Retrieval phase diagrams for attractor neural networks with optimal interactions

1990 J. Phys. A: Math. Gen. 23 3361

(http://iopscience.iop.org/0305-4470/23/14/032)

J. Phys. A: Math. Gen. 23 (1990) 3361–3381. Printed in the UK

Retrieval phase diagrams for attractor neural networks with optimal interactions

Daniel J Amit†||, M R Evans‡, H Horner§ and K Y M Wong†¶

† Department of Physics, Imperial College, London SW7 2BZ, UK
‡ Department of Physics, University of Edinburgh, Edinburgh EH9 3JZ, UK
§ Institute for Theoretical Physics, Heidelberg University, Heidelberg, Federal Republic of Germany

Received 15 January 1990

Abstract. We consider the retrieval properties of attractor neural networks whose synaptic matrices have been constructed to maximise the number of patterns which can be stored in a *perceptron* satisfying certain constraints. Retrieval is studied in the absence as well as in the presence of fast noise (temperature). The discussion is restricted to dilute networks, for which dynamical equations for the overlaps are available.

When the patterns are stored with a prescribed lower limit on the stability parameter on every site, the full stability of the perceptron ensures the existence of an attractor with perfect retrieval. It is found that the curve of critical storage capacity (α) against temperature (T) is a line of first-order transitions for high values of α and becomes second order for low α, at a point of a tricritical nature. The phase diagram is compared with the dilute Hopfield model. It is found that at high synaptic noise levels the diluted Hopfield net stores more effectively than the network trained for optimal perceptron storage.

When a given fraction of sites is allowed to violate the stability bound, the solution of the perceptron 'learning' problem does not ensure the existence of an attractor of finite overlap even in the absence of noise. This case is studied separately for $T = 0$ and for finite T. For $T = 0$ it is shown that despite the fact that errors increase dramatically the capacity at the perceptron level, there are no attractors above $\alpha = 2$. For finite T we compute phase diagrams for several situations, emphasising the sensitivity to the level of errors. It was found that learning with errors gives at best a marginal improvement of the retrieval storage capacity in a certain temperature range, but storage drops dramatically as the temperature is lowered. This leads to a re-evaluation of the Gardner–Derrida network optimisation in the presence of errors.

1. Introduction

One of the main attractions of attractor neural networks is the impressive robustness of their performance to noise. The performance is associative recall, i.e. the fact that the state of the network drifts to the neighbourhood of one of a set of stored patterns (the one associated with the stimulus) and remains there for a long time. In the

|| On leave from the Racah Institute of Physics, Hebrew University, Jerusalem (also present address).
¶ Present address: Department of Theoretical Physics, University of Oxford, 1 Keble Road, Oxford OX1 3NP, UK.

0305-4470/90/143361+21$03.50

biological context this expresses itself in the appearance of bursts of spikes in a well defined subset of the neurons in the network.

The noise can have different origins. It can be 'fast', rendering the dynamical behaviour of every neuron indeterminate on the basic neuronal time scale (see e.g. [1–4]), or it can be 'slow', representing fixed randomness in the synaptic structure (see e.g. [5–7]). For both types of noise the performance of the Hopfield synaptic structure, has been shown to be extremely robust (see e.g. [3,4,7–9]).

1.1. Optimal storage and robustness

Gardner [10] opened the way for studying networks with synaptic matrices which are not explicitly prescribed, but which instead optimise the storage capacity for sets of random patterns. In this approach one arrives at a collection of synaptic matrices *all* satisfying the condition that *all* patterns are stored with a stability parameter greater than some pre-assigned value, for as many patterns as such matrices exist.

The Gardner approach gives on the one hand an evaluation of the storage capacity, i.e. the highest number of random patterns for which a non-vanishing set of interactions exists. On the other hand, it provides a probability distribution of the local fields, when the network is in a stored pattern, for a generic matrix in the volume of interactions satisfying the storage constraints [11,12]. This distribution of fields provides the dynamical effect following one time step [11] upon any initial input state, but bold steps have to be taken in order to go beyond [13].

Clearly, a matrix which stores the patterns perfectly, satisfying the stability requirement (equation (8), below), will have the patterns as fixed points of the dynamics in the absence of noise. But will it retrieve at finite noise levels? How will the storage capacity of the network and its retrieval quality vary with the noise level? The first question is concerned with the maximal number of patterns which are stored perfectly at zero temperature and whose generic matrix produces a non-zero fixed point value for the overlap with single patterns at finite temperature. The second question is concerned with the magnitude of the overlap at the critical storage level. What one is after is the equivalent of the phase diagram in the T–α plane. These questions are important for the estimation of the robustness of an optimally storing matrix. Yet they have found no technical solution for fully connected networks to date. The reason is that on the one hand the matrices are not symmetric, precluding the application of equilibrium statistical mechanics. On the other, dynamical equations become very complicated beyond the first step [14], although recently progress has been made on the dynamics of fully connected networks [13,15].

1.2. Overlap dynamics without errors

A way toward a qualitative understanding of the overlap dynamics and attractor structure is the study of the dilute network, for which one has equations of motion for the overlaps [8]. It has been used on several occasions to study the behaviour of networks with intrinsically asymmetric couplings. In some cases the dilution was superposed on an explicit synaptic prescription. See e.g. [16,17]. But, as was recently shown in a posthumously published paper of Gardner [18], optimal storage matrices can be studied for the dilute system with the same effectiveness as for the fully connected one. Using the ensuing distribution of fields and the existence of dynamical equations for the overlap, m, Gardner has studied the structure of the basins of attraction in a noiseless, optimally connected network near saturation. It was found that if the

storage fraction $\alpha = p/C < 0.42$, where p is the number of stored patterns and C is the number of connections per neuron, the basin of attraction of the retrieval state, is unity. Above this value of α the $m = 0$ fixed point becomes an attractor at the expense of the basin of retrieval.

In this paper we extend this study to include fast noise (temperature). The dynamics of a neuron is stochastic; its new state is determined according to a probability which depends on the local field generated by the previous network state and on the temperature. Following the formulation in section 2, we compute, in section 3 the phase diagram in the α–T plane, at maximal storage, i.e. when the volume in interaction space shrinks to zero. It is found that the line separating retrieval from non-retrieval is a line of continuous transitions, with the retrieval overlap (m) vanishing continuously for $\alpha < 0.26$. Above this value of α the transition is discontinuous. The value $\alpha = 0.26$ is determined by an analytic condition on the distribution of the local stability parameter. This point is the analogue of a 'tricritical' point. The noisy performance of the optimal network is compared with the performance of networks with dilute Hopfield coupling matrices [8]. It is found that for high noise levels the dilute Hopfield network has retrieval attractors at higher storage levels than the network of optimal storage at $T = 0$. Section 4 presents some analytical results in the vicinity of the continuous line and near the 'tricritical' point.

1.3. Storage with errors and overlap dynamics

The optimal storage computation of Gardner was extended to allow for a fixed fractional number of violations of the stabilty constraints [19]. This is done by defining a cost function to be the number of violations of the stability constraint (irrespective of the magnitude of the violation) and optimising the cost in the space of interactions. The behaviour of the noiseless perceptron under this relaxed set of constraints was investigated in detail. However, in this case the existence of an appropriate matrix of weights does not guarantee the existence of a finite overlap attractor even at $T = 0$. And then there is the whole question of the structure of the attractors in the presence of noise.

In section 5 we use the Abbott–Kepler results for the universality in the space of optimal interactions [20] to obtain the field probability distribution in a pattern for a network whose couplings are optimised with the Gardner–Derrida constraints. This distribution function has a rather peculiar nature. In fact, the distribution of the fields in the optimal Gardner–Derrida network develops, in addition to a distribution starting at the imposed stability parameter κ, a Gaussian tail at the extreme negative values of the stability parameter, wherever constraint violations exist (see e.g. figure 1). This implies that the network enlarges the number of patterns it can store, at a fixed error fraction, by making the errors in the field, i.e. the violation of the constraints wherever they are violated, larger and larger.

The consequences of this field distribution for the attractor structure is analysed in section 5, for the noiseless and for the noisy case. It is found that in some parameter regions the sensitivity to the number of allowed errors is rather extreme. Phase diagrams for several values of the error fraction are presented and discussed, in parameter regions in which replica symmetry (in weight space) is stable. In particular, it is shown that, despite the fact that in the single layer perceptron the storage capacity increases dramatically above $\alpha_p = 2$ with the allowed fraction of errors, there are no attractors for $\alpha > 2$. In what follows we shall denote the perceptron storage capacity by α_p and the storage capacity for attractors by α_c.

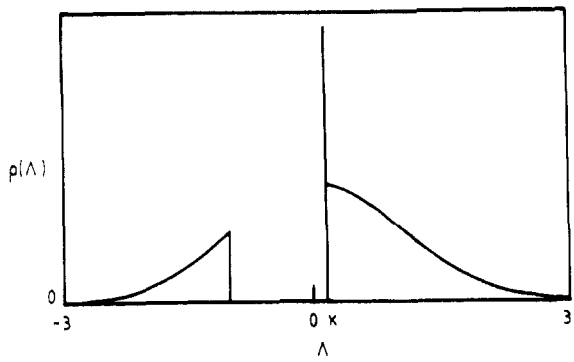

Figure 1. The probability distribution of the stability parameter in a network optimised with error fraction [19] $f = 0.15$ and $\kappa = 0.1$.

2. Formulation

We denote the state of the network at time t by $\{S_i(t) = \pm 1\}$, with $i = 1, \ldots, N$ listing the neurons. The dynamics of a neuron in the presence of fast noise is given by the probability that the node i will take on the state S_i at time $t + 1$, e.g.

$$\Pr[S_i(t+1)] = \frac{1}{1 + \exp[-2\beta h_i(t) S_i(t+1)]} \tag{1}$$

when it sees a local field h_i due to the activity in the network at time t. The field at node i is determined by the states of the neurons connected to it via the synaptic matrix, which will be denoted by $J_{ij}/\sqrt{C}$, namely

$$h_i(t) = \frac{1}{\sqrt{C}} \sum_j J_{ij} S_j(t) \tag{2}$$

and J_{ij} is normalised by:

$$\sum_{j=1}^{N} J_{ij}^2 = C \tag{3}$$

where C is the mean connectivity per neuron. In (1) the magnitude of the fast noise is denoted by $T = \beta^{-1}$.

When storing p N-bit patterns $\{\xi_i^\mu = \pm 1\}$ with $\mu = 1, \ldots, p$ in the network, one monitors the degree of associative recall of a single pattern ν by the overlap

$$m(t) = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\nu S_i(t). \tag{4}$$

When the network settles in an attractor, associative recall of pattern ν is signalled by the fact that the temporal average of the overlap

$$\langle m(t) \rangle = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\nu \langle S_i(t) \rangle \tag{5}$$

is large, of order unity. Overlaps with all other patterns will fluctuate around zero. The angular brackets denote a temporal average over the states of the individual neurons.

For the network to retrieve associatively a set of patterns, information about the patterns has to be encoded in the synaptic matrix J_{ij}. One form of encoding, which has been found to be extremely robust, is Hopfield's synaptic structure, i.e. in the notation of (2)

$$J_{ij} = \frac{1}{\sqrt{\alpha C}} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{6}$$

see, e.g. references [3,4,7–9]).

On the other hand, in Gardner's approach the synaptic matrices are not explicitly prescribed. Instead, one considers generic matrices for which a stability requirement is satisfied for *every* pattern at *every* node. The stability parameter at a site, when the network is in a pattern is, to recapitulate,

$$\Delta^\mu = \frac{1}{\sqrt{C}} \xi_i^\mu \sum_{j=1}^{N} J_{ij} \xi_j^\mu. \tag{7}$$

The stability requirement, which is the constraint defining the volume in interaction space, reads

$$\Delta_i^\mu > \kappa \qquad \text{for } \mu = 1, \ldots, p;\ i = 1, \ldots, N. \tag{8}$$

This approach provides a probability distribution $\rho_\kappa(\Lambda)$ when the network is in a stored pattern, for a generic matrix in the volume of interactions satisfying (8). In other words, one can compute

$$\rho_\kappa(\Lambda) = \overline{\delta(\Delta^\nu - \Lambda)} \tag{9}$$

where the bar indicates an average over the volume of storing matrices as well as over the distribution of patterns stored and the subscript κ refers to the particular stability constraint (equation (8)). Note that since the connectivity is considered optimal in the absence of synaptic noise, the κ determines α and will be used below as the independent variable. The distribution of stability parameters provides, in turn, dynamical equations for the overlap of the dilute network [11, 12], in terms of the of the signal and Gaussian noise in the local field [5, 8]. It reads,

$$m(t+1) = \int_{-\infty}^{\infty} \mathrm{d}\Lambda\, \rho(\Lambda) \int_{-\infty}^{\infty} \mathrm{D}y \tanh\left\{\beta\left[m(t)\Lambda + \sqrt{1 - m^2(t)}\, y\right]\right\} \tag{10}$$

where ρ is the probability distribution of the stability parameters and $\mathrm{D}y$ is the Gaussian measure

$$\mathrm{D}y \equiv \frac{\mathrm{d}y \exp(-y^2/2)}{\sqrt{2\pi}}.$$

Equation (10) describes the parallel (synchronous) dynamics of a dilute network. Here we shall not consider the dynamical features implied by this equation but only

the structure of its fixed points. The fixed points are, of course, the same as those of asynchronous dynamics, which is more realistic and more robust.

For a fixed value of the stability bound κ, maximal storage is attained when the volume of the interactions satisfying the stability constraint shrinks to a point, which is indicated by

$$q = \overline{\frac{1}{C}\sum_{i=1}^{N} J_{ij}^{a} J_{ij}^{b}} \to 1. \tag{11}$$

The indices a and b refer to different replicas. Storage capacity, in the above sense, α_{P}, depends on κ via [10]

$$\alpha_{\mathrm{P}}^{-1} = \int_{-\infty}^{\kappa} \mathrm{D}t(\kappa - t)^2. \tag{12}$$

The field distribution function at maximal storage is given [11,12] as

$$\rho_{\kappa}(\Lambda) = \frac{\exp(-\Lambda^2/2)}{\sqrt{2\pi}}\Theta(\Lambda - \kappa) + \tfrac{1}{2}[1 + \mathrm{erf}(\kappa/\sqrt{2})]\delta(\lambda - \kappa) \tag{13}$$

where κ is defined in (8).

3. Noisy dynamics of errorless optimal network

The fixed points of (10) have been studied numerically and the results are summarised in figures 2 and 3. In figure 2(a) we present the phase diagram, the composite line separating the regions in the α–T plane of retrieval (under the curve) from non-retrival (above), as well as regions of 'wide retrieval' and 'narrow retrieval' (see below for explanation of these terms), to the left and to the right of the dotted curve, respectively. Recall that α is determined by the $q = 1$ optimisation condition, as a function of the stability parameter κ. The lines are actually lines of $T_c(\kappa)$, which is why κ is given on the top horizontal axis. Non-retrieval is characterised by the absence of a fixed point with non-zero overlap m. For low α and high T (broken curve) the transition is continuous. At $\alpha \approx 0.26$, it turns into a discontinuous transition (full curve), at a 'tricritical' point.

The dotted curve is the analytic continuation of the continuous transition. It reaches $\alpha = 0.42$ at $T = 0$. In section 4 we shall see that this line, together with the line of continuous transitions, separates the regions in which the $m = 0$ fixed point is stable (on the right) and unstable (on the left). Thus, in the upper right region, $m = 0$ is the only fixed point, whereas in the lower right region, attractors corresponding to retrieval and to no retrieval coexist. But, because of the presence of the $m = 0$ attractor, the basin of the retrieval attractor is narrowed. Therefore, the dotted curve is also the transition line from 'narrow' to 'wide' retrieval, namely to its left the basin of attraction of the attractor is the entire interval (0–1), while to its right, only part of this interval constitutes the basin of attraction and $m = 0$ retains a

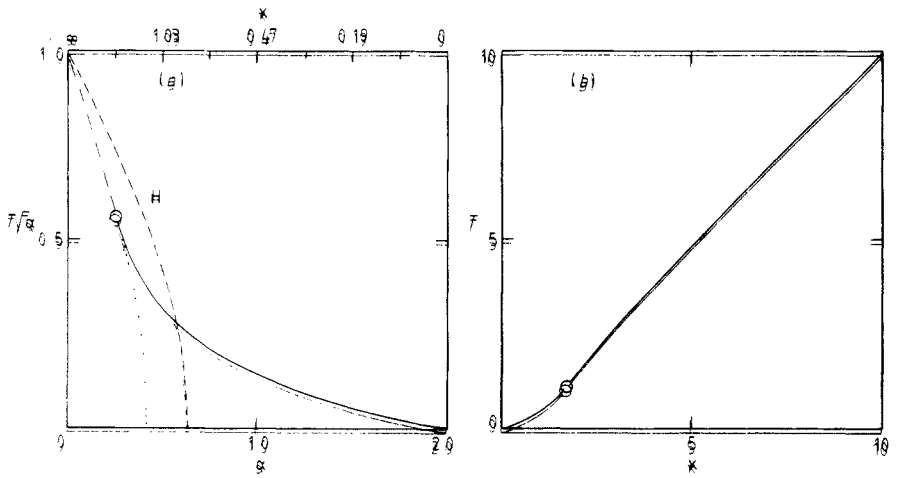

Figure 2. (*a*) Phase diagram, α–T plane, for dilute, optimally connected network. Full curve: discontinous transition; broken curve: continuous transition; dotted curve: transition from 'wide retrieval' (100% basin of attraction) to 'narrow retrieval' (less than 100%). The small circle is a 'tricritical' point. Curve (H) is phase separation curve for dilute Hopfield network [8], with couplings normalised according to (3). The top horizontal axis gives the value of κ which corresponds to the α on the bottom axis. (*b*) Transition temperature against κ. Note the relation $t = \kappa$ over almost the entire range of continuous transitions (see e.g. section 4).

finite basin of its own. We point out that the basin boundary undergoes a continuous transition along this line.

In figure 2(*b*) we plot transition temperature against stability parameter κ. One observes that for a wide region of high κ, beyond the tricritical point, the transition temperature $T_c \cong \kappa$ (see, e.g. section 4).

The existence of both the 'narrow' and 'wide' retrieval phases can be summarised as the result of interplay between fast noise (temperature) and slow noise (pattern interference). Roughly speaking, interference due to increasing storage level shifts the local field distribution downwards, resulting in a drop in the output overlap $m(t+1)$ which is more drastic for a small input overlap $m(t)$ than for a large one (see, e.g. (10)). On the other hand, an increase in temperature causes a drop in the output overlap less dependent on the input overlap. These effects lead to the phase of no retrieval for high T and high α, and make the phase line seperating the 'narrow' and 'wide' regions within the retrieval phase virtually vertical.

The phase line for the dilute asymmetric Hopfield network is also plotted in figure 2(*a*). It represents a continuous transition between retrieval and non-retrieval at all storage levels, and the entire retrieval phase is 'wide'. At $T = 0$, it reaches $\alpha = 0.64$ [8]. It is interesting to note that while the Gardner prescription has a higher storage capacity for attractors at low temperatures, the Hopfield network stores more attractors at high temperatures. High temperature series expansion in the following section shows that for both types of network, the asymptotic behaviour for T_c is

$$T_c \sim \frac{1}{\sqrt{\alpha}}. \tag{14}$$

Hence the difference in performance of the two networks vanishes at very high temperatures, resulting in the common vertical intercept at $T\sqrt{\alpha} = 1$ in figure 2(*a*).

Concerning the behavioural difference between the Gardner and Hopfield networks, we draw attention to a recent study on training noises [21,22]. At $T = 0$, the Hopfield network maximises, over the entire interaction space, the output overlap for very noisy input overlap, while the Gardner prescription maximises the output overlap for a clean input. The ability of the Hopfield network to process noisy inputs accounts for its larger basins of attraction, and consequently its stronger propensity to 'wide' retrieval. It also accounts for the better associative performance of the Hopfield network in a noisy environment such as high temperature. In fact, the Hopfield network has the best storage capacity of attractors at sufficiently high temperatures [23].

In figure 2(b) we plot the transition temperature against the stability parameter κ. One observes that for sufficiently high κ, beyond the tricritical point, the transition temperature $T_c = \kappa$ (see, e.g. section 4). In the physical analogue of the single node dynamics (equation(1)), one identifies κ as the 'energy gap', and hence a thermal agitation of magnitude $T \sim \kappa$ is required to cause a significant fraction of the Ising spins to flip against their local fields and become errors. This implies $T_c \sim \kappa$.

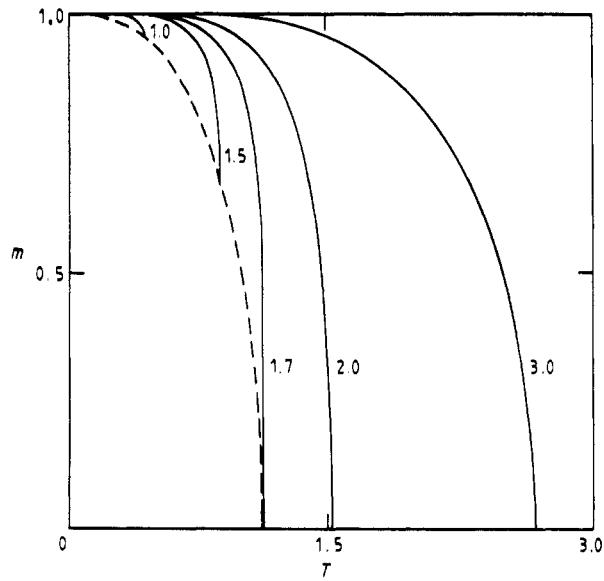

Figure 3. Retrieval quality m against noise T for several values of κ (and hence of $\alpha_c(\kappa)$), marked in the figure.

In figure 3 we show the retrieval quality m against T for several values of $\alpha_c(\kappa)$, i.e. several values of κ. For high κ (low $\alpha_c(\kappa)$), m vanishes continuously at the transition. Hence the horizontal intercepts of the retrieval quality curves correspond to points on the phase line separating wide retrieval from no retrieval. This transition is continuous up to the tricritical storage $\alpha_{tr} = 0.26$ ($\kappa_{tr} = 1.7, T_{tr} = 1.1$).

For lower κ, m vanishes discontinuously, and the broken curve shows m at the discontinuous transition. The touching points of the retrieval quality curves with the broken curve correspond to points on the phase line separating narrow retrieval from no retrieval.

It should be noted that while for low α_c (high κ) m vanishes continuously at the transition, wherever both networks retrieve the high κ network gives better retrieval quality.

4. Analysis of the order parameter map

The order parameter map given by (10) is quite general and can be studied without specifying either the form or the parametrisation of $\rho(\Lambda)$. We first proceed, therefore, to analyse the fixed point structure of the map without specifying $\rho(\Lambda)$. In particular, one can obtain conditions on $\rho(\Lambda)$ for various types of transitions from retrieval to no retrieval, which would, in turn, yield critical values of the parameters of the model. All these questions go back to the dependence of the fixed point structure of (10) and of the stability of these fixed points on the properties of $\rho(\Lambda)$ and on the noise T.

The fixed points, $m^*(\kappa,T)$, of (10) are given by the roots of $g(m,\kappa,\beta)$ where

$$g(m,\kappa,\beta) = f(m,\kappa,\beta) - m \tag{15}$$

$$f(m,\kappa,\beta) = \int_{-\infty}^{\infty} \mathrm{D}y \int_{-\infty}^{\infty} \mathrm{d}\Lambda\, \rho(\Lambda) \tanh\left\{\beta\left[m\Lambda + \sqrt{1-m^2}\, y\right]\right\} \tag{16}$$

with $T = 1/\beta$. This function is odd so we can focus our discussion on $m > 0$. For an attractor, we require $g'(m^*) < 0$ where $g(m^*) = 0$, and for an unstable fixed point $g'(m^*) > 0$.

For the optimal neural network considered here there are three fixed point structures: $m = 0$ is always a fixed point, due to the antisymmetry of $g(m)$. It may be the only fixed point or there may be either one or two additional fixed points with $m > 0$.

1. When $m = 0$ is the only fixed point it must be stable (figure 4, curve (1)). This is the situation of no retrieval. This is the case, for example, in Gardner's model at very high loading, i.e. $\alpha > 2$ [18].
2. When there is only one additional positive fixed point it must be stable and $m = 0$ must be unstable (figure 4, curve (2)). This is, for example, the case at low loading ($0 < \alpha < 0.42$) in Gardner [18].
3. Finally, when there are two additional positive fixed points, as for intermediate loading ($0.42 < \alpha < 2$) in Gardner [18], $m = 0$ and the m with the high absolute value are stable, while the fixed point in the middle is unstable (figure 4, curve (3)). The intermediate fixed point delimits the basin of attraction of the high-m (retrieval) fixed point.

One may conceive of neural networks with more complex fixed point structures [23]. Those will not be considered here.

As the structure of g changes with the noise level, at fixed learning parameters (e.g. κ or α), the fixed point structure may change between any two of the three alternatives. For purposes of associative retrieval, we are primarily interested in changes between retrieval dynamics and no retrieval, i.e. (2) → (1) or (3) → (1). The first transition is continuous, in the sense that the finite m fixed point disappears as $m \to 0$ continuously. The second transition is discontinuous, as the two finite m fixed points coalesce and disappear at finite m. In addition there is a transition (2) → (3), through which the retrieval attractor survives, but its basin of attraction, which is unity in (2), is reduced in (3).

4.1. The continuous transition and the 'tricritical' point

A necessary condition for a continuous transition from retrieval to no retrieval is the appearance of the $m = 0$ attractor. This requires,

$$g'(0) = 0 \tag{17}$$

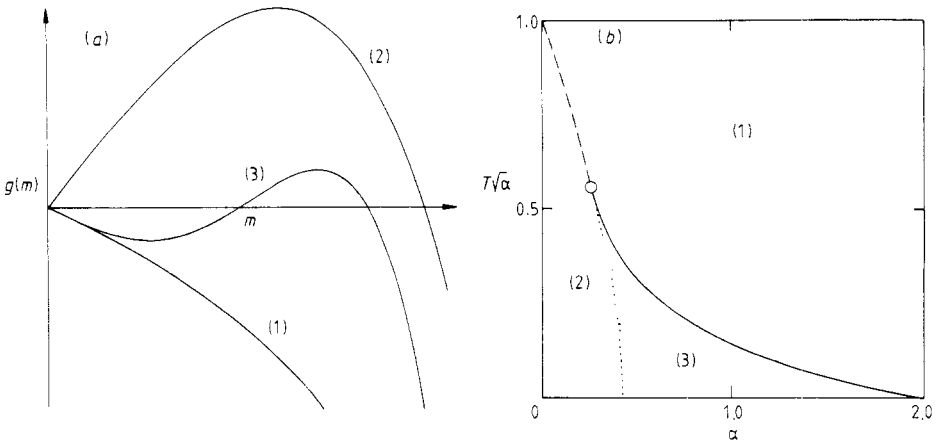

Figure 4. (*a*) Schematic sketch of the three possible fixed point structures in g: (1) no retrieval, (2) wide retrieval, (3) narrow retrieval. (*b*) Schematic representation of different retrieval regions in the phase diagram.

which, using (16), reads:

$$\langle \Lambda \rangle = \frac{1}{\beta \int_{-\infty}^{\infty} \mathrm{D}y \, \mathrm{sech}^2(\beta y)} \tag{18}$$

where the angular brackets denote an average over the distribution $\rho(\Lambda)$, i.e.

$$\langle F(\Lambda) \rangle = \int_{-\infty}^{\infty} \mathrm{d}\Lambda \, \rho(\Lambda) F(\Lambda). \tag{19}$$

But this condition is not sufficient since the retrieval attractor may still remain stable. Thus, $g' = 0$ may also indicate a transition (2) $\rightarrow$ (3). We come back to this issue below.

The additional condition for a continuous transition is:

$$g'''(m = 0) < 0 \tag{20}$$

to ensure a stable fixed point at arbitrarily small m as T is raised to the transition value. Whereas for

$$g'''(m = 0) > 0 \tag{21}$$

$m = 0$ is a stable fixed point and the retrieval fixed point disappears discontinuously, when the two roots at finite positive m coalesce and disappear. The two cases are separated by the point at which

$$g'''(m = 0) = 0 \tag{22}$$

when all five (positive and negative) roots of $g(m)$ coalesce at $m = 0$. The special point is determined by the two simultaneous equations (18) and (22). It is the analogue of a thermodynamic tricritical point. When (16) is substituted in (22), one has

$$3\langle \Lambda \rangle - \langle \Lambda^3 \rangle = 0. \tag{23}$$

Note that this condition is temperature independent, which implies that it must reflect itself also in the $T = 0$ dynamics (see below).

4.2. Transitions near the 'tricritical' point

The line of discontinuous transitions is defined by the appearance of a double zero of g at non-zero m, i.e. when the maximum in figure 4, curve (3) crosses $g = 0$. Thus, in addition to the equation $g(m) = 0$ we must satisfy $g'(m) = 0$, both at $m \neq 0$. These are two equations for the three unknowns m, β and κ, whose solution is the equation for the line of discontinuous transitions, $\beta_c(\kappa)$, and for the discontinuity in the retrieval amplitude $\Delta m(\kappa)$.

To investigate the neighbourhood of the 'tricritical' point, where m is small on both the continuous and the discontinuous sides, we expand (16) for small m, up to fifth order. Analysing, in the appendix, the expanded map we find that the retrieval amplitude m grows as $\sqrt{\Delta T}$, as one goes below the line of continuous transitions†. The line given by (18) can be evaluated beyond the 'tricritical' point and all the way to $T = 0$. Yet when $g'''(0) > 0$, as one lowers T or raises β from the state of no retrieval, two non-zero roots appear before (at higher temperature) $g'(0)$ vanishes. In the language of the corresponding thermodynamic analogue: the first-order transition takes place when the zero field susceptibility is still positive. Thus, two lines start at the 'tricritical' point. They are the full and the dotted curves in figure 2, for example. The first is the line of discontinuous transitions and the second is the continuation to $T = 0$ of the line of instability at $m = 0$. Near the 'tricritical' point the two lines can be computed and compared, see the appendix. The result is that the two lines start with equal slopes and diverge at second order. This implies, *inter alia*, that the line of discontinuous transitions has the same slope as the actual line of continuous transitions.

Returning to the Gardner case, using $\rho_\kappa(\Lambda)$ of [11,18], we find the continuous transition line from (18) as

$$\frac{\kappa}{2}\left(1+\operatorname{erf}\left(\kappa/\sqrt{2}\right)\right)+\frac{\exp(-\kappa^2/2)}{\sqrt{2\pi}}=\frac{1}{\beta\int_{-\infty}^{\infty}\mathrm{D}y\ \operatorname{sech}^2(\beta y)} \tag{24}$$

from which one can construct the dotted curve in figure 2. Moreover, expanding the left-hand side of (24) for large κ and the right-hand side for small β, one finds the relation $T_c = \kappa$. Since the correction to the left-hand side is exponentially small and to the right-hand side it is of relative magnitude β^2, the relation holds over a wide region (see e.g. figure 2(*b*)).

The condition given by (23) for the tricritical point becomes

$$\left(1+\operatorname{erf}(\kappa/\sqrt{2})\right)(3\kappa-\kappa^3)+\sqrt{2/\pi}\exp(-\kappa^2/2)(1-\kappa^2)=0. \tag{25}$$

Equations (24) and (25) can be solved numerically to give the tricritical point as $\kappa_{tr} = 1.700$, $\beta_{tr} = 0.909$ which corresponds to $\alpha_{tr} = 0.258$, $T_{tr} = 1.100$ in the phase diagram, figure 2.

4.3. The transition from wide to narrow retrieval

The dotted curve in figure 2 was described above as the continuation of the line of continuous transitions. Its dynamical significance is that in the region of retrieval it separates two sub-regions: to its left there is 'wide retrieval', marked (2) in figure 4(*b*),

† At the tricritical value of α the growth of m is like $\Delta T^{1/4}$.

and to its right is a region of 'narrow retrieval', marked (3). At $T = 0$, 'wide retrieval' is delimited by $0 < \alpha < 0.42$ [18] and for $0.42 < \alpha < 2$ retrieval is narrow [18]. As T is raised, the region of 'wide retrieval' shrinks until at the temperature of the tricritical point it is limited to $0 < \alpha < 0.258$.

The identification of the three regions in figure 4(b) with the three types of map dynamics is justified by the fact that on crossing the full curve the form of g must change from that of figure 4(a) curves (1) to (3), as was explained in the discussion of the continuous transition. Similarly, upon crossing the broken curve, g must change from the form figure 4(a) curve (1) to (2), as is implied by the continuous transition. The changes in the form of g as implied by the transition type coincide with the classification of the retrieval dynamics into no retrieval, 'wide retrieval' and 'narrow retrieval'. It remains to be shown that the subdivision of the retrieval region is made by the dotted curve, i.e. the line $g'(0) = 0$, together with $g'''(0) > 0$.

To see this recall that below the dotted curve is a region in which $g'''(0) > 0$, since $\alpha > \alpha_{tr}(= 0.258)$. Thus above (and near) this curve $g'(0) < 0$ and the form of g is that of (3), with the middle (unstable) fixed point approaching $m = 0$, as one approaches the line $g'(0) = 0$. Upon crossing this line $g'(0)$ becomes positive, leading to the form (2). The basin of attraction reaches the full interval (0–1) continuously as the dotted curve is approached.

It is interesting to observe that part of the region which at $T = 0$ gave wide retrieval, i.e. $0.258 < \alpha < 0.42$, gives narrow retrieval as T is raised. The precursor of this change can be detected already in the form of g at $T = 0$. Note that along the $T = 0$ axis, as α is increased there is a changeover, at 0.258, from a shape qualitatively like (2) to (3), although there is still only one finite fixed point. Yet, for low α, $g'''(T = 0, m = 0) < 0$, and since $g'(0) > 0$, the function $g(m)$ is convex, for small m. But, g''' changes its sign at the value of α (or κ) which corresponds to the 'tricritical' point. This condition, (23), is independent of T. As one crosses this point, g at $T = 0$, becomes concave in the neighbourhood of $m = 0$, and eventually as α crosses the value at which

$$\langle \Lambda \rangle = \sqrt{\pi/2} \tag{26}$$

where $g'(T = 0, m = 0)$ vanishes, g takes on the form (3). To summarise, for $\langle \Lambda \rangle > \sqrt{\pi/2}$, as the temperature is raised, the transition to no retrieval will be continuous or discontinuous depending on whether $g(m, T = 0)$ is concave or convex, respectively, in the neighbourhood of $m = 0$.

5. Dilute networks with minimised error number

Having investigated the behaviour of a network which optimally stores random patterns without errors in the presence of fast noise, it is natural to proceed to investigate networks in which random patterns are stored with a given fraction, f, of the stability constraints, (8), violated. One way of describing such networks was put forth by Gardner and Derrida [19]. In this approach one searches for the normalised coupling matrices which minimise the cost function

$$C = \sum_{\mu=1}^{p} \Theta(\kappa - \Delta^{\mu}) \tag{27}$$

for a given set of random patterns, where Δ^μ is given by (7). The actual computation of the volume in interaction space proceeds by first observing that it can be fully mapped on the original Gardner [10] volume computation, if the stability constraints are modified according to:

$$\Theta(\Delta^\mu - \kappa) \rightarrow G(\Delta^\mu - \kappa) = \exp[-h\Theta(\kappa - \Delta^\mu)] \tag{28}$$

and then taking the limit $h \rightarrow \infty$, to project the matrices of minimal error. This limiting procedure is slightly subtle [19], as it must follow the rate at which $q \rightarrow 1$, namely

$$h = \frac{x^2}{2(1-q)}. \tag{29}$$

When the optimal storage allows for violations of the stability condition at some sites, the question of the existence of attractors becomes non-trivial, even in the absence of noise ($T = 0$). This is particularly true because of the special way in which the local fields are distributed in the Gardner–Derrida scheme. It might have been the case that all the violations of the stability condition produce sites at which, while $\Delta^\mu < \kappa$ still $\Delta^\mu > 0$. Such a situation would have left the state with $m = 1$ a fixed-point attractor at $T = 0$. But as we shall see below the distribution has the form displayed in figure 1 and hence any site which violates the stability condition actually produces an error in the first dynamical step. Errors can converge or proliferate on subsequent steps, even at $T = 0$.

The dynamics of the dilute network is described again by (10). In particular, for $T = 0$ this equation reduces to

$$m(t+1) = \int_{-\infty}^{\infty} \mathrm{d}\Lambda\, \rho_\kappa(\Lambda)\, \mathrm{erf}\left[m(t)\Lambda/\sqrt{2[1-m^2(t)]}\right]. \tag{30}$$

For both zero and finite temperature we need the distribution function ρ_κ.

To obtain this distribution function we resort to the general results of Abbott and Kepler [20], who give ρ for any constraint function on the stability parameters†. Their central result is worth recapitulating: if on the right-hand side of (28) one writes $G(\Delta^\mu - \kappa)$, to replace the no error constraint, then the function

$$A(z,y,q) = \frac{G(y)\exp[-(z-y)^2/2(1-q)]\exp(-z^2)/2q}{\sqrt{2\pi q}\int_{-\infty}^{\infty}\mathrm{d}x\, G(x)\exp[-(x-z)^2/2(1-q)]} \tag{31}$$

determines both α_{P} and ρ. One has

$$\alpha_{\mathrm{P}}^{-1} = \int_{-\infty}^{\infty}\mathrm{d}z\int_{-\infty}^{\infty}\mathrm{d}y\,(z-y)^2\, A(z,y,1) \tag{32}$$

and

$$\rho(\Lambda) = \int_{-\infty}^{\infty}\mathrm{d}z\, A(z,\Lambda,q). \tag{33}$$

† Recently Wong and Sherrington [22] have developed another general technique for the study of optimisation in interaction space.

Hence, given $G(y)$, the calculation reduces to the evaluation of the integral

$$B(z) \equiv \int_{-\infty}^{\infty} \mathrm{d}x\, G(x) \exp[-(x-z)^2/2(1-q)] \tag{34}$$

in the limit $q \to 1$.

In the present context, with $G(y)$ given by (28), this integral is

$$B(z) = \Theta(z-\kappa)\left(\sqrt{2\pi(1-q)} + \mathrm{e}^{-h}\frac{1-q}{z-\kappa}\exp[-(\kappa-z)^2/2(1-q)]\right) + \Theta(\kappa-z)\left(\frac{1-q}{\kappa-z}\exp[-(\kappa-z)^2/2(1-q)] + \mathrm{e}^{-h}\sqrt{2\pi(1-q)}\right) \tag{35}$$

for $q \approx 1$. When this expression is substituted in A, and the limits $q \to 1$ and $h \to \infty$ are taken at a fixed value of x^2, (29), one finds:

$$A_\kappa(z,y,1)) = \frac{\exp(-z^2/2)}{\sqrt{2\pi}}[\Theta(z-\kappa)\delta(z-y) + \Theta(\kappa-x-z)\delta(z-y) + \Theta(z-\kappa+x)\Theta(\kappa-z)\delta(y-\kappa)] \tag{36}$$

which, in turn, substituted in (33), gives

$$\rho_\kappa(\Lambda) = \frac{\exp(-\Lambda^2/2)}{\sqrt{2\pi}}[\Theta(\Lambda-\kappa) + \Theta(\kappa-x-\Lambda)] + \delta(\Lambda-\kappa)\int_{\kappa-x}^{\kappa} \mathrm{d}z\,\frac{\exp(-z^2/2)}{\sqrt{2\pi}} \tag{37}$$

which is the field distribution function shown in figure 1. The value of α_{P} and the fractional number of stability errors, f, are given as in [19], namely

$$\alpha_{\mathrm{P}}^{-1} = \int_{\kappa-x}^{\kappa} \mathrm{D}z\,(z-\kappa)^2 \tag{38}$$

and

$$f = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{x-\kappa}{\sqrt{2}}\right)\Theta(x-\kappa)\right]. \tag{39}$$

Note the sharp gap, $\kappa - x < \Lambda < \kappa$, in the field distribution. It is a result of the particular cost function employed. This interval may not be present or may fill up if the cost function becomes more sensitive to the size of the errors, or if the training patterns contain errors [22]. Figure 1 indicates that, for a Gaussian distribution, the sites at which errors in the stability parameter occur violate the stability in a maximal way. In other words, the fields at those sites, when the network is in a pattern, are as large as possible and opposite in sign to the pattern. The corresponding cost function (27) does not penalise for it. Clearly, an error contributes one to the cost irrespective of the size of the constraint violation. The distribution will also become smoother if the storage were not pushed to its extreme., i.e. $q < 1$.

The dynamical equation corresponding, in a dilute network and parallel dynamics, to this field distribution is:

$$m(t+1) = \left[\int_{\kappa}^{\infty} + \int_{-\infty}^{\kappa-x}\right]\mathrm{D}\Lambda \int_{-\infty}^{\infty} \mathrm{D}y\,\tanh\left[\beta\left(m\Lambda + \sqrt{1-m^2}y\right)\right] + \int_{-\infty}^{\infty}\mathrm{D}y\,\tanh\left[\beta\left(m\kappa + \sqrt{1-m^2}y\right)\right]\left[\mathrm{erf}\left(\frac{\kappa}{\sqrt{2}}\right) + \mathrm{erf}\left(\frac{x-\kappa}{\sqrt{2}}\right)\right]. \tag{40}$$

5.1. Retrieval in networks with errors in learning

We now consider the implications for attractors of a network of the Gardner–Derrida type discussed in the previous section. The formula for $\alpha_P(\kappa, f)$ of Gardner–Derrida shows a dramatic increase in the number of patterns that can be learnt at each site if a fraction f of the patterns are learnt incorrectly. For example, for $f = 0.1$ and $\kappa = 0$, $\alpha_P \approx 5.7$, if we extrapolate the replica symmetric results†. An error at a site implies that the imposed stability condition (8) is violated at that site. From figure 1 we can see that for the Gardner–Derrida cost function the constraint violations ($\Lambda < \kappa$) always give stability violations ($\Lambda < 0$), so that we are not guaranteed an attractor, even at $T = 0$, as would be the case if all $0 < \Lambda < \kappa$. If the maximal number of patterns per neural connection stored in a perceptron with error fraction f is given by $\alpha_P(\kappa, f)$, one would like to know what would be the maximal α, $\alpha_c(\kappa, f, T)$, for which attractors still exist, for field distributions of the form figure 1.

Figure 5 presents the results of solving the fixed point equations of (40). Each point in the α–κ plane corresponds to a value of f where $\alpha = \alpha_P(\kappa, f)$. Several curves of constant f are drawn in the figure. The broken curves are lines of constant critical temperature. A point on such a line gives:

1. The temperature at which the retrieval attractor of the map (40) is destabilised for the particular values of α and κ.
2. The error fraction, $f(\kappa, \alpha)$, corresponding to optimal perceptron storage.

No distinction is made in the figure between continuous and discontinuous transitions to no retrieval. From figure 5 one can read off $\alpha_c(\kappa, T)$ which is the value of α beyond which learning with errors, at fixed κ, no longer gives attractors at temperature T. Note that to each $\alpha_c(\kappa, T)$ corresponds a value of f, so once the critical α has been crossed there are no attractors because at fixed κ an increase in α leads to an increase of f. To obtain $\alpha_c(\kappa, T)$ one draws a line parallel to the α axis, at the chosen value of κ, to find the intersection with the chosen temperature curve. One then projects down to the α axis to find α_c. To obtain the error fraction corresponding to α_c, one interpolates between neighbouring curves of constant f.

An alternative way of reading figure 5 is to select a value of f and a temperature and to find the minimum value of κ that can support this error fraction and still give an attractor at that temperature. This is done by finding the intersection between the relevant f and T curves.

5.2. The noiseless case

For $\kappa = 0$ the Gardner result, that the perceptron of connectivity C can store as many as $2C$ random patterns, implies that an ANN with connectivity C can have as many as $2C$ uncorrelated fixed point attractors at zero temperature. It also implies that errors cannot possibly increase this number. If at $T = 0$ the attractors of [8] are fixed points, then every attractor is also a configuration stored by the perceptron. But the perceptron cannot store more than $2C$ uncorrelated patterns, and the dynamics of the dilute network does not generate correlations [8]. Hence, the possibility of the existence of attractors for $\alpha > 2$ is ruled out, even when the error fraction allows

† The behaviour of α_P is reduced to a numerical example because the figures to be displayed below are restricted to $\alpha \leq 2$. Higher values of α are irrelevant for retrieval (see below).

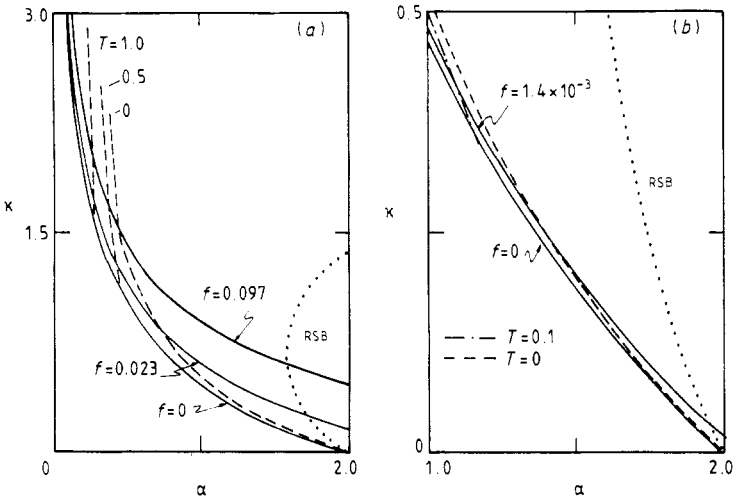

Figure 5. Phases of retrieval attractors for perceptron training with errors. Full curves: perceptron storage capacities α_P against training stability constraint κ for several values of the error fraction f. Broken curves: retrieval storage capacities α_c against κ for several values of the synaptic noise parameter T. The dotted curve delimits to its right the region in which replica symmetry breaks. (b) is an expansion of the large α small κ corner of (a), for very small error fractions.

$\alpha_P > 2$. However, for learning at a fixed positive value of κ there is no reason, *a priori*, to exclude the existence of attractors for values of α with $2 > \alpha > \alpha_P(\kappa, 0)$.

At zero temperature and $\kappa = 0$ we see in figure 5 that $\alpha_c = 2$ at $f = 0$ which agrees with our *a priori* argument. However, for $\kappa > 0$, $\alpha_c(\kappa, T = 0)$ is always greater than $\alpha_P(\kappa, f = 0)$, i.e. the Gardner limit, showing that learning with errors at fixed positive κ increases the storage capacity for retrieval. In figure 5(b) the region of low κ is highlighted to show that for small positive κ, α_c is always less than 2 and attractors exist only for extremely low values of f. For example, we can take $f = 1.4 \times 10^{-3}$ and $T = 0$ to find that we must have $\kappa = 0.25$ to have retrieval fixed points. This shows that for low κ learning with errors will be detrimental to the possibility of attractors unless the error fraction is kept extremely small. This is in agreement with recent studies [21], which show that similar two-band distributions for the fields lead to deterioration in retrieval performance.

Finally, it is important to notice in figure 5 that in the region where replica symmetry breaking occurs, there are no attractors present. The point $\alpha = 2$, $\kappa = 0$ is on the boundary. In the entire region where attractors exist replica symmetry is stable.

5.3. The presence of noise

We now consider the effect of finite temperature. At $T = 0$ we have argued that the maximum value of α_c is fixed at 2 and this is given by the perceptron algorithm without errors. At finite T, α_c will be reduced due to the disordering nature of noise. Robustness against temperature demands a sacrifice of capacity for stability. Indeed, figure 2(b) shows that for $f = 0$ in a wide range of κ, almost the entire range of the continuous transitions to no retrieval, $T_c \simeq \kappa$. Thus high temperature requires a high

κ for retrieval. It is in this high κ regime that learning with errors gives a significant increase in α over the Gardner case. For fixed α, learning errors allow a higher value of κ with the sacrifice of stability at a small fraction of sites. One might expect that at finite T this trade-off will allow attractors to be retained at α values at which there are no attractors in the absence of learning errors. In other words, one might expect that for a fixed finite temperature, the maximal α_c will be reached by learning *with* errors. This possibility is studied below.

Returning to figure 5 one observes, inspecting the lines of constant critical temperature, that as T increases they become more vertical. This implies that as T increases, α_c becomes essentially independent of κ, over a wide range of κ. Equivalently, learning with no errors gives approximately the same storage capacity as learning at a higher value of κ and allowing errors. For $T = 1.0$ the fixed temperature curve even bends back slightly on itself, which implies that the maximum α_c for $T = 1.0$ is given by learning with errors. However, the maximal α_c is found to be 0.279 at $f = 0.037$, whereas for no errors $\alpha_c = 0.275$. Thus the improvement in α_c at finite temperature is only marginal.

A more significant improvement is achieved by the dilute Hopfield ANN, for which at $T = 1.0$ $\alpha_c = 0.367$. It is intriguing to note that at this point the field distribution in the Hopfield network is a Gaussian of width 1 and mean $\alpha^{-1/2} = 1.65$. This distribution has a very significant tail of negative fields. The reason for the better performance of the Hopfield network is that at sufficiently high T, it is better to have more nodes with negative fields distributed around zero. In that case the incorrect bits can be more readily flipped to their correct states by the noise. This implies that learning with errors should improve upon the performance of the errorless network, provided the cost function for the errors produces field distributions with negative tails near zero. One such possible cost function is the sum of the *magnitudes* of the errors, rather than their mere number, as used above [24].

In order to compare retrieval phase diagrams of networks trained as perceptrons with and without errors, we present in figure 6 a phase diagram for a network storing patterns with a fixed fraction of errors $f = 0.023$, superposed on the phase diagram figure 2 for which $f = 0$. The two diagrams are qualitatively similar. In figure 6 the continuous transition line for $f = 0.023$ (short chain curve) begins, for $\alpha = 0.0$, at a lower T than the equivalent curve with $f = 0$. The curve with the errors crosses the errorless curve at a temperature that is higher than that of both tricritical points. The wide to narrow retrieval transition lines are the dotted and short broken curves for $f = 0.023$ and $f = 0$, respectively.

The significant points to note are as follows.

(i) In the case of finite, even very small, f the regime of narrow retrieval is drastically curtailed, which reflects the sensitivity of the attractors to errors at low values of κ. This is clearly seen by comparing the discontinuous transition lines for $f = 0$ (full curve) and $f = 0.023$ (long chain curve), e.g. for $T = 0$, α_c is reduced to 0.86 for $f = 0.023$.

(ii) The continuous transition lines remain very close, although the exact positions of the tricritical points may be well separated. This closeness of the lines of continuous transition implies that, at high temperature, α_c does not vary significantly with f, which is reflected in the vertical nature of the $T = 1.0$ curve in figure 5, as discussed above.

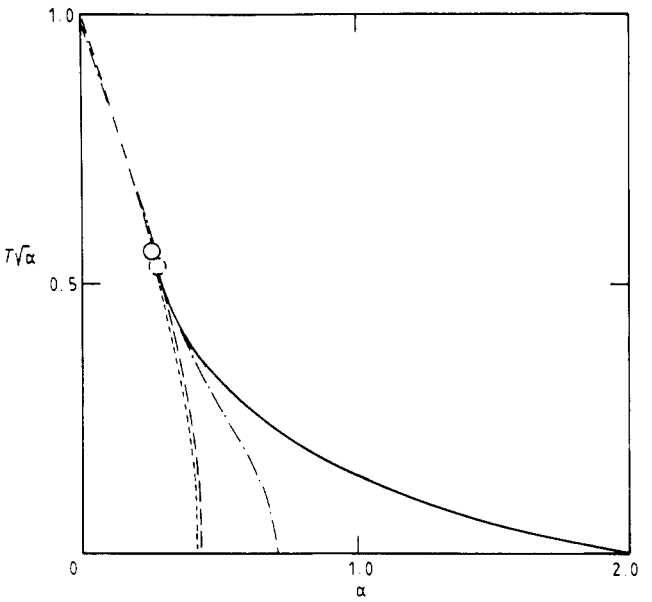

Figure 6. Phases of retrieval attractors for perceptron training with errors $f = 0.023$ and without errors. The conventions for $f = 0$ are as for figure 2. For $f = 0.023$ the long chain curve is the line of discontinuous transitions to no retrieval; the short broken line is the transition from wide to narrow retrieval; the short chain curve is the line of continuous transition to no retrieval; the tricritical point is marked by a dotted circle.

6. Conclusion

We have investigated the effects of fast noise (temperature) on attractor neural networks with optimal interactions which are diluted asymmetrically, and have found three types of behaviour. When both temperature and storage levels are high, the system is unable to retrieve. When either or both of them are lowered, the network enters a retrieval phase, which in turn divides into two sub-regions. Roughly speaking, the major effect of slow noise (pattern interference) is to narrow the basin of attraction. Hence, for high storage level and low temperature, we have a phase of narrow retrieval where the retrieval and non-retrieval attractors coexist, and transition to non-retrieval is discontinuous. On the other hand, fast noise (temperature) reduces retrieval in a way less dependent on the input overlap. As a result, we have a phase of wide retrieval for high temperature and low storage level. The transition to no retrieval is continuous, as is the transition to narrow retrieval.

It is interesting to compare this system with other networks whose phase diagrams are already known. Consider the fully connected Hopfield model [3–5]. There, the transition between retrieval and no retrieval depends in a qualitatively similar way on temperature and storage. At $T = 0$ the phase line reaches $\alpha = 0.14$ [4]. Since the transition is always discontinuous, there is no equivalent to the transition between narrow and wide retrieval. However, even within the retrieval phase, there is a first-order transition when the global minimum of the free energy shifts between the retrieval state and the spin-glass state. At $T = 0$ this phase line reaches $\alpha = 0.05$. It is natural to expect that the basin of attraction of the retrieval state shrinks considerably when

it ceases to be the global minimum, which may be the analogue of narrowing retrieval, although this speculation has yet to be confirmed by further studies. It is also worth recalling that the phase of no retrieval, of the Hopfield model, is further separated into the spin-glass and paramagnetic (ergodic) phases [4]. In the present work we have not studied the analogues of these phases. In any case, here we have dealt with an asymmetric model, for which the analogue of the concept of a spin glass phase has yet to be defined.

The appropriate comparison is between models of the same connection topology, which is why we introduce the dilute, asymmetric Hopfield model in section 3. Diluting and asymmetrising the Hopfield model modifies the phase diagram drastically: the state of no retrieval is no longer metastable relative to the retrieval phase, and the retrieval state has the widest possible basin of attraction right up to its continuous transition to non-retrieval. This is in contrast to the existence of the narrow retrieval phase in the optimal model. We also found that the Hopfield model has a higher storage capacity at high temperature than the model trained to be optimal at $T = 0$. Although the reverse is true at low temperature. This leads us to conclude that the Hopfield model is more 'specialised' to process noisy inputs, and to operate in the presence of extensive ambient noise, whereas the optimal model is more 'specialised' in the opposite situations.

We have also investigated whether it was possible to improve the storage of the optimal model in the presence of fast noise, by increasing the stability at the majority of nodes while allowing errors at the remaining small fraction of nodes. We have found that using the Gardner–Derrida cost function for the errors [19] there is only a marginal improvement at intermediate temperatures, but the storage is dramatically reduced at low temperature; it is impossible to increase the storage of attractors beyond the Gardner $\alpha = 2$ limit [10].

Perhaps the comparison of these models demonstrates some principle of specialisation, namely different models perform better in specific environments, but not in others: the optimal model at low temperature; learning with errors at intermediate temperature; and the Hopfield model at high temperature. Furthermore, these specialisations are characterised by their corresponding field distributions: at low temperature the field distribution should be maximally stable; at intermediate temperature it may be better to have a small tail at the far negative extreme; at high temperature a broad single-band distribution may be more suitable. Thus, although our attempt to use optimal synaptic coding to improve retrieval storage capcity in the presence of fast noise has not succeeded, our study did shed some light on the form of possible solutions [23].

Acknowledgments

We would like to acknowledge the stimulation, encouragement and comments on the manuscript of Professor D Sherrington, which were instrumental to this study. DJA is indebted to the SERC for a fellowship which has made his stay at Imperial College possible and to Dr S Franz for a reminder that replica symmetry breaking is a threat.

Appendix. Expansions about the 'tricritical' point

The function g will be parametrised by β and another parameter κ, which determines the loading level α and may even be α itself. For example, in Gardner's optimal

network [18], κ is the site stability parameter. Or, it may be related to the number of errors allowed in the process of storage (see e.g. section 5), etc. Ultimately we wish to consider a phase diagram in the κ–T or α–T plane.

The condition for a fixed point reads:

$$m = f^{(1)}(\beta,\kappa)m + f^{(3)}(\beta,\kappa)m^3 + f^{(5)}(\beta,\kappa)m^5. \tag{A1}$$

The line of continuous transitions, determined by (18) together with $g'''(m=0) < 0$, gives β, or T, as a function of κ. As T is lowered below this line, at fixed κ, one has the usual mean field result for the developing retrieval amplitude m:

$$m^2 = \frac{f^{(1)}_\beta(c)}{f^{(3)}(c)}\Delta\beta \tag{A2}$$

where the subscript indicates partial differentiation; the variable c implies that the function is evaluated on the continuous line, i.e. at κ and $\beta_c(\kappa)$ and

$$\Delta\beta \equiv \beta - \beta_c(\kappa).$$

At the tricritical point the line of continuous transitions, $g'(0) = 0$, which becomes the line of transitions from wide to narrow retrieval, is expanded as

$$\Delta\beta_c(\kappa) = a\Delta\kappa + b(\Delta\kappa)^2 \tag{A3}$$

and the line of discontinuous transitions to no retrieval as

$$\Delta\beta_{dc}(\kappa) = a_1\Delta\kappa + c(\Delta\kappa)^2 \tag{A4}$$

where $\Delta\kappa = \kappa - \kappa_{tr}$. On substituting the first expansion into (18) one finds:

$$a = -\frac{f^{(1)}_\kappa(\mathrm{tr})}{f^{(1)}_\beta(\mathrm{tr})}$$

$$b = -\frac{a^2 f^{(1)}_{\beta\beta}(\mathrm{tr}) + 2a f^{(1)}_{\beta\kappa}(\mathrm{tr}) + f^{(1)}_{\kappa\kappa}(\mathrm{tr})}{2f^{(1)}_\beta(\mathrm{tr})} \tag{A5}$$

and (tr) indicates evaluation at the 'tricritical' point. The equation for the discontinuous transition, near the 'tricritical' point, reads

$$1 = f^{(1)}(\beta,\kappa) + 3f^{(3)}(\beta,\kappa)m^2 + 5f^{(5)}(\beta,\kappa)m^4. \tag{A6}$$

Combined with the condition for the fixed point, it gives the relation:

$$(f^{(3)})^2 = -4f^{(5)}(1 - f^{(1)}) \tag{A7}$$

which is an equation relating β and κ along the discontinuous transition line. After expanding this expression about the tricritical point at which $f^{(3)} = 0$, $f^{(3)}_\beta = 0$, $f^{(1)} = 1$, and substituting A3 and A4, one finds

$$\begin{aligned} 0 = \Delta\kappa\big[&4f^{(5)}(\mathrm{tr})(f^{(1)}_\kappa(\mathrm{tr}) + a_1 f^{(1)}_\beta(\mathrm{tr}))\big] \\ &+(\Delta\kappa)^2\big[[f^{(3)}_\kappa(\mathrm{tr})]^2 + 4c\big(f^{(5)}(\mathrm{tr})f^{(1)}_\beta(\mathrm{tr}) + f^{(5)}(\mathrm{tr})_\beta[f^{(1)}_\kappa(\mathrm{tr}) + a_1 f^{(1)}_\beta(\mathrm{tr})]\big) \\ &- 8bf^{(5)}(\mathrm{tr})f^{(1)}_\beta(\mathrm{tr}) + 4f^{(5)}_\kappa(\mathrm{tr})\big(f^{(1)}_\kappa(\mathrm{tr}) + a_1 f^{(1)}_\beta(\mathrm{tr})\big)\big]. \end{aligned} \tag{A8}$$

This equation gives

$$a_1 = a \tag{A9}$$

$$c = b - \frac{[f_\kappa^{(3)}(\mathrm{tr})]^2}{4f_\beta^{(1)}(\mathrm{tr})f^{(5)}(\mathrm{tr})}. \tag{A10}$$

With a and b given by A5 and A6. Hence the line of transition from retrieval to no retrieval is continuous and has a continuous slope at the 'tricritical' point. The discontinuity is in the curvature.

References

[1] Little W A 1974 *Math. Biosci.* **19** 101
[2] Peretto P 1984 *Biol. Cybern.* **50** 51
[3] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev.* A **32** 1007
[4] Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30
[5] Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
[6] Weisbuch G and Fogelman-Soulie F 1985 *J. Physique Lett.* **46** 623
[7] Sompolinsky H 1987 *Heidelberg Coll. on Glassy Dynamics* eds J L Van Hemmen and I Morgestern (Heidelberg: Springer)
[8] Derrida B, Gardner E and Zippelius 1987 *Europhys. Lett.* **4** 167
[9] Amit D 1989 *Modeling Brain Function* (Cambridge: Cambridge University Press)
[10] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[11] Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
[12] Krauth W, Nadal J-P and Mézard M 1988 *J. Phys. A: Math. Gen.* **21** 2995
[13] Henkel R and Opper M *Preprint* Giessen
[14] Gardner E, Derrida B and Mottishaw P 1987 *J. Physique* **48** 741
[15] Horner H, Bormann D, Frick M, Kinzelbach H and Schmidt A 1989 *Z. Phys.* B **76** 381
[16] Gutfreund H and Mézard M 1988 *Phys. Rev. Lett* **61** 235
[17] Gardner E, Mertens S and Zippelius A 1989 *J. Phys. A: Math. Gen.* **22** 2009
[18] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969
[19] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[20] Abbott L F and Kepler T B 1989 *J. Phys. A: Math. Gen.* **22** 2033
[21] Wong K Y M and Sherrington D 1989 *Europhys. Lett.* **10** 419
[22] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** L175
[23] Wong K Y M and Sherrington D 1990 *Optimally Adapted Attractor Networks in the Presence of Noise* in preparation
[24] Wong K Y M 1989 unpublished