Enlarging the attractor basins of neural networks with noisy external fields

1991 J. Phys. A: Math. Gen. 24 5639

(http://iopscience.iop.org/0305-4470/24/23/026)

J. Phys. A: Math. Gen. 24 (1991) 5639–5650. Printed in the UK

Enlarging the attractor basins of neural networks with noisy external fields

H W Yau† and D J Wallace

Department of Physics, James Clerk Maxwell Building, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ U.K.

Received 9 April 1991

Abstract. A neural network model with optimal connections trained with ensembles of external, discrete, noisy fields is studied. Allowing for non-zero errors in the storage, novel behaviour is observed which is reflected in the model's retrieval map. Improvement in the model's content addressability is determined by comparing the maximum storage level at which there is a near 100% basin of attraction. The cases presented here have the external field applied during training, during retrieval, and during both with statistically equal parameters. In all three the content addressability is improved over the zero external field network, with the equal training and retrieval fields case having the largest improvement. However, the apparent domination of the retrieval over the training field suggests this simple equality is perhaps not the optimal relationship.

1. Introduction

The key feature of statistical mechanical models of neural networks is their ability to function as associative memories. This is a two-stage process with the network first *trained* to store a set of memory patterns, which are then later *retrieved* by the neurons' update dynamics. Retrieval of the desired pattern will be successful if the system is initiated sufficiently close to it. That is, if it is started inside the *basin of attraction* of the memory pattern's *attractor*. Expressed in this way, content addressability is merely the consequence of having finite basins of attraction.

The principle aim of this work is to examine how the basin of attraction can be enlarged by the use of *external fields* which are noisy representations of the memory patterns stored. Independent work has shown the beneficial effects of applying noisy external fields throughout retrieval (Engel *et al* 1990) but as stated above, a network is defined in two stages and their role during the training phase should be explored. Moreover, both simulation (Gardner *et al* 1989) and analytical (Wong and Sherrington 1990a,b) results have shown that training a network with *ensembles* of noisy representations also improves content addressability.

For these reasons this work calculates the properties of a network trained with ensembles of noisy external fields. The retrieval dynamics under a persistent, noisy external field is then examined, and the effects of the two fields compared. By looking at the fixed-point behaviour of the dynamics, the attractor structure is revealed, and from this it is judged whether content addressability has been improved. Finally

† E-mail: hwyau@edinburgh.ac.uk

0305-4470/91/235639+12$03.50

comparisons are made for the three cases when external fields are applied during training only, during retrieval only, and during both stages.

2. The model

The model is a single-layered network of N time-dependent binary spin neurons $S_i(t) = \pm 1, i = 1 \ldots N$, required to store P uncorrelated patterns $\{\xi_i^\mu\}, \mu = 1 \ldots P$. Each neuron is dilutely connected to $C \ll \ln N$ other sites via the matrix $J_{ij}/\sqrt{C}$ and normalized by $\sum_j^C {J_{ij}}^2 = C$. The high degree of dilution in the bonds is necessary for the dynamics at each time step to be self-averaging over all time steps, considerably simplifying its calculation (Derrida *et al* 1987).

The dynamics of the network is conducted by zero-temperature parallel update, with each site acting deterministically on the sign of its local field. Retrieval of a typical pattern $\{\xi_i^\nu\}$ is measured at each time step t by the overlap

$$m^\nu(t) \equiv \frac{1}{N} \sum_i^N \xi_i^\nu S_i(t) \tag{1}$$

such that $m^\nu(t \to \infty) = 1$ when $\{S_i(t \to \infty)\} = \{\xi_i^\nu\}$.

In deference to Gardner, the network studied has anneal-optimized connections with the stored patterns quenched-disorder averaged (Gardner 1988). These connections are optimized with respect to a *performance function*, in much the same way a magnetic spin system optimizes by seeking out its lowest-energy configuration. Moreover, Gardner gave a convergent iterative algorithm to actually train the network with these optimal connections, one which directly reflects the performance function used. Hence the performance function can, and will be, intuitively better referred to as the *training function*, emphasizing its role in determining the network's properties.

The training function chosen in the original Gardner model required the network to be invariant to the neurons updating, once the sites match a memory pattern. This is more concretely expressed by defining the *stability field*

$$\Lambda_i^\mu \equiv \xi_i^\mu \sum_j^C \frac{J_{ij}}{\sqrt{C}} \xi_j^\mu \tag{2}$$

and requiring it to be positive definite for all $i = 1 \ldots N$ sites and $\mu = 1 \ldots P$ patterns. This demand can be made more stringent by requiring it to be larger than some positive constant κ, the increasing of which allows the basin of attraction to be enlarged at the price of lowering the storage capacity (Forrest 1988).

Further enlarging the basin of attraction by an improved training function is the goal of this work, with the additions influenced by two observations: firstly, that a network trained with the original Gardner requirement has improved retrieval upon adding a noisy external hint of the pattern to be retrieved; and secondly that training with *ensembles* of noisy versions of the memory patterns improves the network's content addressability to the clean patterns. The desire then is to train the network with ensembles of noisy external fields, in anticipation of later retrieval with a (statistically) similar field. This can be achieved with the training function

$$g = \sum_\mu^P \frac{1}{Q} \sum_s^Q \theta[\Lambda^\mu + \tau_T \zeta_i^{\mu,s} - \kappa] \tag{3}$$

where τ_T is the external training field strength and $\{\zeta_i^{\mu,s}\} = \pm 1$ the noise factor. These quenched-disorder noise terms are enumerated over the $s = 1 \ldots Q$ ensembles for each pattern, and follow the probability distribution

$$\mathcal{P}(\zeta) = (1 - f_T)\delta[\zeta - 1] + f_T\delta[\zeta + 1] \tag{4}$$

with the mean fraction of wrong bits in the training field given by f_T.

This completes the definition of the model, and the remainder of this section is devoted to the relevant quantities calculated.

The first quantity to look at is the distribution of the stability field as defined in equation (2). It transpires that the key characteristics of a network are largely given by this distribution (Abbott 1990), which as an extensive quantity is also a suitable object over which to perform the quenched-disorder averages of the patterns $\{\xi_i^\mu\}$ and noises $\{\zeta_i^{\mu,s}\}$. Dropping the superfluous i-index, the distribution for a typical pattern $\{\xi^{\mu=1}\}$ can be written in terms of the performance function (3) as

$$\rho(\Lambda) = \left\langle\!\!\left\langle \frac{\int \prod_j \mathrm{d}J_j \ \delta[\Lambda - \Lambda_1]\exp(\beta g)}{\int \prod_j \mathrm{d}J_j \ \exp(\beta g)} \right\rangle\!\!\right\rangle \tag{5}$$

where the integration is throughout the spherically constrained connections space, and the $\langle\langle \cdots \rangle\rangle$ brackets denote a quenched average over all patterns and noises. The inverse annealing temperature β controls how 'strictly' the training function should be obeyed, such that in the $\beta \to \infty$ zero-temperature limit the connections will find the optimal solution. Averaging over the $\{\zeta^{\mu,s}\}$ ensembles of noises, the training function becomes a multitude of $(Q + 1)$ binomial terms which in the large-Q (but $\ll \beta$) limit can be replaced by the mean

$$\sum_{\mu}^{P}\{(1 - f_T)\theta[\Lambda^\mu + \tau_T - \kappa] + f_T\theta[\Lambda^\mu - \tau_T - \kappa]\}. \tag{6}$$

The distribution function is then calculated by assuming the replica-symmetric ansatz, and taking the large connectivity and zero annealing temperature limits (Gardner and Derrida 1988) to give

$$\rho_{\kappa,x}(\Lambda) = \int_{-\infty}^{\infty} \frac{\mathrm{d}z}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)\delta[\Lambda - \hat{\lambda}(z)] \tag{7}$$

(Wong and Sherrington 1990a) where $\hat{\lambda}(z)$ maximizes the function

$$(1 - f_T)\theta[\lambda + \tau_T - \kappa] + f_T\theta[\Lambda - \tau_T - \kappa] - \frac{1}{x^2}[z - \lambda]^2. \tag{8}$$

The new variable x is a result of the mathematics, related to the annealing temperature in determining how strictly the training function is to be enforced. It can by connected with a meaningful quantity by parametrizing the fraction of wrong bits to be stored per pattern $\mathcal{F}$ as the integral over the unwanted part of the stability distribution

$$\mathcal{F}_{\kappa,x} = \int_{-\infty}^{0} \mathrm{d}\Lambda \ \rho_{\kappa,x}(\Lambda). \tag{9}$$

Evaluation of the integrals over the mean-field order parameters is done by the method of steepest descent, resulting in a saddle-point equation which gives an expression for the storage capacity (patterns stored ÷ connectivity)

$$\alpha_{\kappa,x} \equiv \frac{P}{C} = \left[\int_{-\infty}^{\infty} \frac{\mathrm{d}z}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)[z - \hat{\lambda}(z)]^2\right]^{-1}. \tag{10}$$

One obvious reservation to all the above is the stability of the replica-symmetric ansatz. As has been pointed out before (Gardner and Derrida 1988) the structure of the relevant stability matrices are pretty much identical to those considered earlier (de Almeida and Thouless 1978) in the treatment of spin-glass models. From this the regime where this ansatz is stable has be calculated for a general training function (Griniasty and Gutfreund 1991), and this result will be used to check the calculation's validity.

Having obtained these properties of this network trained in the presence of noisy external fields, the model's retrieval dynamics can finally be written down. The use of diluted connections allows this to be calculated as a time-iterative map of the overlap measure, with respect to an arbitrary pattern $\{\xi_i^\nu\}$

$$m^\nu(t+1) = \int_{-\infty}^{\infty} \mathrm{d}\Lambda \rho_{\kappa,x}(\Lambda)\left\{(1-f_R)\mathrm{erf}\left[\frac{m^\nu(t)\Lambda + \tau_R}{\sqrt{2(1-(m^\nu(t))^2)}}\right] + f_R\mathrm{erf}\left[\frac{m^\nu(t)\Lambda - \tau_R}{\sqrt{2(1-(m^\nu(t))^2)}}\right]\right\}. \tag{11}$$

This equation contains additional parameters to do with the application of a persistent external field during retrieval. This field is a noisy representation of the pattern to be retrieved, with f_R the mean fraction of erroneous bits, and τ_R the field strength. Taking the field strengths τ_T and τ_R, and the fractional storage error $\mathcal{F}$ all to zero restores the retrieval dynamics for the original Gardner model (Kepler and Abbott 1988).

The above expression (11) is exact for the first time step, and also in the case of low connectivity as mentioned earlier in this section. The validity of this simplification has recently been confirmed by direct numerical simulations (Heindrich 1991) of dilute netoworks. Furthermore, earlier work simulating fully connected networks (Kepler and Abbott 1988) has stressed the importance of the first time step dynamics as being highly indicative of the network's ultimate fate, a result with broadens the generality of these iterative map calculations.

In summary, the distribution of the stability fields for a network trained with ensembles of noisy external fields is calculated. From this equation for the storage capacity α, the fractional storage error $\mathcal{F}$, and the iterative map for the dynamics are given. The first two are treated as parameters for relating to the somewhat non-intuitive pair (κ, x). Finally, another two sets of parameters determine the fraction of noise (f_T and f_R) and strength (τ_T and τ_R) of the external fields, as used in the training and retrieval phases.

3. The stability field distribution

Before delving into the model's dynamics, it is worthwhile to first examine the stability field distribution (7). This distribution gives an indication of the training function's

effect as the following parameters are varied: the training field noise level f_T, the field strength τ_T, the stability constant κ, and the fraction of storage error $\mathcal{F}$.

Upon maximizing equation (8), it becomes apparent that the network's properties are not uniform throughout the range of the three parameters listed above. These parameters interplay in the mean training function (6) to produce three distinct regimes of behaviour in the network, as exemplified by the stability field distributions shown in figures 1(*a*)–1(*c*). Intuitively, these three regimes are a reflection of how well the two terms in the training function—a 'correct' $(1-f_T)\theta[\Lambda^\mu+\tau_T-\kappa]$ term and an 'incorrect' $f_T\theta[\Lambda^\mu-\tau_T-\kappa]$ term—are satisfied. This is affected by the fractional storage error because demanding it to be low is related to setting a strict adherence to the training function. As for the external training field parameters, the noise level weights between the two terms, while the actual 'difference' between the competing step-functions is simply the noise strength $2\tau_T$.

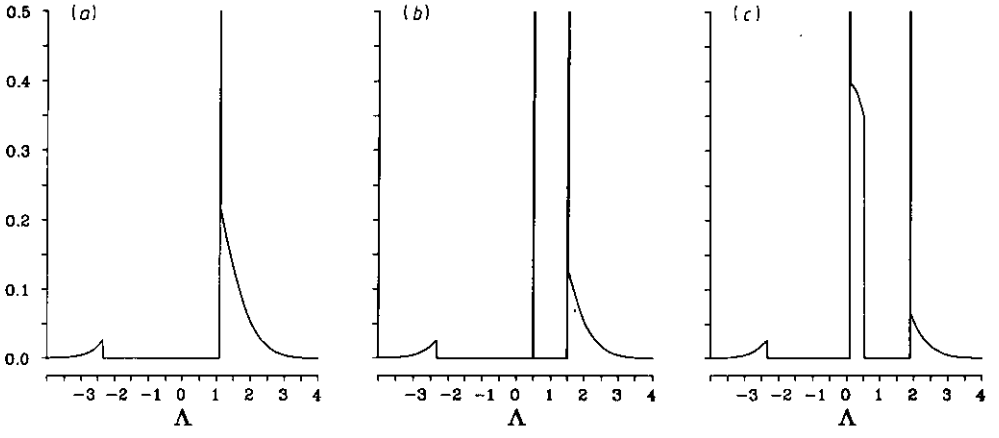

Figure 1. Stability field distributions for fixed storage error $\mathcal{F}=1\%$ and stability requirement constant $\kappa=1.0$. The three regimes shown are for three increasing training field strengths $\tau_T=0.10$ (*a*), 0.50 (*b*) and 0.90 (*c*) with the noise level fixed at $f_T=0.24$. Conversely, these these same regimes can be qualitatively reproduced with the noise levels decreasing $f_T=0.50$ (*a*), 0.30 (*b*) and 0.10 (*c*) and the field strength fixed at $\tau_T=0.50$; the major difference being the location of the delta-peaks at $\kappa\pm\tau_T$.

Figure 1(*a*) shows that the first regime is essentially the original Gardner training function distribution with a non-zero fractional storage error (Amit *et al* 1990), except for a trivial rescaling in the stability requirement constant $\kappa\rightarrow(\kappa+\tau_T)$. This occurs when the 'erroneous' f_T term in the training function (6) dominates over the 'correct' $(1-f_T)$ part. This happens for low storage errors $\mathcal{F}$, as strictly requiring the erroneous part to be trained automatically ensures adherence of the easier correct part. Indeed, for the strictest zero storage error case, the distribution is the same as the original Gardner case regardless of the other parameters and no new behaviour will occur. With a finite storage error, this figure 1(*a*) regime can still be visited by a low enough external field strength τ_T and/or a high field noise level f_T—the former because there is then little difference between the two training terms, and the latter because the function becomes weighted in favour of the erroneous part.

Novel behaviour is shown in figure 1(*b*), which is made by fixing the storage error to a non-zero value while further raising the external field strength and/or lowering the fractional noise. This new regime has an additional delta peak at $(\kappa-\tau_T)$ and

is the manifestation of stability fields that satisfy the correct term in the training function, but not the tougher erroneous term, as aided by suitably weighting towards the former.

Further raising and/or lowering the external field parameters produces the final regime shown in figure 1(*c*). Here the correct term in the performance function continues to grow in importance over the erroneous part, expanding its share of the distribution at the expense of the delta peak at $(\kappa + \tau_T)$. As the external field strength further increases, this diminishing delta peak eventually disappears and the curve returns to the Gardner case but with the rescaling $\kappa \rightarrow (\kappa - \tau_T)$.

4. Critical-points of the dynamics

The principle aim here is to discover the attractor structure of the network's retrieval dynamics, and this is achieved by numerically seeking out the fixed-points of the iterative map equation (11) for the overlap measure m. The stable solutions indicate attractor centres, while unstable ones are defined as the attractor boundaries. From these two values of the overlap, the fidelity of the attractor to its training pattern and the size of the basin of attraction are revealed.

Ignoring the external field parameters, the network's iterative dynamics is essentially described by equation (11), for the storage capacity α in equation (10) and fractional storage error $\mathcal{F}$ of (9). The effect of the training and retrieval external fields will be discussed in detail in subsequent subsections, but for now the focus is on the effects of varying the storage capacity and storage error.

The storage capacity α is a particularly interesting quantity to examine since it gives a readily accessible indication of a network's performance. Using the original Gardner case with no external fields and zero error in the storage, two important phases in the storage-overlap (α, m) space can be identified.

At very low storage capacities the basin boundary has an overlap of essentially zero, implying retrieval of the pattern is guaranteed for any positive, infinitesimal initial overlap. As the number of patterns stored by the network is increased, a storage load is reached where the attractors to each memory pattern can no longer avoid the others, and shrink their basin boundaries in response. This storage load signals the upper bound for the region of *wide retrieval*, beyond which a macroscopic initial overlap is necessary for retrieval.

Further increase in the storage load squeezes the basin of attraction until it and the attractor vanish altogether, making retrieval of a pattern is impossible. Between this saturation point and the end of the wide retrieval region is the region of *narrow retrieval.*

For the cases to be examined, much emphasis is placed on how these regions of wide and narrow retrieval are affected by the external fields. Operationally, these are marked out by points in the storage-overlap space where stable and unstable fixed-points meet. Henceforth they will be referred to as *critical points* and denoted by hats: $(\hat{\alpha}, \hat{m})$, such that the region of wide retrieval is from $\alpha = 0$ to the critical point $(\hat{\alpha}_0, \hat{m}_0)$, and the region of narrow retrieval from $(\hat{\alpha}_0, \hat{m}_0)$ to $(\hat{\alpha}_1, \hat{m}_1)$.

For the zero storage error Gardner case, the wide- and narrow-retrieval regions are bounded by $(\hat{\alpha}_0 = 0.42, \hat{m}_0 = 0.0)$ and $(\hat{\alpha}_1 = 2.0, \hat{m}_1 = 1.0)$ respectively (Gardner 1989). From zero storage to the end of the narrow-retrieval region there is a stable fixed-point at $m = 1$ for the memory attractor, but for storage $\alpha > \hat{\alpha}_0$ there is another spurious attractor with overlap $m = 0$.

The fraction of erroneous bits per pattern stored is parametrized by equation (9). The effect of increasing this value $\mathcal{F}$ is to decrease the quality of the memory attractor, and a corresponding squeezing of the narrow-retrieval region (Amit *et al* 1990). This simple (and somewhat unproductive) response to storage error is qualitatively reproduced for this model at errors of 0.1, 1, 5, and 10% of the bits per pattern. Consequently, in the cases that follow the storage error is always fixed at the somewhat arbitrary value of 1%.

Finally this leaves the external field parameters to examine. The cases presented in the following subsections look at the effect on the critical points with a training field (f_T, τ_T) only, with a retrieval field (f_R, τ_R) only, and with statistically equal training and retrieval fields $(f_T = f_R, \tau_T = \tau_R)$. The principal means of showing their effect is by looking at the regions of wide and narrow retrieval, by plotting the critical points' overlap (– – –) and storage (——).

4.1. Training field only

Figures 2(*a*) and 2(*b*) show the critical-points as the training noise strength is increased, for two levels of training noise at $f_T = 0.20$ and $f_T = 0.24$, respectively.

The straightforward behaviour shown in figure 2(*a*) is typical for low noise levels (below $f_T \sim 0.05$ the network is essentially insensitive to the training field). Given a sufficient noise level, however, the improved content addressability of the system becomes readily apparent by the increased size of the wide-retrieval region as bounded by the critical point $(\hat{\alpha}_0, \hat{m}_0)$, peaking at $\hat{\alpha}_0 = 0.52$ for a field strength $\tau_T = 0.52$. As the number of patterns that can be retrieved with a microscopic initial overlap has increased, the basin of attraction can therefore be said to have widened. Unfortunately this improvement is seemingly at the expense of the narrow-retrieval region, whose decrease worsens the number of patterns that can be stored before saturating the network.

Increasing the training noise above $f_T \sim 0.21$ complicates the critical-point plots considerably, as the typical example at $f_T = 0.24$ in figure 2(*b*) shows.

For low field strengths up to $\tau_T \sim 0.36$ the wide- and narrow-retrieval regions increase and decrease respectively, as in the example discussed above. Additional structures indicating new critical points appear as the field strength is further increased, and are best understood by referring to their fixed-point diagrams. The important 'snapshots' for this example are given in figures 2(*c*)–2(*e*), which plot the fixed-points against increasing storage levels, for increasing field strengths.

Figure 2(*c*) shows that as the field strength is increased the $(\hat{\alpha}_1, \hat{m}_1)$ critical-point marking the narrow-retrieval region starts to merge into the wide-retrieval point $(\hat{\alpha}_0, \hat{m}_0)$. Meanwhile two new critical points appear near overlap $m \sim 1$, indicating the formation of a stable attractor in addition to the memory pattern's, but of lower fidelity. In figure 2(*d*) the old $(\hat{\alpha}_1, \hat{m}_1)$ critical-point vanishes but its role marking out the region of wide-retrieval is quickly taken over by one of the two new critical points. Finally in figure 2(*e*) the other critical point coalesces into the wide-retrieval point, taking with it the extraneous attractor.

The extra structure brought along by the appearance of the extra attractor seems to have little practical consequence. Indeed if anything it appears to degrade the region of wide retrieval over some range of external field strengths, and hence may be considered as indicative of the maximum noise level to choose.

This structure is not, however, a result of instability in the replica-symmetric solution and cannot be dismissed on those grounds. The indications are that these

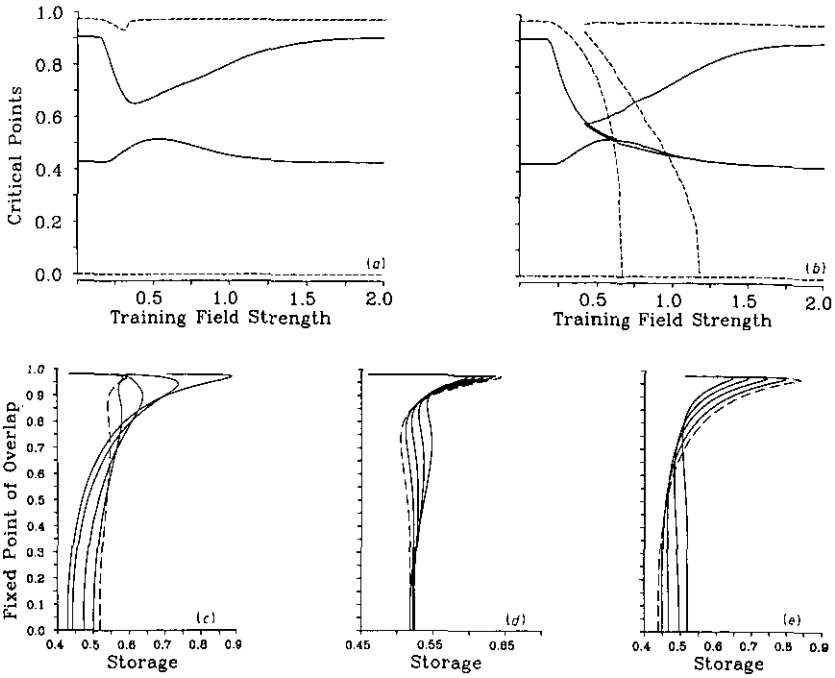

Figure 2. Training field only, for fixed 1% storage error with field noise levels $f_T=0.20$ (*a*) and 0.24 (*b*),(*c*) and (*e*). Parts (*a*) and (*b*) plot the effect of increasing field strength on the critical points' overlaps (– – –) and storage (——). Parts (*a*)–(*e*) are 'snapshots' showing the fixed-point maps which causes the added structure of (*b*).

In (*a*) there are just two critical-points being tracked: the $(\hat{\alpha}_0, \hat{m}_0)$ line with zero overlap and storage starting at 0.42 which marks the upper-bound of wide-retrieval, and the $(\hat{\alpha}_1, \hat{m}_1)$ line starting with overlap $\simeq 0.98$ and storage $\simeq 0.90$ which marks the end of narrow-retrieval. The increase in the wide-retrieval region as indicated by $\hat{\alpha}_0$ shows content addressability is improved, but the decrease in $\hat{\alpha}_1$ shows a worsening of the network's saturation limit, *i.e.*, the narrow retrieval region.

With a higher noise level (*b*) the critical-points plot appears to be far more complicated. However, this additional structure has a straightforward origin and is due entirely to the creation, and later disappearance of, an additional stable fixed point of lower quality than the memory attractor. The evolution of this structure can be explicitly seen by examining the fixed-point maps from which the critical-points are extracted, as shown in (*c*)–(*e*).

(*c*)–(*e*) 'snapshots' showing the fixed-point maps which cause the added structure of (*b*). The evolution is shown by plotting the fixed-point of overlap against storage capacity, for several increasing field strengths with the largest strength plot drawn in broken lines.

In (*c*) the plots are drawn with training field strengths of $\tau_T = 0.20$, 0.28, 0.36, 0.44 and (in broken lines) 0.52. For low field strengths $\tau_T = 0.20$, 0.28 and 0.36 the wide and narrow retrieval regions as marked out by the critical-points $(\hat{\alpha}_0, \hat{m}_0)$ and $(\hat{\alpha}_1, \hat{m}_1)$ increase and decrease respectively, as in (*a*). However as the field strength increases from $\tau_T = 0.36$ to 0.44, a new attractor is created with a stable fixed-point in overlap from $m=0.00$ to $m \simeq 0.90$.

Next in (*d*) the field strength is increased to $\tau_T = 0.52$, 0.56, 0.60, 0.64 and 0.68. Here the original $(\hat{\alpha}_1, \hat{m}_1)$ critical point merges into the $(\hat{\alpha}_0, \hat{m}_0)$ point at the abscissa at $\tau_T = 0.60$–0.64. Its role marking out the region of narrow retrieval having already been usurped by one of the critical-points associated with the new attractor created in (*c*).

Finally in (*e*), as the training field strength is further increased by $\tau_T = 0.68$, 0.84, 1.00, 1.16 and 1.32, the extraneous attractor's stable fixed-point also disappears into the abscissa, merging the final extraneous critical point into the $(\hat{\alpha}_0, \hat{m}_0)$ point.

solutions are unstable at very high storage loads which are never approached in the cases discussed.

4.2. Retrieval field only

As mentioned above, the effect of an external persistent field upon the retrieval dynamics has recently been studied for the zero storage error (Engel *et al* 1990), but it is still useful to consider the 1% error case for the sake of direct comparisons.

For low retrieval noises (figure 3(*a*)) below $f_T \sim 0.20$ the structure is straightforward with no new attractors appearing. However, unlike the training-field only case, introduction of the retrieval field breaks the invariance of the dynamical equation (11) to $m \to -m$ overlap flips, hence raising the critical point's overlap $\hat{m}_0$ above zero. Consequently the stable zero-overlap attractor at $\alpha > \hat{\alpha}_0$ is replaced with a macroscopic one induced by the external field. Meantime the wide-retrieval region $\hat{\alpha}_0$ increases, then eventually coalesces with the falling narrow-retrieval point $\hat{\alpha}_1$. Beyond this the retrieval map has no critical points, and instead there is just one attractor of steadily decreasing quality with increasing storage, signalling the dominance of the external field in the network's dynamics.

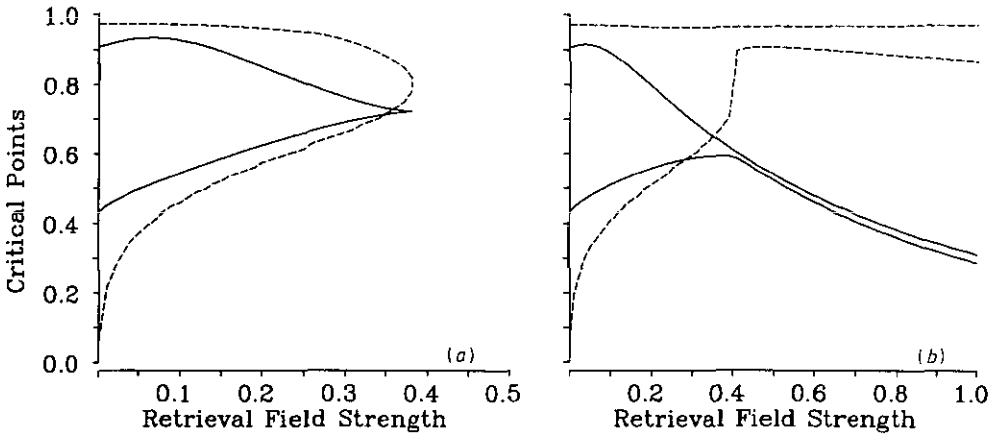

Figure 3. Retrieval field only. Plot of critical-points for 1% storage error with a low noise level $f_R = 0.20$ (*a*), and a higher one $f_R = 0.30$ (*b*). As with figures 2(*a*) and (*b*) these are joint plots of the critical points' overlaps (– – –) and storage (——). The critical points of wide and narrow retrieval eventually come together at $\tau_R = 0.38$ whereupon with the lower noise example (*a*) the network has a single (stable) fixed point of decreasing quality with storage, preventing a simple demarcation of the retrieval regions. However, study of the fixed-point maps indicate that the basin of attraction does not improve with further increases in the retrieval field strength, and this is corroborated by the second plot (*b*) where the critical points do survive.

Retrieving with the slightly noisier field shown by figure 3(*b*) retains the critical points throughout all the field strength range. Consequently it is possible to see how the regions of wide and narrow retrieval decrease for large field strengths. In comparison to the example in figure 3(*a*) this is able to show the eventual polarization of the spin sites to the external retrieval field.

Lastly, both the plots given here are within the regime where replica-symmetry is stable.

4.3. Equal training and retrieval fields

Since the motivation for this work is the training of the network with ensembles of external noisy patterns, it seems reasonable to expect the best performance with (statistically) identical training and retrieval fields. The plots of such an assumption are presented in figures 4(*a*) and 4(*b*), for two levels of external field noise. The most immediately apparent observation is their similarity to the pure retrieval field cases—identical for low field strengths $f_{T/R} <\sim 0.20$. For higher strengths differences do occur, manifesting as higher wide-retrieval regions. These differences are compared in table 1 which shows the maximum wide-retrieval storages $\max\{\hat{\alpha}_0\}$ and the corresponding 'best' field strength, amongst the three cases discussed above for three field noise levels. Nonetheless the improvements are such it may be implied that the retrieval field has a disproportionate effect on the dynamics, and hence an 'optimal' combination of training and retrieval fields may involve weakening the latter. This hypothesis is supported by the disparate best external field strengths to use.

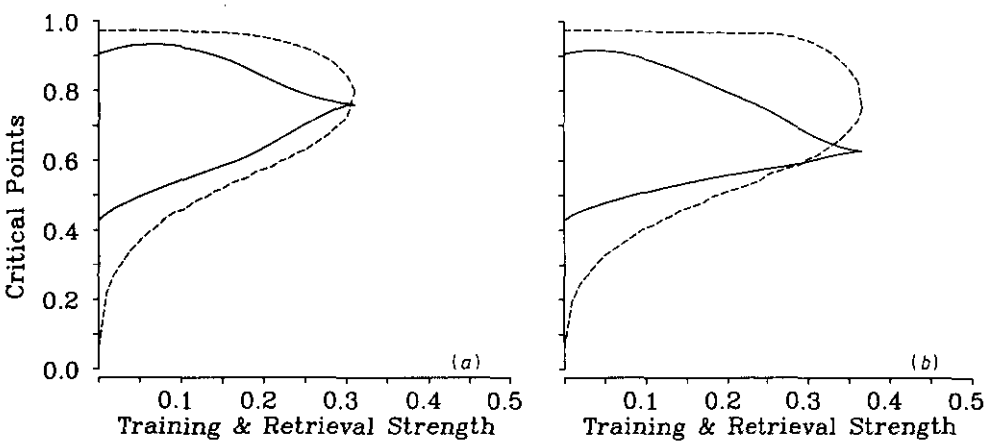

Figure 4. Equal training and retrieval external fields parameters. Plot of critical points for storage error once again at 1%, and training and retrieval noise levels are $f_{T/R}=0.20$ (*a*) and 0.30 (*b*) respectively. The plots' similarity with the pure-retrieval case suggests the dynamics is overly dominated by that field.

Table 1. Comparison of the three training and retrieval cases, for three field noise levels. The entries refer to the largest maximum wide-retrieval storage capacities and the field strengths used to obtain that. For comparison, the region of wide-retrieval in the original Gardner model is bounded by $\hat{\alpha}_0 = 0.42$.

	Mean External Field Noise Level					
	0.20		0.24		0.30	
	$\max\{\hat{\alpha}_0\}$	best$\{\tau\}$	$\max\{\hat{\alpha}_0\}$	best$\{\tau\}$	$\max\{\hat{\alpha}_0\}$	best$\{\tau\}$
Training field only	0.52	0.52	0.52	0.60	0.50	0.50
Retrieval field only	0.72	0.38	0.66	0.38	0.60	0.38
Equal training and retrieval fields	0.76	0.30	0.71	0.32	0.63	0.36

Once again the results presented here lie within the replica-symmetrically stable regime.

5. Concluding remarks

The properties of a neural network model with optimal connections, trained and later retrieved with noisy external fields, is calculated. Novel behaviour from the original Gardner model is found for non-zero storage errors with certain ranges in the training field parameters. This affects the iterative dynamics retrieval equation via three distinct regimes for the stability field distribution, storage capacity and fractional storage error.

Improvements in the basins of attraction are looked for in three cases: training field only, retrieval field only, and statistically equal training and retrieval fields. In all three cases the region of wide retrieval can be improved above the original Gardner model's $\hat{\alpha}_0 = 0.42$, with the equal field case marginally highest; e.g. $\max\{\hat{\alpha}_0\} = 0.76$ for training and retrieval fields at strength 0.30 and noise level 0.20. However, this slight improvement over the corresponding retrieval-field only case (0.72), and the differing value for the best field strength (0.38), perhaps suggests the retrieval field is dominating the dynamics and that a simple equality is not the optimal relationship between the training and retrieval field parameters.

Finally, stability of the replica-symmetric ansatz appears to be respected in all the cases discussed.

Acknowledgments

HWY wishes to thank the groups in Edinburgh and Oxford for many useful discussions, in particular Michael Wong, Martin Evans and David Sherrington. Financial support from the SERC and MEiKO Scientific is gratefully acknowledged.

References

Abbott L F 1990 Learning in neural network memories *Network* **1** 105–22

Amit D J, Evans M R, Horner H and Wong K Y M 1990 Retrieval phase diagrams for attractor neural networks with optimal interactions *J. Phys. A: Math. Gen.* **23** 3361–81

de Almeida J R L and Thouless D J 1978 Stability of the Sherrington-Kirkpatrick solution of a spin glass model *J. Phys. A: Math. Gen.* **11** 983–90

Derrida B, Gardner E J and Zippelius A 1987 *Europhys. Lett.* **4** 167

Engel A, Bouten M, Komoda A and Serneels R 1990 Enlarged basin of attraction in neural networks with persistent stimuli *Phys. Rev.* A **42** 4998–5005

Forrest B M 1988 Content-addressability in neural networks *J. Phys. A: Math. Gen.* **21** 245–56

Gardner E J 1988 The space of interactions in neural network models *J. Phys. A: Math. Gen.* **21** 257–70

—— 1989 Optimal basins of attraction in randomly-sparse neural network models *J. Phys. A: Math. Gen.* **22** 1969–74

Gardner E J and Derrida B 1988 Optimal storage of neural network models *J. Phys. A: Math. Gen.* **21** 270–84

Gardner E J, Stroud N and Wallace D J 1989 Training with noise and the storage of correlated patterns in a neural network model *J. Phys. A: Math. Gen.* **22** 2019–30

Griniasty M and Gutfreund H 1991 Learning and retrieval in attractor neural networks *J. Phys. A: Math. Gen.* **24** 715–34

Hendrich N 1991 Associative memory in damaged neural networks *J. Phys. A: Math. Gen.* **24** 2877–87
Kepler T B and Abbott L F 1988 Domains of attraction in neural networks *J. Physique* **49** 1657–62
Wong K Y M and Sherrington D 1990a Optimally adapted attractor neural networks in the presence of noise *J. Phys. A: Math. Gen.* **23** 4659–72
—— 1990b Training noise adaptation in attractor neural networks *J. Phys. A: Math. Gen.* **23** L175–82