IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Optimally adapted attractor neural networks in the presence of noise

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

1990 J. Phys. A: Math. Gen. 23 4659

(http://iopscience.iop.org/0305-4470/23/20/026)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 129.252.86.83
The article was downloaded on 01/06/2010 at 09:22

Please note that terms and conditions apply.

J. Phys. A: Math. Gen. **23** (1990) 4659-4672. Printed in the UK

Optimally adapted attractor neural networks in the presence of noise

K Y M Wong and D Sherrington

Department of Theoretical Physics, Oxford University, 1 Keble Road, Oxford OX1 3NP, UK† and Department of Physics, Imperial College, London SW7 2BZ, UK

Received 22 June 1990

Abstract. By adapting an attractor neural network to an appropriate training overlap, we optimize its attractor overlap, and subsequently the storage capacity, when retrieval noise (temperature) is present in the system. The training overlap is determined self-consistently by the optimal attractor overlap. The phase diagram of the optimal attractor overlap in the temperature-storage space is found. A novel co-existence phase of strong and weak retrievers is present. The maximum storage capacity deviates from the storage capacity of the maximally stable network on increasing temperature, and in the high-temperature regime ($T \geqslant 0.38$ for Gaussian noise), the Hopfield network yields the maximum storage capacity. Our analysis demonstrates the principles of specialization and adaptation in neural networks.

1. Introduction

It is well known that retrieval in attractor neural networks is very robust against noise disruption [1-3]. In the presence of a moderate amount of noise, which flips the states of the nodes with a probability dependent on the local field and the noise magnitude (or temperature), the network configuration can still drift towards the neighbourhood of a stored pattern. Phase diagrams in the space of temperature (T) and storage level (α) have been obtained for various synaptic prescriptions. However, the optimization of network performance in the presence of noise remains an open question.

Following the work of Gardner and Derrida [4, 5], we have recently studied the effects of introducing noises in the *training* stage, rather than the *retrieval* stage of an attractor neural network [6, 7]. In the presence of *training* noise, the associativity (i.e. the size of the basin of attraction) of the stored patterns increases, although an excessive amount of training noise also reduces the retrieval quality.

In this paper we extend the work of [6] and introduce another use of training noises: by adapting an attractor neural network to an appropriate amount of training noise, we can optimize its performance, and hence the storage capacity, when retrieval noises are present.

Already in [3], there have been indications that training noises are required to improve the retrieval performance in noisy networks. The so-called maximally stable network (referred to as the MSN hereafter) has the maximum storage capacity in the absence of retrieval noise, but at high temperature it performs worse than the Hopfield network. (Indeed, we have shown in [8] that the Hopfield synaptic prescription gives

† Present address.

0305-4470/90/204659+14$03.50 © 1990 IOP Publishing Ltd

the best performance, among *all* Boolean functions, in the high-*training*-noise limit at zero temperature. Hence it is natural to expect that it performs better than other networks in general high-noise situations.) At moderate temperature, 'learning with error' can also improve marginally the performance of the MSN. No doubt, the network with maximal stability minimizes the output error for the case of perfect input of a pattern [6], but the input signals to a node in the presence of retrieval noises can never be perfect, even in the attractor of a stored pattern. Optimizing the performance of the network therefore requires its adaptation to imperfect input signals. The notion of training noise adaptation [6] is therefore relevant.

To determine the amount of training noise to be introduced, we make use of the principle that when a system optimizes its performance in a training environment, then its performance is optimized, among other systems, in the same retrieval environment. This concept is similar to the notion of adaptation in biology. Thus we can optimize the network performance by adjusting the training noise to be at the same level as the error in the retrieval attractor. Since the retrieval error again depends on the training noise, they can only be determined self-consistently.

We emphasize the distinction between learning and adaptation. Ordinary learning involves optimizing the network performance in a fixed *training* environment. On the other hand, adaptation involves optimizing the network performance in a fixed *retrieving* environment. An adapted performance can only be determined self-consistently by the retrieval performance, and an adaptive process involves continually optimizing the network performance in the (adiabatically evolving) environment created by its own retrieval stage, so that the attractor performance of the network is eventually optimized.

Consequently, the retrieval performance and phase diagrams in this paper do *not* correspond to those of networks with *fixed* synaptic prescriptions, say the MSN or the Hopfield network, which were usually studied in previous literature, say [3]. Instead, we are searching the entire space of interactions for the optimal performance. The network for each value of T and α we study corresponds to a unique interaction configuration, or a unique *retriever*. To emphasize this distinction, we call a phase diagram such as figure 5 a *retriever* phase diagram, in contrast to *retrieval* phase diagrams in previous literature.

The principle of adaptation is very general. Apparently, it will have far-reaching implications to the training of neural networks beyond the particular example considered here.

2. Formulation

With this background discussion, let us turn to a more detailed analysis. There are two ways of introducing noises to the retrieval dynamics.

(i) Discrete noise. The output state S_i of a node i at time $t+1$, which takes the possible Ising values ± 1, is updated according to the probability

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

where $T = \beta^{-1}$ is the temperature, quantifying the amount of retrieval noise, $h_i(t)$ is the (normalized) local field at node i, given by

$$h_i(t) = \frac{1}{\sqrt{C}} \sum_{j=i_1}^{i_C} J_{ij} S_j(t) \tag{2}$$

with C being the connectivity of a node, and $j = i_1, \ldots, i_C$ the nodes feeding node i. Here the interactions J_{ij} satisfy the spherical constraint $\Sigma_j J_{ij}^2 = C$, and we shall be interested in the case of large connectivity $C \gg 1$.

(ii) Gaussian noise. The output state S_i of node i at time $t+1$ is stochastically updated according to

$$S_i(t+1) = \mathrm{sgn}(h_i(t) + Tz) \tag{3}$$

where z is a Gaussian variable of mean 0 and width 1.

Below we shall focus on the case of Gaussian noise.

Next, we consider optimizing the averaged output overlap g_i^μ at node i of the stored patterns $\{\xi_i^\mu;\ 1 \leq i \leq N;\ 1 \leq \mu \leq p\}$ for a temperature T and training overlap m_t,

$$g_i^\mu = \langle \langle \xi_i^\mu S_i(t+1) \rangle_{\mathrm{th}} \rangle_{m_t}. \tag{4}$$

Here $\langle \ldots \rangle_{\mathrm{th}}$ denotes thermal averaging at temperature T, and $\langle \ldots \rangle_{m_t}$ represents averaging over input states $\{S_i(t)\}$ having overlap m_t with pattern μ. It turns out that this quantity is a function of Λ_i^μ, which is the local field at node i, in the aligning direction of the output state ξ_i^μ, when the input state is that of the stored pattern μ [9-11], i.e.

$$\Lambda_i^\mu = \frac{\xi_i^\mu}{\sqrt{C}} \sum_{j=1}^{C} J_{ij} \xi_j^\mu \tag{5}$$

(Λ_i^μ will henceforth be called the aligning field, and the subscript i will be implicit hereafter). Since in this case the local field for pattern μ is a Gaussian variable with mean $m_t \Lambda^\mu$ and width $\sqrt{1-m_t^2}$, the performance function to be optimized is then the averaged output overlap given by

$$\begin{aligned} g_{m_t}(\Lambda^\mu) &= \int \mathrm{D}z \int \mathrm{D}y\ \mathrm{sgn}(m_t \Lambda^\mu + \sqrt{1-m_t^2}\, y + Tz) \\ &= \mathrm{erf}\left(\frac{m_t \Lambda^\mu}{\sqrt{2(1-m_t^2+T^2)}} \right) \end{aligned} \tag{6}$$

with $\mathrm{D}z = \exp(-z^2/2)\, \mathrm{d}z/\sqrt{2\pi}$.

Following [6], we shall take this as the appropriate performance function, and maximize it in the space of interactions, quenched-averaged over the stored patterns $\{\xi_i^\mu\}$. This is done by defining an energy function equal to minus the performance function, introducing a free energy corresponding to a thermodynamic average at an annealing temperature T_{an} and then taking the limit $T_{\mathrm{an}} \to 0$ to give the 'ground-state energy'/minimum cost. (Note that the annealing temperature T_{an} is unrelated to the noise temperature T.) Using the replica method [12] for averaging over random patterns, the results of [6] can easily be generalized, in the replica symmetric ansatz, to optimize an *arbitrary* performance function g which is dependent on the variables Λ^μ. This procedure of optimization is derived in appendix 1, and summarized as follows.

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

$$\rho(\Lambda) = \int \mathrm{D}t \delta(\Lambda - \lambda(t)) \tag{7}$$

where $\lambda(t)$ is the inverse function of $t(\lambda)$ defined by

$$t(\lambda) = \lambda - \gamma g'(\lambda) \tag{8}$$

where γ is a constant determined by the condition

$$\int \mathrm{D}t(\lambda(t)-t)^2=\alpha^{-1} \tag{9}$$

with $\alpha=p/C$ being the storage level. When the function $\lambda(t)$ is multi-valued, we choose the argument which gives the largest value of $g(\lambda)-(\lambda-t)^2/2\gamma$. This is equivalent to discarding the range of argument $[\lambda_<, \lambda_>]$ given by Maxwell's construction

$$\int_{\lambda_<}^{\lambda_>} \mathrm{d}\lambda t(\lambda)=t_0(\lambda_>-\lambda_<) \tag{10}$$

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

This completes our formulation of the training stage.

To study the retrieval stage of this optimized network, we consider the output overlap $f_{m_t}(m)$ with a stored pattern for an *arbitrary* input overlap m at the same temperature T. Again, this mapping is determined by the aligning field distribution $\rho_{m_t}(\Lambda)$, which in turn is determined by the training overlap m_t via equations (7)-(10). Following the argument of [9-11], which we have already used in deriving equation (6), namely that the local field for an input overlap m with pattern μ is a Gaussian variable with mean $m\Lambda^\mu$ and width $\sqrt{1-m^2}$, we have

$$f_{m_t}(m)=\int \mathrm{d}\Lambda \rho_{m_t}(\Lambda) g_m(\Lambda) \tag{11}$$

where $g_m(\Lambda)$ is given by equation (6) with m_t replaced by m. Note that the output overlap is a function of both the retrieval overlap m and the training overlap m_t, and when m becomes m_t, equation (11) reduces to the maximum performance function.

Finally, we have to determine the training overlap m_t which gives the optimal performance for a constant temperature T and storage level α. To this end we invoke a principle of adaptation relating the training overlap m_t and the retrieval overlap m. This means that if we consider a fixed retrieval overlap m and search the space of interactions, the network which gives the best output overlap is the one corresponding to the training overlap $m_t=m$. Conversely, if we consider a network with fixed training overlap m_t and search the space of state configurations, the output overlap which is better than those of any other networks is found at the retrieval overlap $m=m_t$. These statements are the direct consequences of the optimization procedure. Thus we can envisage a family of retrieval curves $f_{m_t}(m)$ each being enveloped by the curve $f_m(m)$ above them (see figure 1). Mathematically, this is equivalent to

$$\frac{\mathrm{d}}{\mathrm{d}m_t} f_{m_t}(m)|_{m_t=m}=0. \tag{12}$$

As a check that our optimization procedure sketched from equations (7)-(10) does lead to this conclusion, this equation is explicitly proved in appendix 2.

So far, the analysis is applicable to attractor neural networks of any connectivity, as well as for feedforward networks for one time step. Henceforth, we shall restrict our discussion to dilute attractor neural networks satisfying $C \ll \ln N$, whose retrieval dynamics is completely determined by the retrieval mapping in (11) [2]. The attractor overlap corresponds to the stable fixed points of the retrieval mapping, and the basin boundary of the attractor is determined by its unstable fixed points.

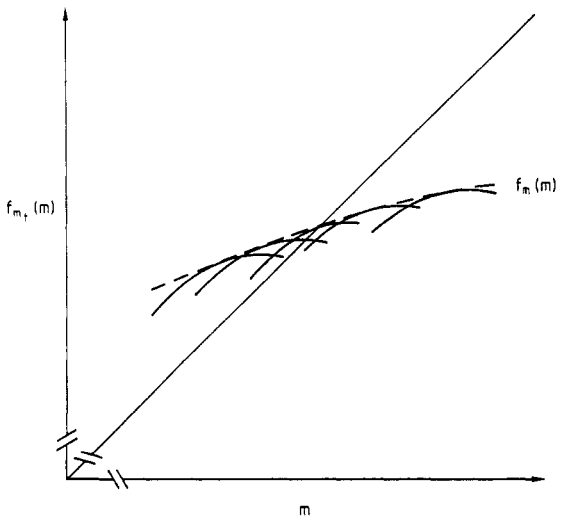

Figure 1. A schematic diagram of the family of retrieval curves $f_{m_t}(m)$ and their envelope $f_m(m)$. The fixed point of the envelope optimizes the network performance.

Now if we consider the training overlap m_t as an adjustable parameter, the principle of adaptation relating the optimal training and retrieval overlaps implies that the stable fixed point of the envelope $f_m(m)$ would give the best attractor overlap. Hence for a constant temperature T and storage level α, we would choose the training overlap m_t to be the fixed point m of the envelope $f_m(m)$:

$$f_m(m) = m. \tag{13}$$

Once we obtain the fixed point of the envelope, the retrieval mapping of the corresponding training overlap will touch the envelope at the same point, and its fixed point will be identical. As schematically shown in figure 1, it will give a greater fixed point overlap than any other network for the particular T and α. Hence the fixed point of the envelope gives, on one hand, the training overlap required to optimize the network performance and, on the other hand, the attractor overlap during retrieval.

3. Results for the optimally adapted retrievers

Thus the best possible attractor overlap is given by the solution of equations (6)-(11), (13) for a constant T and α. In practice, we solve equations (6)-(8), (10), (11), (13) for a constant T and m, and find the corresponding α by equation (9). In general, the retrieval envelope $f_m(m)$ has at most two convex regions in the range $0 \leqslant m \leqslant 1$, and hence there exist at most two non-zero stable fixed points, in contrast to the retrieval curve of the MSN, which has at most one non-zero stable fixed point in the same range. (See figure 2.) As we shall see, this interesting difference gives rise to novel co-existence phases in the phase diagram. The results, on increasing T, are presented below. (See figure 3.)

First consider $T = 0$. For $0 \leqslant \alpha \leqslant 0.60$, the perfect retriever with $m = 1$ is the only stable fixed point, and $m = 0$ is the only unstable fixed point. This means that a network capable of perfect retrieval can be constructed, if we use a training overlap $m_t = 1^-$.

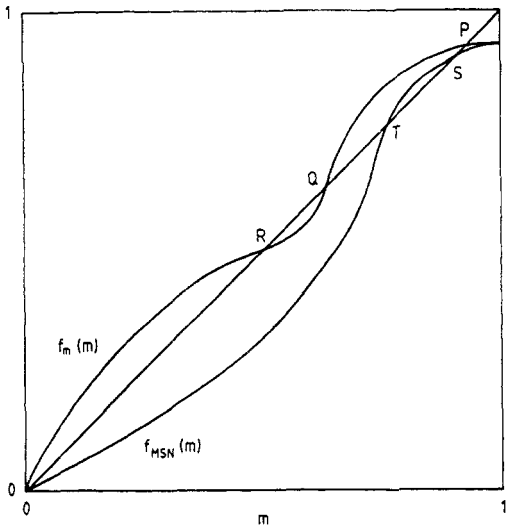

Figure 2. A schematic plot of the retrieval envelope $f_m(m)$ having two non-zero stable fixed points P and R in the range $0 \leqslant m \leqslant 1$. This envelope is typical in region II of the phase diagram in figure 5. (Regions I and III have only one non-zero stable fixed point in the same range.) The retrieval curve $f_{MSN}(m)$ of the MSN, which has only one non-zero stable fixed point S, is also shown for comparison. It touches the envelope at $m = 1$, with its stable fixed point S smaller than the greater one P of the envelope, and unstable fixed point T larger than the corresponding Q of the envelope.

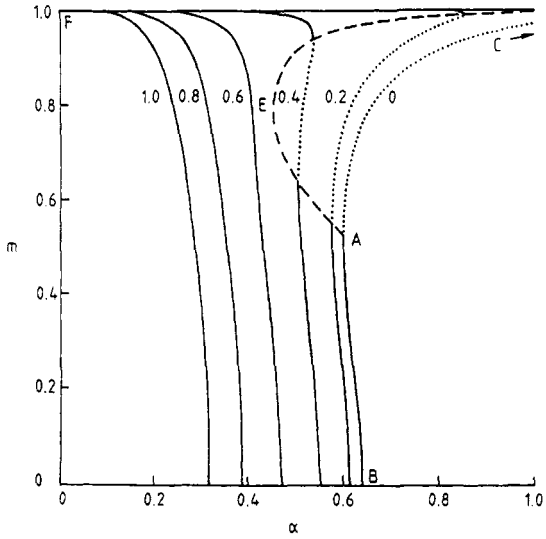

Figure 3. The dependence of the fixed point overlap of the optimally adapted retriever on the storage level for $T = 0, 0.2, 0.4, 0.6, 0.8$ and 1.0. The stable fixed point of $f_m(m)$ is shown in solid curve, and the unstable in dotted curve. The dashed curve corresponds to the discontinuous transition between stable and unstable fixed point overlaps. The alphabetical labels correspond to points in the phase diagram of figure 5.

This retriever is the MSN [6]. As α increases above 0.60, an extra pair of stable and unstable fixed points appears and bifurcates. (Refer to R and Q respectively in figure 2.) Thus besides the perfect retriever, we have another retriever of weaker attractor overlap but, nevertheless, which has a locally maximal performance when compared with other networks in its neighbourhood of the interaction space. As α increases above 0.64, this weak retriever vanishes, leaving behind the perfect retriever and the non-retriever with $m = 0$ as the stable fixed points. At $\alpha = 2$, when the network reaches its storage capacity, the perfect retriever is also destabilized.

The second stable fixed point of the retrieval envelope should not be interpreted as a second weaker retrieval state *at the optimal network configuration* which, if it exists, should correspond to a second stable fixed point of a *single* retrieval mapping. Here we have, instead, a second stable fixed point of the envelope of a *family* of retrieval mappings. Rather, this stable fixed point can be interpreted as an attractor of self-adaptation, in the sense that if we start with a network in its neighbourhood of the interaction space, and allow it to adiabatically adapt its interactions to optimize the performance at the retrieval attractor, then the attractor overlap will converge to this stable fixed point.

Likewise, the unstable fixed point of the retrieval envelope should *not* be interpreted as the basin boundary of retrieval attraction. Instead, it corresponds to the basin boundary of self-adaptation. A further significance of the unstable fixed point will be discussed below.

As T increases from 0 to 0.38, the picture is essentially the same: only a single optimal retriever is present at low storage level; both strong and weak retrievers are present at intermediate storage level; and only the strong retriever survives with a narrowed basin of self-adaptation at high storage level. The attractor overlap of the strong retriever, however, is no longer perfect, but drops with both T and α. At some $\alpha = \alpha_c(T)$, the attractor overlap of the strong retriever and the boundary overlap of self-adaptation coalesce, and the optimal attractor overlap of the network vanishes discontinuously. Since this attractor overlap corresponds to the maximal performance in the interaction space, we argue that $\alpha_c(T)$ is the maximum storage capacity attainable by attractor neural networks. $\alpha_c(T)$ also decreases with T.

As T increases, the strong retriever of high training overlap is less and less favourable when compared with the weak retriever of lower training overlap. For $0.38 \leqslant T \leqslant 0.51$, the strong retriever vanishes before the weak on increasing α, and for $T \geqslant 0.51$, there is at most one retriever for each α. Hence, near $\alpha_c(T)$ the optimal attractor overlap is attained by the weak retriever, and $\alpha_c(T)$ now corresponds to a training overlap which vanishes continuously. Since from (6), $g(\Lambda) \sim \Lambda$ in this limit, we can easily verify that the aligning field distribution $\rho(\Lambda)$ is again a Gaussian distribution of mean $1/\sqrt{\alpha}$ and width 1, corresponding to a dilute Hopfield network with Hebbian synapses, i.e. $J_{ij} \sim \Sigma_\mu \xi_i^\mu \xi_j^\mu$, whose storage capacity is $\alpha_c(T) = 2/[\pi(1+T^2)]$. Thus the Hopfield network gives the maximum storage capacity for attractor neural networks in the high-temperature regime ($T \geqslant 0.38$), where the optimal attractor overlap undergoes a continuous transition.

We have also shown for comparison in figure 4 the attractor overlap and the basin boundary of attraction for the MSN, which corresponds to a vanishing amount of training noise [6]. While the attractor overlaps of the two systems are close in the low-noise regime (i.e. low T or low α), the performances of the MSN (i.e. both the attractor overlap and the storage capacity) are increasingly inferior in the high-noise regime. No weak retrieval states are present in the MSN.

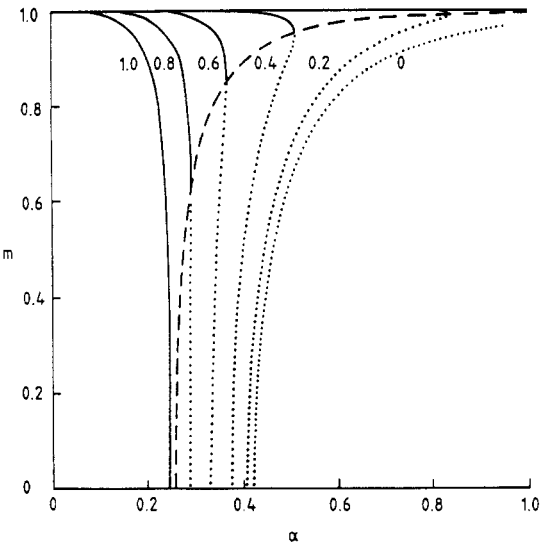

Figure 4. The same curves as in figure 3 for the MSN.

Figure 5 shows the phase diagram in the $T-\alpha$ space. The curve AE corresponds to the bifurcation line on which the weak retriever separates from the strong one, CE corresponds to the transition line on which the strong retriever vanishes discontinuously, and BF that on which the weak retriever vanishes continuously. Hence the regions I to IV are, respectively, the single retriever, strong and weak retrievers, strong retriever, and non-retriever phases. The maximum storage capacity of attractor neural networks is given by the discontinuous transition line CD for $T \leqslant 0.38$, and the continuous transition line DF for $T \geqslant 0.38$. When compared with the storage capacity of the MSN,

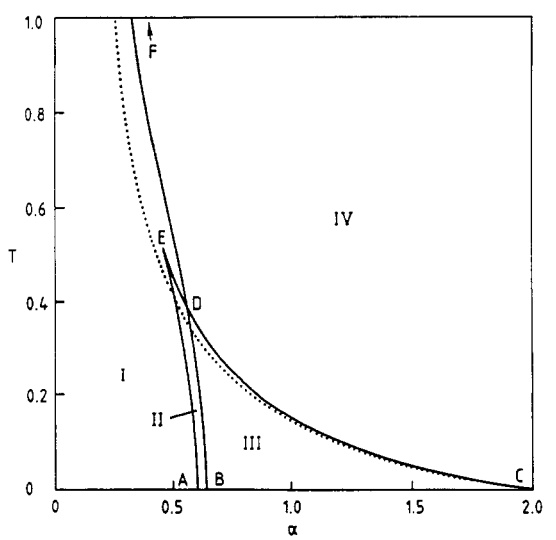

Figure 5. The retriever phase diagram of the optimal network in the temperature-storage space. The retrieval phase boundary of the MSN (dotted curve) is also shown for comparison.

the maximum storage capacity is increasingly superior for increasing T, but the storage capacity is increasingly superior for increasing T, but the storage capacity curves eventually come together asymptotically in the very-high-temperature limit.

Note, however, that this phase diagram should *not* be interpreted as a phase diagram of *attraction*, but one for *self-adaptation*. Phase diagrams of attraction, which usually appeared in previous literature, say [3], describe the attractor behaviour of networks with fixed synaptic prescriptions, say the MSN or the Hopfield network. Here, through the process of self-adaptation, each point in the $T-\alpha$ space has its own optimal value of the training overlap. The phase diagram thereby illustrates the (global or local) maximum attractor overlaps that can be attained for a particular value of T and α if we, instead of considering a fixed synaptic prescription, search in the interaction space for the optimum.

Consequently, we note an important difference between phase diagrams of attraction and of self-adaptation. When we deal with networks of fixed synaptic prescriptions, the aligning field distribution $\rho(\Lambda)$, which determines the retrieval behaviour, is determined by α, and not T. For a given α, the network configuration is independent of T. On the other hand, when we consider self-adapted networks, the aligning field distribution is optimally determined by both α and T. If we consider networks self-adapted at a fixed temperature T_1, the corresponding phase diagram of attraction will have the phase boundaries touching those of self-adaptation at T_1, but not necessarily at other temperatures. (The exception, however, is the case when the maximum storage capacity $\alpha_c(T_1)$ is given by that of the Hopfield network, which is the optimal retriever for all temperatures above 0.38. Hence the phase boundary of attraction coincides with the vanishing line of the weak retriever for $T_1 \geqslant 0.38$. Another case of interest is $0.38 \leqslant T_1 \leqslant 0.51$, when there is a discontinuity in the self-adapted overlap as α is varied. This implies that when we consider the phase diagram for self-adapted networks at fixed T_1, there will also be a corresponding discontinuity in the phase boundaries of attraction, each branch touching one of the phase boundaries of self-adaptation at T_1. These interesting cases are, however, beyond the scope of this work.)

It is interesting to note that the bifurcation line AE and the discontinuous transition line CE meet at the critical point E with a common tangent. Using series expansion around E, we see that the cusp has a critical exponent $\frac{2}{3}$, as shown in appendix 3.

4. Discussions

We have found the optimal attractor overlap as a function of T and α, and the maximum storage capacity as a function of T. A related issue is the optimization of the associativity of the attractor neural network (i.e. maximizing its basins of attraction). Similar arguments lead us to conclude that the optimal basin boundary for a constant T and α is given by the unstable fixed points of the envelope $f_m(m)$, and a training overlap equal to this basin boundary should be introduced to attain this optimum. Since the stable and unstable fixed points coalesce at the phase boundaries, the phase diagram in this case is identical to figure 4, except that the regions I to IV are now, respectively, the wide retriever, wide and narrow retrievers, narrow retriever, and non-retriever phases. This suggests that the maximum storage capacity $\alpha_c(T)$ can be attained by either maximizing the retrieval overlap or the basin of attraction; the two requirements are equivalent at the phase boundary. In regions I and II, the Hebbian synaptic

prescription gives the maximum possible associativity, although it may not give the best retrieval overlap.

Despite the similarity of phase nomenclatures with those in the phase diagram of attraction for the MSN in [3], we again caution that here the phase diagram of self-adaptation bears a very different meaning, for it reflects the best possible associativity in the interaction space for each value of T and α. Again, if we consider networks with maximum associativity at a fixed temperature T_1, the corresponding phase diagram of attraction will have the phase boundaries touching those of self-adaptation at T_1, but not necessarily at other temperatures (except for $T_1 \geqslant 0.38$ when the phase boundaries of attraction and self-adaptation coincide. Also, for $T_1 \leqslant 0.38$, there is a discontinuity in the phase boundaries of attraction.) We further caution that while the phase diagrams for maximum overlap and associativity are identical, they correspond to networks of different training overlaps for a general value of T and α; the training overlaps of the two cases become identical only at the phase boundaries. Furthermore, the networks in the two cases converge to different attractor overlaps.

It should, furthermore, be remarked that when the wide retriever exists, optimal associativity does not correspond to a unique retriever (except at the phase boundary). This is because when the wide retriever exists, the origin becomes an unstable fixed point of the retrieval envelope $f_m(m)$. Since retrieval curves of any training overlap pass through the origin, there exists a range of retrievers having the origin as an unstable fixed point, and each of them has the same optimized associativity, in the sense that each is a wide retriever. In this case, the Hopfield network, which corresponds to a training overlap of $m_t = 0^+$, can still have the best associative power among the wide retrievers, in the sense that its ascendent in overlap with a stored pattern is still greatest near the origin. However, the Hopfield network may or may not give the best retrieval overlap among the possible wide retrievers. As an example, figure 2 of [6] shows that for $\alpha = 0.5$ at $T = 0$, retrievers with m_t between 0 and 0.77 are all wide retrievers, and while the Hopfield network gives the greatest overlap ascendent near the origin, the retriever with $m_t = 0.77$ has the best retrieval overlap.

The optimal performances in the case of discrete noise can be obtained similarly, except that the performance function in (6) has to be replaced by

$$g(\Lambda) = \int \mathrm{D}z \tanh[\beta(m_t\Lambda + \sqrt{1-m_t^2}z)]. \tag{14}$$

The behaviour of the optimal attractor overlap and the corresponding phase diagram are qualitatively the same as for the case of Gaussian noise.

Our work has demonstrated the principle of specialization for different environments in neural networks. At zero temperature the MSN gives the best retrieval overlap and storage capacity, but at high temperatures the MSN is no longer the best. In the high-temperature regime ($T \geqslant 0.38$) the Hopfield network yields the maximum storage capacity, but at lower temperatures its storage capacity is no longer optimal and at zero temperature it is much worse than the MSN. In the low-temperature regime ($0 < T \leqslant 0.38$) neither the MSN nor the Hopfield network has the maximum storage capacity. Instead, the network with the maximum storage capacity at a temperature in this regime corresponds to a training overlap close to (but not exactly) 1^-, which in turn stores less patterns than other networks at other temperatures. Therefore one cannot attain the best storage at all temperature ranges for a single network. In other words, networks are *specialized.* Interestingly, this is roughly the picture anticipated in [3], where it is found that the storage capacity of the MSN is superseded by 'learning

with error' at moderate temperature, and by the Hopfield network at high temperature. The aligning field distribution $\rho(\Lambda)$ for 'learning with error' [3] has a two-band structure: a positive MSN-like band and a small negative tail; this in fact is very similar to the $\rho(\Lambda)$ for training overlaps close to 1 at low temperature [6].

The notion of specialization also applies to different performance requirements in neural networks. In [6] we have already seen that the best retrieval overlap and associativity are given at $T=0$ by the MSN and the Hopfield network, respectively. (See also [8] for similar arguments in the space of neuronal Boolean functions.) For $T \neq 0$, we see that the best retrieval overlap and associativity are generally given by a higher and lower training overlap, which correspond respectively to stable and unstable fixed points of the retrieval envelope. One cannot attain both the best retrieval overlap and associativity for a single network, except at the phase boundaries. Again, networks can be described as specialized. The implication to network design is that separate or modular networks have to be considered in order to achieve both objectives.

The procedure of optimization in equations (7)-(10), and the duality of the optimal training and retrieval overlaps are, in fact, very general. The averaged output overlap, which is used here as the performance function, can be replaced by other performance functions, and the training overlap by other environmental parameters. The important point is the principle of adaptation: if we intend to optimize a performance function g for an environment m during retrieval, the network should be adapted to the same performance function g for the same environment m during training.

Our studies have opened the gateway to examination of the behaviour of networks which optimize an arbitrary performance function for an arbitrary environment, but general algorithms or explicit formulae for the optimal network remain unknown (except, perhaps, a few special cases, e.g., the perceptron learning algorithm [13, 14] or the Hebbian learning rule). This work is now in progress. As a point of further interest, we remark that the concept of self-adaptation, which provides a theoretical interpretation for the retrieval envelope, may also serve as a basis for *practical learning procedures* which optimize the network performance in the presence of noise (or in another retrieval environment). It may be possible that the learning procedure involves relaxing the network to an attractor, and then updating the synaptic interactions to optimize the performance at the retrieval attractor. This idea has been accommodated in some recent work on symmetric networks [15], and surely deserves further exploration.

Acknowledgments

We thank H Horner, M Evans and D Amit for meaningful discussions. We also thank A Zippelius for drawing our attention to [15]. This work is supported by SERC.

Appendix 1.

In this appendix we shall derive the optimization procedure outlined from (7)-(10). Following [6] the partition function corresponding to the performance $\Sigma_\mu g(\Lambda^\mu)$ can be written as

$$Z = \prod_j \int \mathrm{d}J_j \delta\left(\sum_j J_j^2 - C\right) \exp\left(\sum_\mu \beta_{\mathrm{an}} g(\Lambda^\mu)\right) \tag{A1.1}$$

where $\beta_{an} = T_{an}^{-1}$. Using the replica method [12], we shall calculate the pattern-averaged free energy by the replica formula

$$\langle\langle \ln Z \rangle\rangle = \lim_{n\to 0} \frac{1}{n} (\langle\langle Z^n \rangle\rangle - 1) \tag{A1.2}$$

where $\langle\langle \ldots \rangle\rangle$ represents averaging over the patterns. Using the techniques of Gardner and Derrida [5], we can show that $\langle\langle Z^n \rangle\rangle$ is given by

$$\langle\langle Z^n \rangle\rangle = \text{extr} \exp\left[C\left(-\sum_{\alpha<\beta} q_{\alpha\beta} F_{\alpha\beta} + G_J(\{E_\alpha\}, \{F_{\alpha\beta}\}) + \alpha G_\xi(\{q_{\alpha\beta}\}) \right) \right] \tag{A1.3}$$

where the extremum is taken over the space of $\{E_\alpha\}, \{F_{\alpha\beta}\}$ and $\{q_{\alpha\beta}\}$. G_J is a term involving integration over the interaction space, and is identical to the corresponding term in [5], which is given by

$$\exp G_J(\{E_\alpha\}, \{F_{\alpha\beta}\}) = \prod_\alpha \int \mathrm{d}J_\alpha \exp\left(-\sum_\alpha E_\alpha (J_\alpha^2 - 1) + \sum_{\alpha<\beta} F_{\alpha\beta} J_\alpha J_\beta \right) \tag{A1.4}$$

and G_ξ is a term involving averaging over the patterns, given by

$$\exp G_\xi(\{q_\alpha\}) = \prod_\alpha \int \frac{\mathrm{d}\lambda_\alpha \, \mathrm{d}x_\alpha}{2\pi} \exp\left(\sum_\alpha [\beta_{an} g(\lambda_\alpha) + \mathrm{i}\lambda_\alpha x_\alpha - \tfrac{1}{2} x_\alpha^2] - \sum_{\alpha<\beta} q_{\alpha\beta} x_\alpha x_\beta \right). \tag{A1.5}$$

In the replica symmetric ansatz, $E_\alpha = E$, $F_{\alpha\beta} = F$ and $q_{\alpha\beta} = q$. In the $n \to 0$ limit, elimination of E and F at the saddlepoint, as done in [5], yields

$$\frac{1}{C} \langle\langle \ln Z \rangle\rangle = \text{extr}_q \left(\tfrac{1}{2} \ln[2\pi(1-q)] + \frac{1}{2(1-q)} + \alpha \int \mathrm{D}t \ln \int \frac{\mathrm{d}\lambda}{\sqrt{2\pi(1-q)}} \exp\left(\beta_{an} g(\lambda) - \frac{(\lambda - t)^2}{2(1-q)} \right) \right). \tag{A1.6}$$

In the low-temperature limit, $\beta_{an} \to \infty$ and $\beta_{an}(1-q) = \gamma$, the integration over λ can be simplified by the method of steepest descent, so that

$$\int \frac{\mathrm{d}\lambda}{\sqrt{2\pi(1-q)}} \exp\left(\beta_{an} g(\lambda) - \frac{(\lambda - t)^2}{2(1-q)} \right) \to \exp\left[\beta_{an} \left(g(\lambda) - \frac{1}{2\gamma} (\lambda - t)^2 \right) \right] \tag{A1.7}$$

where λ is related to t via (8) at the saddle point. The averaged maximum performance per pattern is then given by

$$f = \lim_{\beta_{an}\to\infty} \frac{1}{\beta_{an} \alpha C} \langle\langle \ln Z \rangle\rangle = \text{extr}_\gamma \left[\int \mathrm{D}t \max_\lambda \left(g(\lambda) - \frac{1}{2\gamma} (\lambda - t)^2 \right) + \frac{1}{2\alpha\gamma} \right]. \tag{A1.8}$$

The extremum condition for γ becomes

$$\frac{1}{2\gamma^2} \left(\int \mathrm{D}t (\lambda(t) - t)^2 - \frac{1}{\alpha} \right) + \int \mathrm{D}t \left(g'(\lambda) - \frac{1}{\gamma} (\lambda - t) \right) \frac{g'(\lambda)}{t'(\lambda)} = 0. \tag{A1.9}$$

By virtue of (8), the second term vanishes, and this condition reduces to (9), and f reduces to $f = \int \mathrm{d}\Lambda \rho_t(\Lambda) g(\Lambda)$, where $\rho_t(\Lambda)$ is identical to $\rho(\Lambda)$ in (7). When the function $\lambda(t)$ is multi-valued, (A1.7) implies that we should choose the value of λ giving the largest exponential argument, which reduces to the Maxwell's construction (10).

It now remains to show that the aligning field distribution

$$\rho(\Lambda) = \lim_{n\to 0} \left\langle\!\!\left\langle \prod_{\alpha j} \int \mathrm{d}J_j^\alpha \prod_\alpha \delta\left(\sum_j (J_j^\alpha)^2 - C \right) \exp\left(\sum_{\alpha\mu} \beta_{an} g(\Lambda_\alpha^\mu) \right) \delta(\Lambda - \Lambda_a^\nu) \right\rangle\!\!\right\rangle \tag{A1.10}$$

is identical to (7). In the $n \to 0$ limit this can be written as

$$\rho(\Lambda) = \prod_{\alpha} \int \frac{d\lambda_\alpha \, dx_\alpha}{2\pi} \times \exp\left(\sum_{\alpha} (\beta_{an} g(\lambda_\alpha) + i\lambda_\alpha x_\alpha - \tfrac{1}{2} x_\alpha^2) - \sum_{\alpha<\beta} q_{\alpha\beta} x_\alpha x_\beta \right) \delta(\Lambda - \lambda_a). \tag{A1.11}$$

In the replica symmetric ansatz and the low-temperature limit, (A1.11) reduces to (7).

Appendix 2.

In this appendix we shall derive equation (12). Using (7) and (11), we obtain

$$\frac{d}{dm_t} f_{m_t}(m) = \int Dt g'_{m_t}(\lambda(t)) \frac{d\lambda(t)}{dm_t}. \tag{A2.1}$$

In the term $d\lambda/dm_t$ above, the differentiation should be carried out at constant t, since it is an integration variable. Thus from (8), λ depends on m_t only via γ and the function $g(\lambda)$. Differentiating (8) with respect to m_t, we obtain

$$\frac{d\lambda(t)}{dm_t} = \lambda'(t) \left(\frac{d\gamma}{dm_t} g'_{m_t}(\lambda(t)) + \gamma \frac{\partial}{\partial m_t} g'_{m_t}(\lambda(t)) \right). \tag{A2.2}$$

To evaluate $d\gamma/dm_t$, we invoke the condition (9), yielding

$$\int Dt(\lambda(t) - t) \frac{d\lambda(t)}{dm_t} = 0. \tag{A2.3}$$

Substituting (A2.2) into (A2.3), we have

$$\frac{d\gamma}{dm_t} = -\gamma \int Dt \lambda'(t)(\lambda(t) - t) \frac{\partial}{\partial m_t} g'_{m_t}(\lambda(t)) \left(\int Dt \lambda'(t)(\lambda(t) - t) g'_{m_t}(\lambda(t)) \right)^{-1}. \tag{A2.4}$$

Substituting (A2.2) and (A2.4) into (A2.1), the result is

$$\frac{d}{dm_t} f_{m_t}(m)\Big|_{m_t = m} = \gamma \left[\int Dt g' \lambda' \frac{\partial g'}{\partial m} - \left(\int Dt g'^2 \lambda' \right) \left(\int Dt \lambda'(\lambda - t) \frac{\partial g'}{\partial m} \right) \times \left(\int Dt \lambda'(\lambda - t) g' \right)^{-1} \right]. \tag{A2.5}$$

Eliminating $\lambda - t$ using (8), we arrive at equation (12). In cases where Maxwell's construction is necessary, surface terms will appear in (A2.1) and (A2.3), but it is easy to prove that they cancel in the final result. Note that the derivation is not dependent on the particular form of the performance function, nor on its functional dependence on the training overlap.

Appendix 3.

In this appendix we shall derive the critical exponent at point E. Since both lines AE and CE involve the coalescence of stable and unstable fixed points, both are described by the equations

$$f_m(m) = m \tag{A3.1}$$

and

$$df_m(m)/dm = 1. \tag{A3.2}$$

The difference between the lines is that $d^2f_m(m)/dm^2 \geqslant 0$ for AE whereas $d^2f_m(m)/dm^2 \leqslant 0$ for CE. The critical point E is therefore given by the conditions (A3.1) and (A3.2), together with

$$d^2f_m(m)/dm^2 = 0. \tag{A3.3}$$

Near E, series expansion of these equations yields

$$\frac{\partial f_0}{\partial T}\Delta T + \frac{\partial f_0}{\partial \alpha}\Delta\alpha + \tfrac{1}{6}f_3(\Delta m)^3 = 0 \tag{A3.4}$$

$$\frac{\partial f_1}{\partial T}\Delta T + \frac{\partial f_1}{\partial \alpha}\Delta\alpha + \tfrac{1}{2}f_3(\Delta m)^2 = 0 \tag{A3.5}$$

where $f_r = d^r f_m(m)/dm^r$, so that along the lines AE and CE we have

$$\Delta m = \pm\left[-\frac{2}{f_3}\left(\frac{\partial f_1}{\partial T}\Delta T + \frac{\partial f_1}{\partial \alpha}\Delta\alpha\right)\right]^{1/2} \tag{A3.6}$$

and the equation of the lines becomes

$$\frac{\partial f_0}{\partial T}\Delta T + \frac{\partial f_0}{\partial \alpha}\Delta\alpha \pm \frac{f_3}{6}\left[-\frac{2}{f_3}\left(\frac{\partial f_1}{\partial T}\Delta T + \frac{\partial f_1}{\partial \alpha}\Delta\alpha\right)\right]^{3/2} = 0. \tag{A3.7}$$

Hence the two lines intersect with a common tangent of slope

$$-(\partial f_0/\partial\alpha)/(\partial f_0/\partial T)$$

and a cusp of critical exponent $\frac{2}{3}$.

References

[1] Amit D, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30
[2] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
[3] Amit D, Evans M, Horner H and Wong K Y M 1990 *J. Phys. A: Math. Gen.* **23** 3361
[4] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[5] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[6] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** L175
[7] Gardner E, Stroud N and Wallace D 1989 *J. Phys. A: Math. Gen.* **22** 2019
[8] Wong K Y M and Sherrington D 1989 *Europhys. Lett.* **10** 419
[9] Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
[10] Krauth W, Nadal J-P and Mézard M 1988 *J. Phys. A: Math. Gen.* **21** 2995
[11] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969
[12] Edwards S F and Anderson P W 1975 *J. Phys. F: Met. Phys.* **5** 965
[13] Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan)
[14] Minsky M and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT Press)
[15] Pöppel G and Krey U 1987 *Europhys. Lett.* **4** 979