overlap $f_{m_t}(m_t)$ on the training overlap m_t for $\alpha=0.4, 0.6, 1.6$ (from top to bottom). Note that in Figs. 3, 4, and 6, m_t is plotted in reverse direction to illustrate the effects of increasing training noise d_t.

the output overlap with a stored pattern when, during retrieval, a perfect input pattern is presented. We call this the *storage overlap*. It measures the extent to which a perfect pattern can be recovered during retrieval, when the network is trained on an ensemble of noisy patterns without the perfect input pattern being presented deliberately. In fact, this ability to recover a perfect pattern from an ensemble of partially similar patterns has been observed in neural networks and is called spontaneous generalization [33]. In feedforward networks, this is related to the rule extraction ability of the networks by filtering out the noises in the training stage.

Again, errorless retrieval is not possible for generic training noises of the order N^0. However, when compared with the performance overlap, the restriction on training noise is less stringent, since there is no error in the input for the case of storage overlap. To estimate the critical amount of training noise, we note that, in the low-training-noise limit, the storage overlap $f_{m_t}(1)$ is given via (2.11) by

$$f_{m_t}(1)=\mathrm{erf}(-t_0/\sqrt{2}) . \tag{4.3}$$

Using (3.10), we obtain $-t_0 \sim \exp[K(\alpha)^2/(8d_t)]$, which further implies

$$f_{m_t}(1) \sim 1-O\left[\exp\left[-\exp\left[\frac{K(\alpha)^2}{8d_t}\right]\right]\right] . \tag{4.4}$$

For attractor networks, retrieval errors become inevitable when $f_{m_t}(1) \sim 1-O(N^{-1})$. This yields the maximum training noise for errorless storage after one step in attractor neural networks as

$$d_t=\frac{K(\alpha)^2}{8 \ln \ln N} , \tag{4.5}$$

whereas the training noise for errorless storage in feedforward neural networks is $d_t=K(\alpha)^2/(8 \ln \ln p)$. The restriction on training noise reveals a disadvantage of the training with noise scheme, when compared with the maximally stable perceptron algorithm, which has been proposed as an alternative method to enhance memory associativity. However, because of the double logarithmic dependence on N or p, this restriction is not too stringent.

For a general training noise, the storage overlap $f_{m_t}(1)$ is given by the averaging of the sign function of the aligning fields, by virtue of (2.11). Inspecting the aligning field distribution determined by (2.7) and (2.8), we see that when $\lambda_>$ is positive, the aligning field as a function of t is positive for $t > t_0$, and negative otherwise. On the other hand, when $\lambda_>$ is negative, the aligning field is positive for $t > t(0)$, and negative otherwise. Here $t(0)=-2\gamma/[2\pi(1-m_t^2)/m_t^2]^{1/2}$ from (2.8). Thus

$$f_{m_t}(1)=\mathrm{erf}\left[\frac{1}{\sqrt{2}}\min(-t_0,-t(0))\right] . \tag{4.6}$$

Figure 4 shows the dependence of the storage overlap $f_{m_t}(1)$ on training noise of the order N^0. It always remains above the performance overlap $f_{m_t}(m_t)$, and decreases from 1 to a nonzero value in the range of training overlaps between 1 and 0. We also notice the presence of a kink in the storage overlap curve. This kink signifies that the bound $\lambda_>$ of the upper band of the aligning field distribution passes from positive to negative. The condition for the occurrence of the kink is therefore

$$t_0=t(0) . \tag{4.7}$$

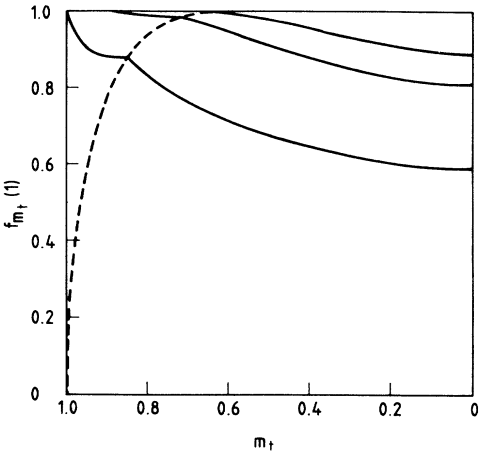

FIG. 4. The dependence of the storage overlap $f_{m_t}(1)$ on the training overlap m_t for $\alpha=0.4, 0.6, 1.5$ (from top to bottom). The dashed line shows the position of the kink when α is varied.

V. ATTRACTOR AND BOUNDARY OVERLAPS

The dynamics of attractor neural networks becomes greatly simplified in networks with dilute connectivity, i.e., $1 \ll \ln C \ll \ln N$. In this case, if the network state has a macroscopic overlap m with only one of the stored patterns, then the dynamics of the dilute network is completely determined by the iterative *retrieval mapping* [24]. This mapping relates the output overlap to the input overlap, and is given by $f_{m_t}(m)$ in (2.11). For parallel dynamics

$$m(t+1)=f_{m_t}(m(t)) , \tag{5.1}$$

and for random sequential dynamics

$$\frac{dm(t)}{dt}=f_{m_t}(m(t))-m(t) . \tag{5.2}$$

In both kinds of dynamics, the attractor overlap corresponds to the stable fixed points of the retrieval mapping

$$m^*=f_{m_t}(m^*) , \tag{5.3}$$

and the basin boundary of the attractor is determined by its unstable fixed points

$$m_B=f_{m_t}(m_B) . \tag{5.4}$$

The storage capacity of the network is reached when the stable fixed point m^* corresponding to pattern retrieval coalesces with one of the unstable fixed points m_B as the

storage level is varied. In our previous work [14] we have found the typical situation in which an increase in training noise causes the attractor overlap m^* to deteriorate, but the basin boundary m_B to decrease, signaling a widening of the basins of attraction, or alternatively an increase in associativity. For sufficiently high training noise, the boundary overlap m_B drops to zero, and the system is said to undergo a transition from the *narrow retrieval* phase to the *wide retrieval* phase. This essentially describes the situation for sufficiently low storage.

Subsequent studies have revealed some intriguingly surprisingly results when the storage level is raised beyond the low storage regime. Algebraically this novel behavior can be traced to the multiple fixed points of the performance overlap $f_m(m)$ for the storage level α between 0.599 and 0.637, when the input overlap m becomes identical to the training overlap m_t. In fact, the existence of multiple fixed points can be deduced from series expansion and continuity arguments.

First we know that for $\alpha=0.637$, the MSN, corresponding to $m_t=1^-$, is in the narrow retrieval regime [31]. By the principle of adaptation, the performance overlap $f_m(m)$, is an envelope for all possible retrieval mappings [9]. Therefore $m=1$ should also be a stable fixed point of the performance overlap $f_m(m)$.

Secondly, series expansion of the performance overlap at $m=0$ yields

$$f_m(m)=\left[\frac{2}{\pi\alpha}\right]^{1/2}\left[m-\frac{m^3}{6\alpha}\right], \tag{5.5}$$

implying that it is convex at $m=0$ for all values of α. We note in passing that this behavior is already very different from the retrieval mapping of the MSN [19] for which the third-order term changes sign at the tricritical point. This precludes the phase diagram involving the performance overlap from having the same structure which is indeed confirmed in Ref. [9].

When α drops slightly below $\sqrt{2/\pi}=0.637$, the convex region of the performance overlap in the immediate neighborhood of $m=0$ lies above the diagonal line $f_m(m)=m$, whereas the region of $f_m(m)$ slightly beyond this neighborhood lies below the diagonal. Since the curve must lie above the diagonal line again in the neighborhood of $m=1$, it follows that the performance overlap must have at least two convex regions separated by at least one concave region. Numerical results confirm that for α between 0.599 and 0.637, the performance overlap has two convex regions separated by one concave region, and it has two stable and two unstable fixed points for $0\le m\le 1$, as shown in Fig. 5.

The relative depression of the performance overlap for intermediate values of training overlap is a manifestation of the difference between networks trained with high and low noises in the examples. In the high-training-noise network, retrieval is achieved from within broad basins with only imprecise asymptotic overlaps, whereas in the low-training-noise network, narrower basins are associated with more precise retrieval. In the intermediate region, the crossover between the two tendencies apparently cause the depression in performance.

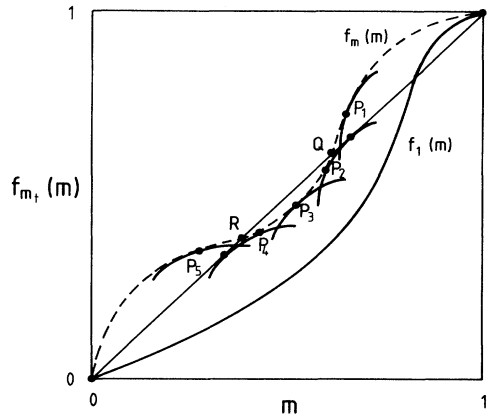

FIG. 5. A schematic plot of the performance overlap $f_m(m)$ having two stable and two unstable fixed points for $0\le m\le 1$. The retrieval mapping $f_1(m)$ of the MSN, which has only one unstable fixed point, is also shown for comparison. The individual retrieval mappings, touching the envelope curve $f_m(m)$ at P_1 to P_5, respectively, show the transition from retrieval to nonretrieval and back to retrieval when the training overlap is reduced. The subregion of nonretrieval spans from P_2 to P_4, and lies within the region of performance depression spanning from Q to R.

A direct consequence of this medial depression is that for a constant α, the attractor and boundary overlaps cannot vary continuously from the maximally stable limit to the Hebbian limit as the training overlap is decreased continuously. This is because the attractor and boundary overlaps must lie on the intersection of the diagonal line and the retrieval mapping for each m, which is bounded above by the performance overlap.

This discontinuity is a prerequisite for the reentrant retrieval behavior discussed here. This means that for constant α between 0.599 and 0.637, there is an intermediate range of training overlap for which the system is in the nonretrieval phase, whereas for higher and lower training overlaps retrieval is possible. As illustrated in Fig. 5, the set of individual retrieval mappings for the particular training overlaps m_t is enveloped by the performance overlap curve at $m=m_t$, by virtue of the principle of adaptation. Thus in the region where the performance overlap is depressed below the diagonal line, there is a subregion for which retrieval is not possible. Note, however, that this nonretrieval subregion is smaller than the depressed region, for individual retrieval mappings with m_t a little within it are still able to intersect the diagonal line, but with stable fixed points outside the depressed region.

Figures 6(a)–6(f) show the dependence of the attractor and boundary overlaps on training overlap for different storage levels. In Fig. 6(a), training noises disrupt the stored patterns and reduce the attractor overlap, as found previously [14]. On the other hand, the storage is sufficiently low that the boundary overlap is always zero. This means that the network is always in the wide retrieval region. Assuming that training noises enhance network associativity as found previously [14], the MSN,

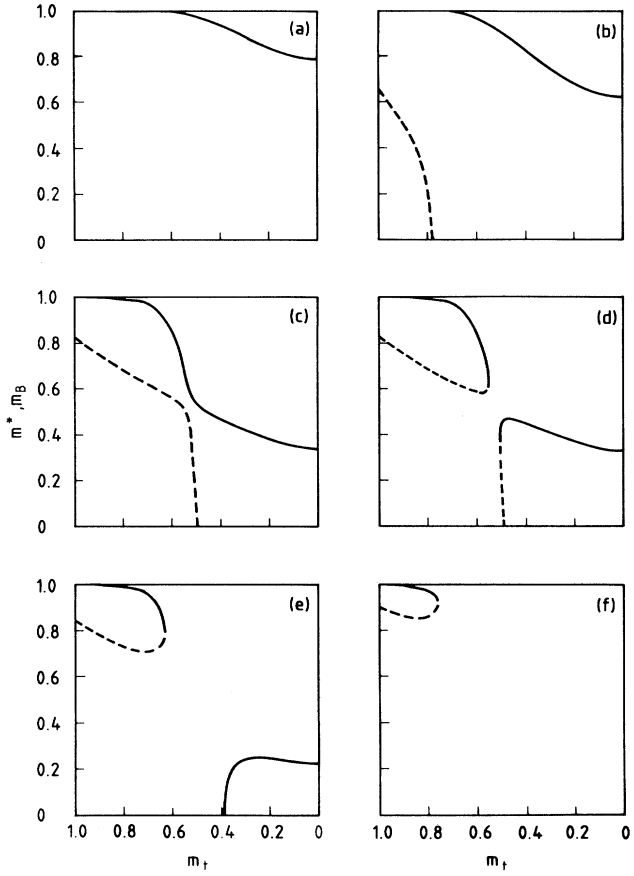

FIG. 6. The dependence of the attractor and boundary overlaps (solid and dashed lines, respectively) on the training overlap for $\alpha=$ (a) 0.4, (b) 0.5, (c) 0.598, (d) 0.6, (e) 0.62, (f) 0.7.

corresponding to a training overlap of $m_t=1^-$, is supposed to have the least associativity. Since even the MSN has wide retrieval basins for α below 0.42 [31], this scenario of wide retrieval basins for all training overlaps extends up to the storage level 0.42.

Figure 6(b) has been shown in Ref. [13]. Here the storage level is above 0.42 but still sufficiently low. The attractor overlap decreases with training noise, and the basin of attraction changes from the narrow retrieval phase of the MSN to the wide retrieval phase of higher training noise.

In Fig. 6(c), the storage level becomes close to 0.599, and the attractor and boundary overlaps approach the value 0.53 at the training overlap of the same value. At this point, the envelope curve $f_m(m)$ becomes very close to the diagonal line, and the curves of m^* and m_B versus m_t can be approximated by a hyperbola. To see this, we consider series expansion around the bifurcation point $(\alpha_p,m_p)=(0.599,0.53)$ of the envelope curve $f_m(m)$ where it touches the diagonal line. At this point,

$$f_{m_p}(m_p)=m_p \quad \text{and} \quad \frac{d}{dm}f_m(m)\bigg|_p=1 \ . \tag{5.6}$$

Furthermore, the principle of adaptation implies that

$$\frac{\partial}{\partial m_t}f_{m_t}(m_r)\bigg|_p=0 \ , \tag{5.7}$$

which, on combining with (5.6), leads also to

$$\frac{\partial}{\partial m_r}f_{m_t}(m_r)\bigg|_p=1 \ . \tag{5.8}$$

Now consider deviations in the training overlap, retrieval overlap, and storage levels, respectively, denoted by $\epsilon_t=m_t-m_p$, $\epsilon_r=m_r-m_p$, $\epsilon_\alpha=\alpha-\alpha_p$. Expanding the fixed-point equation $f_m(m)=m$, we obtain

$$\frac{1}{2}\frac{\partial^2 f}{\partial m_t^2}\epsilon_t^2+\frac{\partial^2 f}{\partial m_t \partial m_r}\epsilon_t\epsilon_r+\frac{1}{2}\frac{\partial^2 f}{\partial m_r^2}\epsilon_r^2=-\frac{\partial f}{\partial \alpha}\epsilon_\alpha \ , \tag{5.9}$$

which corresponds to a hyperbolic equation.

At $\alpha=0.599$, the envelope curve $f_m(m)$ touches the diagonal line. Near the critical point, the curves of m^* and m_B versus m_t become a pair of straight lines, with m^* always above m_B. Thus both overlaps exhibit a discontinuity in slope.

In Fig. 6(d) where α is slightly greater than 0.599, the curves of m^* and m_B versus m_t become hyperbolas with the gap opening in the conjugate direction. Reentrant retrieval behavior appears. At low training noise, the stored patterns have narrow basins. As the training noise increases, the attractor and boundary overlaps coalesce and the patterns are no longer retrievable. But as the training noise further increases, a pair of stable and unstable fixed points appears again and bifurcates. Thus the attractors of the stored patterns appear again with narrow basins of attraction, which become wide basins on further increase of training noise.

Figure 6(e) shows a similar reentrant behavior at a higher storage level, but the reentrant retrieval phase has a wide basin of attraction for all the training noise levels where it exists. The nonretreival gap widens, and the extent of the reentrant retrieval phase shrinks with increasing storage level.

For storage levels above 0.637, even the limit of the Hebbian network, corresponding to extremely high training noise, cannot sustain the reentrant retrieval phase. Thus the reentrant phase disappears altogether. Narrow retrieval is possible at low training noise, and no retrieval is possible at high training noise, as illustrated in Fig. 6(f).

The fixed-point curves in Figs. 6(a)–6(f) further describe the self-adaptation processes proposed in Ref. [9]. In self-adaptation, the performance overlap is considered to be an envelope of retrieval mappings. Its stable fixed points give the networks, or retrievers, with optimal attractor overlap attained by processes of self-adaptation. We have found that for α between 0.599 and 0.637, a strong retriever with a higher attractor overlap coexists with a weak retriever with a lower attractor overlap. The strong retriever has the globally maximum attractor overlap, and is given by the MSN at zero temperature. The weak retriever has a locally maximum attractor overlap. From Figs. 6(d) and 6(e), this local maximum is

simply the maximum of the reentrant retrieval phase. Similarly, if we consider maximizing the basins of attraction, we have a wide retriever coexisting with a narrow one. The wide retriever is the Hebbian network, and the narrow retriever simply corresponds to the minimum m_B in the retrieval branch of higher training overlap. Furthermore, the maxima and minima of the fixed-point curves all lie on the diagonal line on which training overlap equals retrieval overlap, by virtue of the principle of adaptation.

We specify that self-adaptation involves the network state first locating a retrieval attractor, and then adiabatically optimizing its performance at the attractor. Then it is obvious that for α between 0.599 and 0.637, the range of training overlaps in the principal retrieval branch defines one basin of adaptation, whereas those in the reentrant retrieval branch define another. For the intermediate range of training overlaps, no retrieval is possible, and hence they do not belong to either basin. For lower α the network has a wide basin of adaptation, and for higher α it has a narrow basin of adaptaton. (Alternatively, if the process of self-adaptation merely involves the network state adiabatically optimizing its performance at the output overlap, i.e., locating the attractor is not a necessary step, then the basins of adaptation are simply delimited by the unstable fixed points of the retrieval envelope.)

Figure 7 summarizes the various retrieval phases in the space of m_t and α. The wide retrieval phase is the region where $m_B=0$. It vanishes when the fixed point $m=0$ of the individual retrieval map $f_{m_t}(m)$ changes stability. The phase boundary is therefore given by

$$f'_{m_t}(0)=1 \ . \tag{5.10}$$

Depending on whether the retrieval map is convex or concave at $m=0$, the wide retrieval phase may undergo a transition to either the nonretrieval or narrow retrieval phase, respectively. Since the retrieval map is an odd function in m, as is evident from (2.11), this is determined by the third derivative. This gives rise to a tricritical point

$$f'_{m_t}(0)=1 \quad \text{and} \quad f'''_{m_t}(0)=0 \ , \tag{5.11}$$

which is located at $(\alpha,m_t)=(0.604,0.48)$.

For $m_t<0.48$, the wide retrieval phase becomes nonretrieval on increasing α, but for $m_t>0.48$, the wide retrieval phase first becomes narrow retrieval. The phase transition is continuous, since a nonzero fixed point either merges with or bifurcates from the fixed point at $m=0$.

The narrow retrieval phase exists up to a higher α when the nonzero stable and unstable fixed points merge and disappear. The network then undergoes a discontinuous transition to the nonretrieval phase. The phase line separating the regions of narrow retrieval and nonretrieval is therefore given by

$$f_{m_t}(m)=m \quad \text{and} \quad f'_{m_t}(m)=1 \ . \tag{5.12}$$

It meets the line of continuous transition at the tricritical point with a common slope [19].

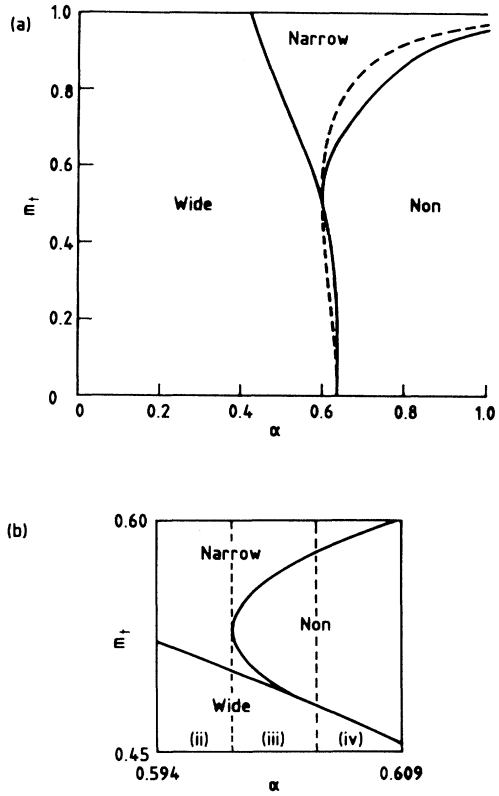

FIG. 7. (a) The phase diagram of retrieval behaviors in the space of m_t and α. The upper right curve extends to the point of maximum storage $(\alpha,m_t)=(2,1)$. The dashed curve shows the fixed points of the retrieval envelope. (b) The amplified phase diagram around the tricritical point, showing three transition behaviors. (Horizontal magnification is 10 times vertical magnification.)

To summarize, a wide retrieval phase exists at low storage level, above which a narrow retrieval phase exists at high training overlap, and a nonretrieval phase exists at low training overlap.

The reentrant behavior can be observed around the tricritical point. This is indicated by a bend of the discontinuous transition line before arriving at the tricritical point. As training noise increases, the transition behaviors at constant α are (i) only wide retrieval for α below 0.42, (ii) narrow $\rightarrow$ wide retrieval for α between 0.42 and 0.599, (iii) narrow $\rightarrow$ nonretrieval $\rightarrow$ narrow $\rightarrow$ wide retrieval for α between 0.599 and 0.604, (iv) narrow $\rightarrow$ nonretrieval $\rightarrow$ wide retrieval for α between 0.604 and 0.64, and (v) narrow $\rightarrow$ nonretrieval for α between 0.64 and 2.

For comparison, we have also plotted in Fig. 7 the fixed points of the retrieval envelope. Note that this curve touches the discontinuous transition line at the onset point of reentrant retrieval, i.e., (α_p,m_p). This curve defines the bounds in the fixed-point retrieval overlaps of the network, whereas the retrieval to nonretrieval transition line determines the bounds in the training overlaps for network retrieval. In other words, the former deter-

mines the vertical bounds in the plots of Figs. 6(a)–6(f), and the latter determines the horizontal bounds.

Thus the presence of at most two stable and two unstable fixed points, in the range $0 \leq m \leq 1$ on the retrieval envelopes of the present problem, gives rise to new features which have not been observed in previous studies of retrieval in dilute networks [19,20], where retrieval mappings consist of at most two stable and one unstable fixed points in the same range. In general, the possibility of even more stable and unstable fixed points cannot be precluded. For example, even more complex fixed-point structures have recently been reported in dilute networks with external stimuli [34].

VI. STABILITY OF THE REPLICA SYMMETRY

The optimization procedure introduced in Sec. II has assumed that the optimal solution is replica symmetric in the space of synaptic interactions. In this section, we consider the stability of the replica symmetric solution. As derived in the Appendix, the condition for stability against replica-symmetry-breaking fluctuations is given by

$$\alpha^{-1} > \int Dt[\lambda'(t)-1]^2 . \tag{6.1}$$

We now proceed to map out the stable and unstable regions, and the phase lines separating them (i.e., the so-called de Almeida–Thouless line), in the space of m_t and α. First consider the high storage limit, in which we expect that the aligning field distribution is single band, and that the parameter γ is small. Equation (2.8) then allows us to write down the inverted function $\lambda(t)$ directly, which is

$$\lambda(t)=t+2\gamma\frac{\exp[-m_t^2t^2/2(1-m_t^2)]}{[2\pi(1-m_t^2)/m_t^2]^{1/2}} . \tag{6.2}$$

The condition (2.9) then reduces to

$$\frac{2\gamma^2m_t^2}{\pi(1-m_t^4)^{1/2}}=\alpha^{-1} , \tag{6.3}$$

and the de Almeida–Thouless condition, which is obtained by the equality of both sides of (6.1), becomes

$$\frac{2\gamma^2m_t^2}{\pi(1-m_t^4)^{1/2}}\frac{m_t^4}{1-m_t^4}=\alpha^{-1} . \tag{6.4}$$

Combining (6.3) and (6.4), the de Almeida–Thouless condition becomes $m_t=2^{-1/4}=0.84$. Thus, in the high storage limit, the replica symmetric solution is stable for $m_t<0.84$, and unstable otherwise.

Next, we consider the behavior when the aligning field distribution changes from two-banded to one-banded. From (3.15), the Taylor expansion around λ_c yields

$$t(\lambda)=t_0+\frac{t'''}{6}(\lambda-\lambda_c)^3 . \tag{6.5}$$

The inverted function $\lambda(t)-\lambda_c\sim(t-t_c)^{1/3}$, and $\lambda'(t)\sim(t-t_0)^{-2/3}$ diverges as t approaches t_0. As a result, the integral in (6.1) diverges, implying that replica symmetry is bound to be broken in the region sufficiently close to the training overlap m_c.

Figure 8 shows the de Almeida–Thouless lines and the band merging line. The band merging line starts at the point $(\alpha,m_t)=(0,0)$, and approaches the high storage limit when m_t approaches 1. There is a de Almeida–Thouless line on each side of this line, and the three lines converge in the low storage limit. As a result, the region of replica symmetry breaking represents a very small fraction of space in regions where retrieval is possible. Our previous observations on the retrieval behaviors are only slightly affected.

For higher storage levels, the two de Almeida–Thouless lines become increasingly distinct. The lower line approaches $m_t=0.84$ in the high storage limit, whereas the upper line reaches an extremum at $(\alpha,m_t)=(6.3,0.94)$ and then bends back to terminate at the point $(\alpha,m_t)=(2,1)$.

The region of replica symmetry breaking effectively separates the regions of replica symmetry into two. Equivalently, reentrant de Almeida–Thouless transition is present in the network. Furthermore, the aligning field distribution in the upper region, of which the MSN is representative, consists of two bands, whereas that in the lower region, of which the Hebbian network is representative, consists of a single band only.

Traditionally, de Almeida–Thouless transitions in neural networks are thought to be driven by pattern interference. When the storage level is too high, the competing tendencies to stabilize all patterns, which usually involve conflicting instructions to encode information in the interactions, result in the multiple ground states characteristic of frustrated systems. Consequently, in studies of optimization in networks, replica symmetry breaking is usually found above some critical storage levels [12,35]. This is illustrated in the high training overlap regime in Fig. 8.

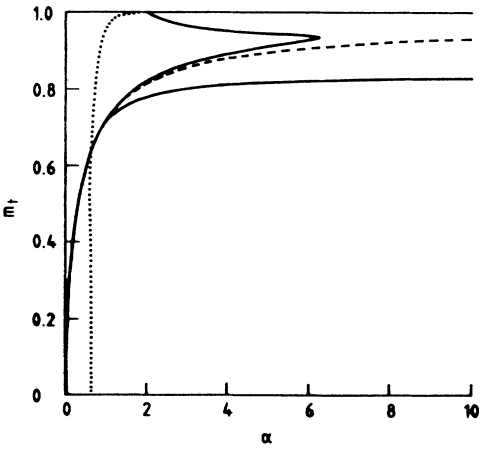

FIG. 8. The de Almeida–Thouless lines and the band merging line (solid and dashed lines, respectively) in the space of the storage level α and the training overlap m_t. The region of replica symmetry breaking is enclosed between the two de Almeida–Thouless lines. For comparison, the storage capacity curve in Fig. 7 is also shown in dotted lines.

However, the present study demonstrates that more subtleties exist in the system. For example, the replica symmetric solution for $m_t < 0.84$ is stable even in the high storage limit, and the optimal solution in the high-training-noise limit is always uniquely the Hebbian network, in which the interactions are uniquely prescribed by $J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu / \sqrt{\alpha C}$.

Apparently, the form of the performance function to be optimized contributes to the replica-symmetry-breaking effects, which are manifested in the banded structure of the aligning field distribution. In the neighborhood of the region where the band gap in the aligning field distribution starts to develop, the space of the aligning fields tends to be topologically disconnected as well. The instability of the replica symmetric solution indicates that there is a multiplicity of possible solutions in the disconnected (or nearly disconnected) space competing to be the optimal solution.

However, it is known that a few disordered systems exhibit a discontinuous transition to the replica-symmetry-breaking phase unrelated to the de Almeida–Thouless line, in much the same way as a soft-mode instability can be preempted by a first-order transition in several conventional phase transitions [35,36]. At this stage we cannot preclude this possibility in noise-optimal networks, and this is definitely a subject for further investigation. Nevertheless, this kind of transition, if existent, broadens the region of replica instability, and the basic picture of a replica-symmetry-breaking phase separating the high- and low-training-noise regimes remains valid even if these complications arise.

VII. CONCLUSION

We have studied the effects of training noise on neural networks. For networks trained with noisy examples, we found features that are characteristic of wide but imperfect basins associated with the stored patterns, as represented by the Hebbian network. These networks have high associativity, low attractor overlaps, and low storage capacities. These wide, imperfect basins are usually associated with aligning field distributions of a single continuous band, in which the extent of storage violations is more homogeneously distributed among all patterns.

On the other hand, networks corresponding to low training noises possess features that are characteristic of perfect but narrow basins, as represented by the MSN. These networks have low associativity, high attractor overlaps, and high storage capacities. The perfect, narrow basins are usually associated with aligning field distributions with two bands, in which a minority of patterns is sacrificially violated to sustain an overall optimal performance.

Because of these competing tendencies, the system exhibits interesting reentrant behaviors, both in the learning and the retrieving stage. In the learning stage, the two-banded aligning field regime is separated from the one-band regime by a region of replica symmetry breaking, signifying the frustration arising from the competition of many possible optima. On the other hand, this results in a slightly depressed performance for intermediate training noises, giving rise to a narrow region of storage levels which exhibit reentrant retrieval behavior in dilute attractor networks. Since reentrant retrieval depends on a very delicate performance depression [for example, the performance overlap $f_{m_t}(m_t)$ is merely at most 0.008 smaller than m_t for $\alpha = 0.62$], it would be interesting to explore whether the same phenomenon can be observed in nondilute networks.

Other recent studies on the dynamical properties reflect the difference between wide, imperfect and narrow, perfect basins when the training noise is varied [37]. An example is the pattern selectivity of neural networks. In dilute Hebbian networks, this issue was first considered by Derrida, Gardner, and Zippelius [24]. They considered a network storing two correlated patterns among a background of p uncorrelated patterns. When the storage level increases, the network undergoes a transition from the distinguishing phase (in which the network retrieves and differentiates the two correlated patterns) to the nondistinguishing phase (in which the network retrieves a common portion of the patterns but fails to differentiate them) and then to the nonretrieval phase. Because of the wide basins, the nondistinguishing phase is quite extensive in the Hebbian network. On the other hand, we have recently considered the pattern selectivity of the MSN [37] and found that the proportion of the retrieving section of the phase diagram corresponding to the nondistinguishing phase is much reduced. This may be attributed to the narrow basins of the MSN.

Another property reflecting the same features of basin structures may be the nature of damage propagation. In the dilute Hebbian network it is found that damage spreads in the basins of attraction [24]. This means that when differences are initially present in two network states, the differences iterate towards some finite nonzero value as the two states evolve. This reflects the fact that the attractor occupies a subspace in the state space, or that a cloud of attractors is present in the basin. On the other hand, the attractor overlap is 1 in the MSN for α below 2, reflecting the presence of a point attractor. Again, this difference can be attributed to the difference in basin structures in the two extreme limits of training noises.

A further interesting dynamical measure is the activity distribution [38], where the activity of a neuron is the time-averaged neuronal state in the attractor. It has been demonstrated that the Hebbian network has a partially frozen activity distribution at low α, but a completely unfrozen activity distribution at high α [38]. On the other hand, the activities of neurons in the MSN are always frozen for α below 2, since the attractor is always the fixed point of the correct pattern. This again illustrates the different basin structures of the networks.

Furthermore, we have recently found that when the training overlap m_t and the storage level α are varied, the optimized networks can be broadly classified into Hebbian-like or MSN-like [39]. The boundary separating the two classes may be defined by either the band merging line in Fig. 8, or the line of minimum interaction susceptibility γ, or the line of maximum deviation of the trajectory between the Hebbian network and the MSN in

the space of synaptic interactions. Although these lines are distinct at high values of α, they converge to the line $m_t=\sqrt{\alpha}$ at low values of α, corresponding to the condition that the signal-to-noise ratio $m_t/\sqrt{\alpha}$ inherent in the training ensemble is equal to unity. Universality classes in the space of interactions have been proposed by Abbott and Kepler [40], in which the Hebbian network and the MSN belong to different classes. Here we have substantiated that because of the contrasting optimization strategies and basin structures, such a classification has physical implications for the learning and retrieving behavior of the network.

Interestingly, similar effects of training noises are also present in Boolean networks [32]. In the low-training-noise limit, the optimal network is prescribed by the nearest-neighbor majority rule, which has a low associativity, high attractor overlap, high storage capacity, high selectivity, and point attractors, indicating that the basins are perfect and narrow. In the high-training-noise limit, the Hebbian network is optimal, and exhibits opposite features.

This shows that the effects of training noise on these measures of the retrieval behavior and the basin structures of neural networks are very universal. They are to a large extent quite independent of the detailed structure of the network. As a corollary, since multilayered networks can be considered as a subset of Boolean networks, we expect that they have similar behaviors too.

A difference with the synaptic networks, however, is that in Boolean networks there is no replica symmetry breaking in the space of network parameters, for the exceedingly large number of adjustable parameters reduces the degree of frustration, in spite of competition in the different terms of the performance function.

Although the notion of basin structures is most directly applicable to attractor neural networks, it is also relevant to feedforward networks. In feedforward networks one is often interested in their generalization ability. This means that a general target relation exists between the input and output states (also called the "teacher network" if it can be realized by the given network architecture [15–17]), and generalization involves reconstructing this target relation using the information provided by a set of examples. In many cases, the target relation is determined by the proximity of the input states to a number of prototype states, as in the "proximity problem" [18]. High training noises lead to wide, imperfect basins associated with each prototype, which correspond to broad but imprecise generalization, meaning that associations of input states far from a prototype are still possible, but the precision of associations is generally weak. Similarly, low training noises lead to perfect but narrow basins, which correspond to more accurate associations but with a more restricted range.

This may have consequences in the training of multilayered networks. When raw data are fed to the input layer, wide basins may be necessary to retrieve signal from a noisy background, and the accurate retrieval of information at the basin centers may not be a high priority. Therefore high training noises may be preferred in the training of this layer. However, the signal-to-noise ratio in the following layers may have progressively increased, and broad associations may no longer be a high priority. Therefore it is preferable to train these layers with progressively lower training noises, so that progressively accurate information can be retrieved.

The discovery that the region of replica symmetry breaking separates the replica symmetric region into two is also relevant to the dynamics of learning. Consider the training of networks using noisy inputs with a fixed training noise level. As the number of patterns is increased, the network passes from the two-band to one-band regime through the region of replica symmetry breaking. If learning is achieved by gradient-descent dynamics, this means that the system will encounter a host of local minima, and this costs extra computational effort.

To conclude, we have demonstrated the effects of tuning the basin structures of neural networks by training noise on both learning and retrieving properties, and explained their underlying physics. It is hoped that this study will help the design of neural networks optimal to various operating conditions and requirements.

ACKNOWLEDGMENTS

We thank D. Wallace, H. W. Yau, and E. Domany for stimulating discussions. This work was supported financially by Grant No. GR/G02727 of the Science and Engineering Research Council of the United Kingdom. Computation facilities were provided by the University of London Computer Centre.

APPENDIX

In this appendix we derive the condition for the stability of the replica symmetric optimal solution in the space of interactions. The derivation is a generalization of Appendix B in Ref. [12]. Following the notations in Appendix 1 of our previous work [9], we evaluate the following parameters required in calculating the eigenvalues of the stability matrix:

$$P=\frac{\partial^2 G_\Lambda}{\partial q_{\alpha\beta}^2}=\langle\langle x^2\rangle_{x\lambda}^2\rangle_t-\langle\langle x\rangle_{x\lambda}^2\rangle_t^2 , \tag{A1}$$

$$Q=\frac{\partial^2 G_\Lambda}{\partial q_{\alpha\beta}\partial q_{\alpha\gamma}}=\langle\langle x^2\rangle_{x\lambda}\langle x\rangle_{x\lambda}^2\rangle_t-\langle\langle x\rangle_{x\lambda}^2\rangle_t^2 , \quad \beta\neq\gamma , \tag{A2}$$

$$R=\frac{\partial^2 G_\Lambda}{\partial q_{\alpha\beta}\partial q_{\gamma\delta}}=\langle\langle x\rangle_{x\lambda}^4\rangle_t-\langle\langle x\rangle_{x\lambda}^2\rangle_t^2 , \quad \alpha,\beta\neq\gamma,\delta , \tag{A3}$$

where $\langle f(x)\rangle_{x\lambda}$ and $\langle f(t)\rangle_t$ are, respectively, defined by

$$\langle f(x)\rangle_{x\lambda}=\frac{\int (dxd\lambda/2\pi)f(x)\exp[\beta_{an}g(\lambda)+ix(\lambda-\sqrt{q}\,t)-(1-q)x^2/2]}{\int (dxd\lambda/2\pi)\exp[\beta_{an}g(\lambda)+ix(\lambda-\sqrt{q}\,t)-(1-q)x^2/2]}\ , \tag{A4}$$

$$\langle f(t)\rangle_t=\int Dt\, f(t)\ . \tag{A5}$$

Following de Almeida and Thouless [23] the eigenvalue corresponding to replica-symmetry-breaking fluctuations is

$$\gamma_\Lambda=P-2Q+R=\langle(\langle x^2\rangle_{x\lambda}-\langle x\rangle_{x\lambda}^2)^2\rangle_t\ . \tag{A6}$$

In the limit $\beta_{an}\to\infty$ and $q\to 1$, the exponential terms in (A4) diverge, facilitating the use of steepest descent. In particular the denominator in (A4) becomes

$$\lim_{\beta_{an}\to\infty}\int\frac{d\lambda}{\sqrt{2\pi(1-q)}}\exp\left[\beta_{an}g(\lambda)-\frac{(\lambda-\sqrt{q}\,t)^2}{2(1-q)}\right]=\frac{\exp(\beta_{an}\{g(\lambda(t))-(1/2\gamma)[\lambda(t)-t]^2\})}{\sqrt{1-\gamma g''(\lambda(t))}}\ , \tag{A7}$$

where $\lambda(t)$ is given by (2.8). Substituting x for $f(x)$ in (A4), the numerator becomes

$$\lim_{\beta_{an}\to\infty}\int\frac{d\lambda}{\sqrt{2\pi(1-q)}}i\left[\frac{\lambda-\sqrt{q}\,t}{1-q}\right]\exp\left[\beta_{an}g(\lambda)-\frac{(\lambda-\sqrt{q}\,t)^2}{2(1-q)}\right]=i\left[\frac{\lambda(t)-t}{1-q}\right]\frac{\exp(\beta_{an}\{g(\lambda(t))-(1/2\gamma)[\lambda(t)-t]^2\})}{\sqrt{1-\gamma g''(\lambda(t))}}\ . \tag{A8}$$

Thus

$$\langle x\rangle_{x\lambda}=i\frac{\lambda(t)-t}{1-q}\ . \tag{A9}$$

Similarly, substituting x^2 for $f(x)$ in (A5), the numerator becomes

$$\lim_{\beta_{an}\to\infty}\int\frac{d\lambda}{\sqrt{2\pi(1-q)}}\left[\frac{1}{1-q}-\left[\frac{\lambda-\sqrt{q}\,t}{1-q}\right]^2\right]\exp\left[\beta_{an}g(\lambda)-\frac{(\lambda-\sqrt{q}\,t)^2}{2(1-q)}\right]$$
$$=\left[\frac{1}{1-q}-\left[\frac{\lambda(t)-t}{1-q}\right]^2-\frac{1}{\beta_{an}(1-q)^2[\gamma^{-1}-g''(\lambda(t))]}\right]\frac{\exp(\beta_{an}\{g(\lambda(t))-(1/2\gamma)[\lambda(t)-t]^2\})}{\sqrt{1-\gamma g''(\lambda(t))}}\ . \tag{A10}$$

Thus

$$\langle x^2\rangle_{x\lambda}=\frac{1}{1-q}-\left[\frac{\lambda(t)-t}{1-q}\right]^2-\frac{1}{(1-q)[1-\gamma g''(\lambda(t))]}\ . \tag{A11}$$

Hence

$$\langle x^2\rangle_{x\lambda}-\langle x\rangle_{x\lambda}^2=\frac{1}{1-q}\left[1-\frac{1}{1-\gamma g''(\lambda(t))}\right]\ . \tag{A12}$$

Using (2.7), we have

$$P-2Q+R=\frac{1}{(1-q)^2}\int Dt[1-\lambda'(t)]^2\ . \tag{A13}$$

The calculation of $\partial^2 G_J/\partial F_{\alpha\beta}\partial F_{\gamma\delta}$ is similar to Ref. [11] yielding, in the notation of Ref. [11],

$$\gamma_J=P'-2Q'+R'=(1-q)^2\ . \tag{A14}$$

The condition for stability of the replica symmetric solution is given by $\alpha\gamma_\Lambda\gamma_J<1$, resulting in (6.1).

*Present address: Department of Physics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Electronic address: PHKYWONG@USTHK.BITNET

†Electronic address: SHERR@THPHYS.OX.AC.UK

[1] J. J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2544 (1982).

[2] D. Amit, H. Gutfreund, and H. Sompolinsky, Ann. Phys. (N.Y.) **173**, 30 (1987); Phys. Rev. **32**, 1007 (1985); **35**, 2293 (1987).

[3] T. Kohonen, *Self-Organization and Associative Memory* (Springer-Verlag, Berlin, 1984).

[4] L. Personnaz, I. Guyon, and G. Dreyfus, J. Phys. (Paris) **46**, L359 (1985).

[5] B. Widrow and M. E. Hoff, in *IRE WESCON Convention Record* (IRE, New York, 1960), Part 4, p. 96.

[6] M. Minsky and S. Papert, *Perceptrons,* Expanded ed. (MIT Press, Cambridge, MA, 1988).

[7] J. K. Anlauf and M. Biehl, Europhys. Lett. **10**, 687 (1989).

[8] E. Gardner, N. Stroud, and D. J. Wallace, J. Phys. A **22**, 2019 (1989).
[9] K. Y. M. Wong and D. Sherrington, J. Phys. A **23**, 4659 (1990).
[10] M. Griniasty and H. Gutfreund, J. Phys. A **24**, 715 (1991).
[11] K. Y. M. Wong (unpublished).
[12] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[13] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, Science **220**, 671 (1983).
[14] K. Y. M. Wong and D. Sherrington, J. Phys. A **23**, L175 (1990).
[15] Ph. Refregier and J.-M. Vignolle, Europhys. Lett. **10**, 387 (1989).
[16] G. Györgyi and N. Tishby, in *Neural Networks and Spin Glasses,* edited by W. K. Theumann and R. Köberle (World Scientific, Singapore, 1990).
[17] G. Györgyi, Phys. Rev. Lett. **64**, 2957 (1990).
[18] D. Hansel and H. Sompolinsky, Europhys. Lett. **11**, 687 (1989).
[19] D. Amit, M. Evans, H. Horner, and K. Y. M. Wong, J. Phys. A **23**, 3361 (1990).
[20] A. Komoda, R. Serneels, M. Bouten, and K. Y. M. Wong, J. Phys. A **24**, L743 (1991).
[21] K. Y. M. Wong and M. Bouten, Europhys. Lett. **16**, 525 (1991).
[22] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[23] J. R. L. de Almeida and D. J. Thouless, J. Phys. A **11**, 983 (1978).
[24] B. Derrida, E. Gardner, and A. Zippelius, Europhys. Lett. **4**, 167 (1987).
[25] K. Y. M. Wong and D. Sherrington, in *Statistical Mechanics of Neural Networks,* edited by L. Garrido, Lecture Notes in Physics Vol. 368 (Springer-Verlag, Berlin, 1990), p. 105.
[26] D. Thouless, P. W. Anderson, and R. G. Palmer, Philos. Mag. **35**, 593 (1977).
[27] S. F. Edwards and P. W. Anderson, J. Phys. F **5**, 965 (1975).
[28] E. Gardner, J. Phys. A **21**, 257 (1988).
[29] T. B. Kepler and L. F. Abbott, J. Phys. (Paris) **49**, 1657 (1988).
[30] W. Krauth, J.-P. Nadal, and M. Mézard, J. Phys. A **21**, 2995 (1988).
[31] E. Gardner, J. Phys. A **22**, 1969 (1989).
[32] K. Y. M. Wong and D. Sherrington, Europhys. Lett. **10**, 419 (1989).
[33] D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel Distributed Processing* (MIT Press, Cambridge, MA, 1986), Vol. 1.
[34] H. W. Yau and D. Wallace, J. Phys. A **24**, 5639 (1991).
[35] W. Krauth and M. Mézard, J. Phys. (Paris) **50**, 3056 (1989).
[36] D. J. Gross and M. Mézard, Nucl. Phys. **B240**, 431 (1984).
[37] A. Rau, K. Y. M. Wong, and D. Sherrington, Europhys. Lett. **17**, 649 (1992).
[38] B. Derrida, J. Phys. A **22**, 2069 (1989).
[39] K. Y. M. Wong, A. Rau, and D. Sherrington, Europhys. Lett. **19**, 559 (1992).
[40] L. F. Abbott and T. B. Kepler, J. Phys. A **22**, 2031 (1989).