Noise-optimal binary-synapse neural networks

1993 J. Phys. A: Math. Gen. 26 3995

(http://iopscience.iop.org/0305-4470/26/16/016)

J. Phys. A: Math. Gen. 26 (1993) 3995–4010. Printed in the UK

Noise-optimal binary-synapse neural networks

R W Penney and D Sherrington
Oxford University, Department of Physics, Theoretical Physics, 1 Keble Road, Oxford OX1 3NP, UK

Received 19 February 1993

Abstract. We examine the possibility of improving the performance of discrete-synapse neural networks, functioning as content-addressable memories, by the inclusion of noise in their training procedure, and study the effects on the training itself. Pattern stability field distributions for optimized networks are illustrated for various levels of training noise, including the noiseless, maximally stable, regime. We show that the clipped Hebb rule is optimal in the high training noise limit, but that simulated annealing cannot be relied upon to identify a well defined optimal network for an arbitrary, finite, training-noise, in contrast to the case for continuous-synapse systems. Training by use of a continuous-synapse network, whose synapses are subsequently clipped, is also addressed.

1. Introduction

Instances of improvements in the performance of neural networks due to controlled introduction of noise into their training procedure are becoming more widespread (e.g. Wong and Sherrington (1991), Holmstrom and Koistinen (1992), Murray (1991) and Györgyi (1990)) with particularly beneficial effects being expected when training environment parallels operational environment (Wong and Sherrington 1990b). One motivation for this strategy is that a network is likely to be better trained by the form of data with which it will ultimately deal, albeit possibly imperfect, than with perfect examples of the the desired behaviour. Stochasticity in the training data presented to a network of limited adaptability also curtails any likelihood that the system will be able to provide an exact representation of these untrustworthy examples, but the network is hoped to abstract from its training a *modus operandi* which will allow it to perform well on new, noisy, input data.

Considering the prototypical neural network function of content addressable memory, both memory associativity and retrieval accuracy of optimally adapted networks surpass those of fixed learning rules which may be best only at particular noise levels (e.g. the Hebb rule is found to be optimal in the high noise limit) (Wong and Sherrington 1991). However, most studies of the training-with-noise procedure have been confined to synaptic networks with real-valued synapses (and therefore, in general, connected weight spaces) for which the influence of the training noise may act smoothly in any form of annealing within this weight space associated with training. For networks having discrete-valued synapses, and therefore having highly disconnected weight spaces, the training noise can less readily usher a network from an unfavourable region of this space towards a superior region, and might only be able to communicate between these domains via vastly inferior states. In the language of optimization theory, a training algorithm is much more likely to become trapped in local†

† Our use of notions of proximity are intended to relate to points between which an iterative learning algorithm may jump in a small number of steps, or that are close in terms of some Hamming distance.

0305-4470/93/163995+16$07.50

minima of any cost function imposed on the weight space. It is therefore relevant to examine in what way the use of discrete-synapse neural networks, which are far more amenable to realization than systems having real-valued weights, tarnishes the benefits of a noisy training environment. We will therefore examine a binary-synapse network functioning as an associative memory, this being the system having the least synaptic flexibility, and therefore expected to highlight the symptoms of discrete-synapse networks in general.

For the synaptic networks of concern here, we will assume a McCulloch–Pitts dynamics

$$S_i(t+1) = \operatorname{sgn}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N} J_{ij}S_j(t)\right) \qquad \text{with } S_k(t) \in \{\pm 1\} \tag{1.1}$$

with the number of neurons, N, taken as being large. As discussed by Wong and Sherrington (1990a), if an ensemble of noise-corrupted versions of a pattern, ξ_j^μ, (one of a set of αN patterns, $\mu \in \{1, \ldots, \alpha N\}$) is presented to the inputs of such a network, then the probability that a neuron, i, immediately connected to these inputs, acquires the incorrect state after updating is given by $\frac{1}{2}\operatorname{erfc}\left([m_t/\sqrt{2(1-m_t^2)}]\Lambda_i^\mu\right)$, in which $\Lambda_i^\mu = (1/\sqrt{N})\sum_j \xi_i^\mu J_{ij}\xi_j^\mu$ is the stability field of pattern μ and m_t is the training overlap, such that $p(S_j(0)) = \frac{1}{2}(1 + m_t S_j(0)\xi_j^\mu)$. Hence the larger the aligning field, Λ_i^μ, the smaller the chance of incorrect updating, and thus the smaller the population of incorrectly updated neurons after the first-step dynamics, when an input is presented having overlap m_t with one of the stored patterns, ξ_j^μ. Although the ideal network would rather maximize the probability of correct asymptotic ($t \to \infty$) retrieval, such a system is in general currently beyond analysis. We will therefore focus on the optimization of the first-step dynamics, whose effects are likely to be highly significant in determining long-term behaviour (cf Kepler and Abbott 1988).

Our general approach to this problem will be discussed in section 2; in section 3 the possibility of obtaining a well defined optimal network is addressed, and section 4 examines the high-training noise limit. In section 5 we consider training a binary network by use of a continuous-synapse system. Our conclusions are offered in section 6.

2. General formalism

We will treat learning as a stochastic minimization process, in which, for a given species of network, the whole of its synaptic weight space is explored, and properties typical of those networks lying in the most favourable regions of this space, according to some imposed criteria, are investigated. For the large networks with which we will be concerned, these average properties are not expected to depend on the exact choice of patterns stored by the network, only on their stochastic properties and number. The distribution of pattern stabilities, $\rho(\Lambda)$, will be the object of central concern as this concisely provides insight into the effects of training on the capabilities of a network. Although the allowed values of Λ_i^μ are strictly discrete, for large N the distribution $\rho(\Lambda)$ becomes quasi-continuous, and is equivalent to a function on real Λ as $N \to \infty$.

Following Gardner and Derrida (1988) we will associate a cost function, $E = \sum_{\mu=1}^{\alpha N} g(\Lambda_i^\mu)$, with each point in the weight space, and use a Gibbs weighting ($e^{-\beta E}$) via which networks may be annealed into the minimum cost regions of the weight space on taking the limit $\beta \to \infty$. Using replica mean field theory, and adopting the replica-symmetric ansatz, the field distribution may be obtained in the form

$$\rho(\Lambda) = \int \mathrm{D}x \, \frac{\int \mathrm{D}y \, \exp\left(-\beta g(y\sqrt{1-q} - x\sqrt{q})\right)\delta(\Lambda - y\sqrt{1-q} + x\sqrt{q})}{\int \mathrm{D}y' \, \exp\left(-\beta g(y'\sqrt{1-q} - x\sqrt{q})\right)} \tag{2.1}$$

where the value of q is chosen, along with the parameter $\hat{q}$, such as to extremize a free energy functional $G(q, \hat{q})$, as given below;

$$G(q,\hat{q}) = \tfrac{1}{2}\hat{q}(q-1) + \alpha \int \mathrm{D}x \, \ln\left\{\int \mathrm{D}y \, \exp\left(-\beta g(y\sqrt{1-q} - x\sqrt{q})\right)\right\} + \int \mathrm{D}x \, \ln\left\{2\cosh x\sqrt{\hat{q}}\right\}. \tag{2.2}$$

(We adopt the standard shorthand $\mathrm{D}x = \exp(-\frac{1}{2}x^2)\,\mathrm{d}x/\sqrt{2\pi}$.) A brief outline of the derivation of (2.1) can be found in appendix A. We will also have use of the replica-symmetric expression for the thermodynamic entropy ($S = \beta^2 \frac{\mathrm{d}}{\mathrm{d}\beta}(-\beta^{-1}G_{\mathrm{extr}})$)

$$S_{\mathrm{RS}} = G_{\mathrm{extr}} + \alpha \int \mathrm{D}x \, \frac{\int \mathrm{D}y \, \exp\left(-\beta g(y\sqrt{1-q} - x\sqrt{q})\right) \beta g(y\sqrt{1-q} - x\sqrt{q})}{\int \mathrm{D}y' \, \exp\left(-\beta g(y'\sqrt{1-q} - x\sqrt{q})\right)}. \tag{2.3}$$

As our aim is to maximize the probability of correct first-step update of a noisy input, the artifice β should ultimately be removed by taking the limit $\beta \to \infty$. In this limit, the effect of using a cost function $g(\Lambda) = \frac{1}{2}\mathrm{erfc}(\beta_{\mathrm{tr}}\Lambda)$ (i.e. the probability of incorrect update) is exactly equivalent to that of $g(\Lambda) = -\mathrm{erf}(\beta_{\mathrm{tr}}\Lambda)$ (in which we define $\beta_{\mathrm{tr}} = m_t/\sqrt{2(1-m_t^2)}$). The latter choice of cost function is slightly more convenient analytically, so is the form that has actually been employed. At finite β the two cost functions are also essentially equivalent in their effects. For definiteness we give the definitions of both error functions

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \mathrm{e}^{-y^2}\,\mathrm{d}y \qquad \mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty \mathrm{e}^{-y^2}\,\mathrm{d}y = 1 - \mathrm{erf}(x).$$

The maximally stable network (MSN), in which all patterns are stable states of the neuron dynamics, and for which $\Lambda_i^\mu > \kappa > 0$, may be considered by taking $g(\Lambda) = \theta(\kappa - \Lambda)$, and maximising κ for the chosen network loading, α. This represents a noise-free training procedure, for which $m_t = 1$.

So far our approach is very similar to that of Wong and Sherrington (1990b). However, the desire to take the annealing temperature, $1/\beta$, to zero, thereby eliminating all but the lowest cost regions of the weight space from the Gibbs averages, highlights the pathologies of discrete synapse networks mentioned in the introduction. For any thermodynamic system having a discrete phase space the entropy cannot be negative, so on increasing β a change of sign of S_{RS} (2.3) would signal the breakdown of the replica-symmetric ansatz (a feature familiar from models of Ising spin glasses, cf Sherrington and Kirkpatrick 1975). This effect is believed to relate to the energy barriers, which separate near-degenerate regions of phase space, becoming infinite in the thermodynamic limit ($N \to \infty$), thereby causing the breakup of phase space into disjoint ergodic components. For the neural networks considered here, this effect would mean that although a training noise might favour one region of weight space over another, any learning algorithm is likely to take infinite time in order to escape from a less than optimal region.

However, the studies of the maximally stable network with binary weights by Krauth and Mézard (1989) suggest that the form of replica symmetry breaking (RSB) exhibited by binary-synapse networks is of a rather different type from the hierarchical scheme seen in spin-glasses (Parisi 1980). The one-step breaking seen in the MSN has the formal effect of clamping the effective temperature of all thermodynamic quantities at that at which

the replica-symmetric entropy reaches zero, independent of the true annealing temperature. Thus, if the replica-symmetric entropy remains non-negative as the annealing temperature is taken to zero, then replica-symmetry will remain intact for all physical temperatures. Whether or not this condition is fulfilled will depend on the loading of the network, α, and one may thereby determine the point at which the network becomes saturated, being unable to sustain greater loading. For the MSN, the value of the replica-symmetric order parameter q (which reflects the mutual similarity of the low-cost networks) is found not to approach unity as the network becomes saturated (a condition signalled by the replica symmetric entropy reaching 0^+ at zero annealing temperature), in contrast to the behaviour of continuous synapse models (e.g. Gardner and Derrida 1988). If, at zero annealing temperature, α is made to approach its saturation limit, α_c, from below, then until α_c is reached S_{RS} will be positive, finite, and of order N^0. Given that q does not approach unity under such circumstances, the number of low-cost networks must be exponentially large (so that $S_{RS} > 0$) with these systems being widely dispersed within the weight space (because $q \not\to 1$). That these networks differ in a very significant fraction of their weights, makes it implausible that the minima of the cost function, which they represent, could lie in a single valley of the energy landscape. This picture is lent weight by the form of replica-symmetry breaking observed for this system. Just beyond saturation replica-symmetry breaks, indicating that these disparate regions of weight space, each then having non-zero energy, become separated by infinite energy barriers.† Although at low annealing temperature a single domain would be favoured in terms of lowest cost, the loss of ergodicity means that it is almost certainly dynamically inaccessible. This type of pathology would encourage caution in taking the limit $\beta \to \infty$ in (2.2) and (2.1).

The techniques used by Wong and Sherrington for the spherical model (in which the synaptic weights are constrained only by an overall normalization condition, $\sum_j J_{ij}^2 = N$) centred on applying the method of steepest descents to integrals involving $e^{\beta g}$. This strategy would seem ill-advised, *a priori*, for the present problem, because it relies on being able to take the limit $\beta \to \infty$ within the replica-symmetric approximation. We have therefore adopted a more cautious, numerical, approach. For particular α and m_t we have numerically extremized $G(q, \hat{q})$ (2.2), and varied β in search of the zero of S_{RS} (2.3). (We have used an adaptive integration routine, based on a step-size controller for a Runge–Kutta differential equation integrator (Press *et al* 1988), in order to handle the highly inhomogeneous integrands in (2.1), (2.2) and (2.3).) According to the results of Krauth and Mézard (1989), these β's would represent a maximum usable β in any simulated annealing method of training the network, and would correspond to a minimum possible step-size in an iterative learning scheme. Given the conventional interpretation of RSB in terms of (exponentially) diverging dynamical timescales, despite the fact that the Gibbsian method strictly makes no direct reference to an underlying dynamics, beyond these limits on β no improvement in network performance should be expected, for practical purposes, even though an optimal network would remain unidentified. These bounds on β are plotted against training overlap, m_t, in figure 1, for various α. The numerical difficulties of calculating these values limits us to a maximum β of about 256, and to a small number of data points (shown as circles). The curves are rational polynomial fittings to calculated values of $\ln(\beta_{max})$.

† The replica-symmetry broken order parameters, q_0, q_1 and m (see appendix A), suggest a number of features of the phase space. The disjoint regions of this space represent the pure thermodynamic states constituent to the full Gibbs state, with state ψ having a weight P_ψ. We infer that these regions are each small (because $q_1 = 1$), widely separated (given that $q_0 \sim q \sim 0.6$) and numerous (because $m = (1 - \sum_\psi P_\psi^2)$ is unity just beyond saturation), but not having a population which diverges exponentially in the system size, N (so that the entropy remains zero, to zeroth order in N).

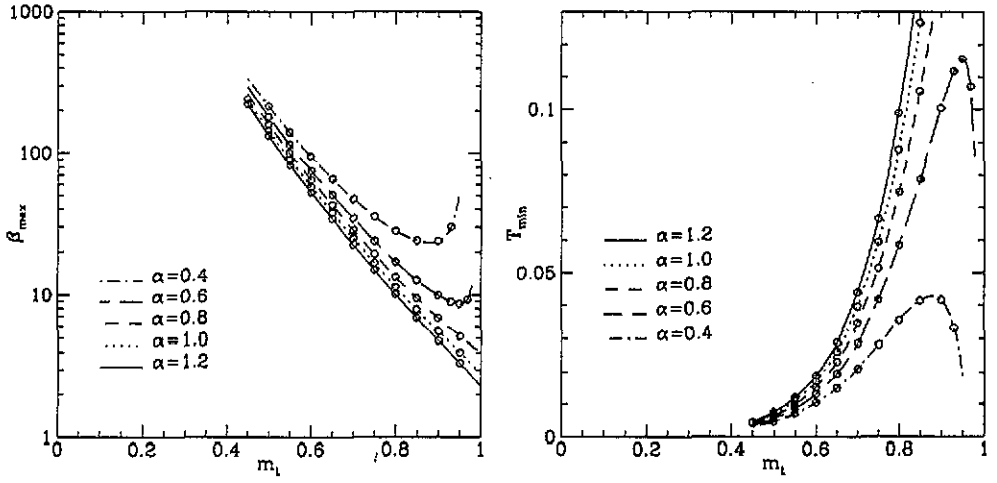

Figure 1. Maximum inverse annealing temperatures, β, and minimum annealing temperatures, $T = \beta^{-1}$, against training overlap, m_t, for various storage ratios, α.

For the MSN, Horner's explicit analysis of the dynamics of learning (Horner 1992) shows that ergodicity breaking can occur even before the replica-symmetric entropy reaches zero. That the curves of figure 1, when extrapolated to $m_t = 1$ (where the error-function cost-function resembles that of the MSN), appear to have finite β_{max} rather than the infinite β allowed for an MSN in the Gibbsian approach for $\alpha < 0.83$ (Krauth and Mézard 1989), would suggest that figure 1 might be at least indicative of the onset of dynamical ergodicity breaking. If real dynamical timescales do not diverge at β_{max} then β_{max} must surely represent an upper bound for the onset of practical difficulties.

Having obtained these bounds on β, one may proceed to examine the distribution of pattern stabilities produced by the network. Some illustrative curves of $\rho(\Lambda)$ are shown in figure 2 and figure 3. In calculating these, where possible the appropriate β_{max} has been

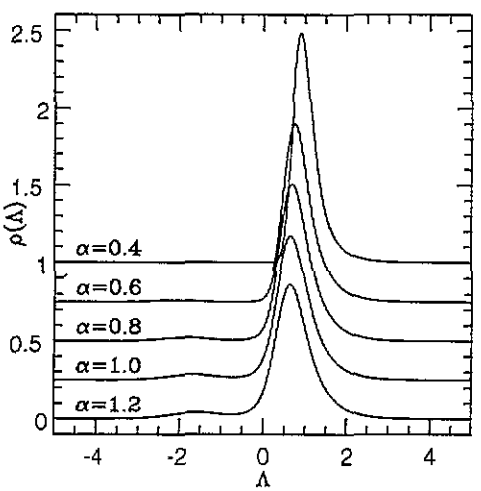

Figure 2. Stability field distributions for various α, all at $m_t = 0.9$. Successive curves are vertically offset by 0.25.

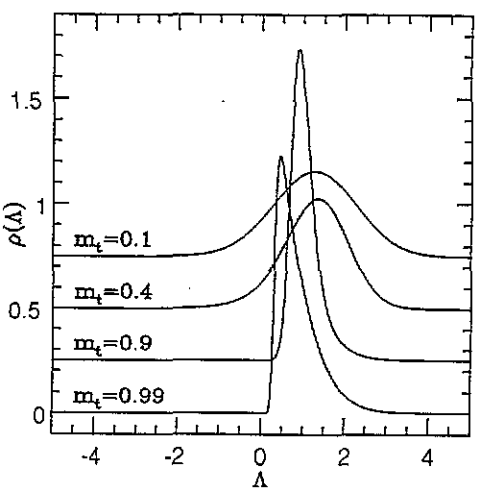

Figure 3. Stability field distributions for various m_t, all at $\alpha = 0.4$. Successive curves are vertically offset by 0.25.

employed, although beyond $\beta_{max} = 256$ we use $\beta = 256$ owing to numerical difficulties. Where forced to limit β in this way, investigation of a number of choices of smaller β for a given α and m_t suggests that the distributions obtained would not change significantly if β was increased.

The field distributions in figure 2 exhibit the same sacrificial storage effects as are seen in the spherical model (Wong and Sherrington 1991), whereby the majority of patterns are given positive stability at the expense of a small fraction of patterns being unstable, i.e. having $\Lambda^\mu < 0$. However, the restriction to finite annealing temperature means that $\rho(\Lambda)$ is always continuous, and no disjoining of the distributions for stabilized and sacrificed patterns occurs. Reducing the network loading facilitates stabilization of all patterns, as reflected in figure 2.

On approaching $m_t = 1$, one would expect $\rho(\Lambda)$ to approach that of the maximally stable network. Comparison of the $m_t = 0.99$ curve in figure 3 with those typical of a binary MSN (shown in figure 4) reveals qualitative similarity, but $\beta_{tr} \sim 5$ is far from infinite, so close quantitative agreement should not be expected. (We note in passing that the aligning-field distributions for the spherical MSN (Kepler and Abbot 1988, Gardner 1989) are markedly dissimilar to those of the binary network, consisting of a δ-function at $\Lambda = \kappa$, and a purely Gaussian tail beyond.) On decreasing the training overlap the cliff in $\rho(\Lambda)$ shallows, and the distribution becomes more rounded, appearing Gaussian in form towards $m_t = 0.1$.

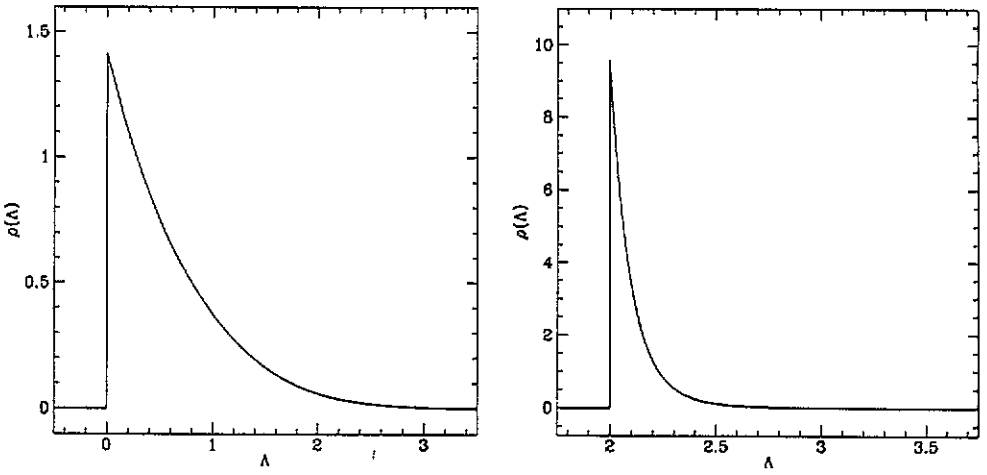

Figure 4. Pattern stability field distributions for maximally-stable (noiseless) training, for $\kappa = 0$ and $\kappa = 2$.

In view of the observation made in Wong and Sherrington (1990a) that the training cost function in the limit $m_t \to 0$ reproduces the Hebb rule, with its associated Gaussian distribution of pattern stabilities, one may wonder what this limit corresponds to for the binary-synapse network, for which a true Hebb rule is obviously inadmissible. The natural suggestion would be that this limit would reproduce the clipped Hebb rule (van Hemmen 1987), for which $J_{ij} = \mathrm{sgn}\big((1/\sqrt{\alpha N})\sum_\mu \xi_i^\mu \xi_j^\mu\big)$. This training rule again produces a Gaussian distribution of pattern stabilities, according to

$$\rho_{cH}(\Lambda) = \frac{1}{\sqrt{2\pi}} \exp -\tfrac{1}{2}\left(\Lambda - \sqrt{\frac{2}{\pi\alpha}}\right)^2. \tag{2.4}$$

Graphs of $\rho(\Lambda)$ for $m_t = 0.1$ do typically appear Gaussian in character (cf figure 3), and in order to examine how closely these might correspond to (2.4) we have calculated cumulant averages over the experimental distributions. For a Gaussian distribution, only the two lowest order cumulants are non-trivial, in general. Comparisons of these cumulants, for various α, are given in table 1, and the agreement is seen to be encouraging. It can be shown to become blatantly less so as m_t is increased.

If one neglects the condition that S_{RS} (2.3) should remain positive if $G(q, \hat{q})$ (2.2) is to be valid, then extremising $G(q, \hat{q})$ for increasing β suggests that the order parameter q would approach unity in the limit $\beta \to \infty$, for all α and m_t. This observation facilitates two analyses presented below. In view of the numerical inaccessibility of some regions of figure 1, and the associated possibility that there might actually be choices of α and m_t which allow zero annealing temperature to be reached, we have tried to analyse, more directly, the replica-symmetric entropy (2.3) in the limit $\beta \to \infty$. If S_{RS} should remain positive in this limit, our method would be self-consistent, and indicate the accessibility of a well defined optimal network. A negative limit would vitiate our approach, and would suggest that zero annealing temperature is inaccessible under the relevant conditions. (The examination of the local stability of the replica-symmetric saddle point, cf de Almeida and Thouless 1978, is believed to be a less reliable indicator, for binary-synapse networks, of the onset of the replica-symmetry breaking that would invalidate our methods, cf Krauth and Mézard 1989).

Table 1. Cumulant averages of Λ, over $\rho(\Lambda)$, for various loadings and for small m_t, along with these quantities for the clipped Hebb rule (cH). All results are based on calculations at $\beta = 256$.

	$\alpha = 0.2$		$\alpha = 0.4$		$\alpha = 0.6$	
	$m_t = 0.1$	cH	$m_t = 0.1$	cH	$m_t = 0.1$	cH
C_1	1.7749	1.7841	1.2583	1.2616	1.0283	1.0301
C_2	0.9429	1	0.9703	1	0.9799	1
C_3	−0.0410	0	−0.0324	0	−0.0276	0
C_4	0.0064	0	0.0042	0	0.0031	0
C_5	−0.0039	0	−0.0002	0	0.0004	0
C_6	−0.0116	0	−0.0033	0	−0.0018	0
	$\alpha = 0.8$		$\alpha = 1.0$			
	$m_t = 0.1$	cH	$m_t = 0.1$	cH		
C_1	0.8909	0.8921	0.7971	0.7979		
C_1	0.9848	1	0.9878	1		
C_1	−0.0245	0	−0.0222	0		
C_1	0.0024	0	0.0020	0		
C_1	0.0006	0	0.0006	0		
C_1	−0.0012	0	−0.0010	0		

3. The accessibility of zero annealing temperature

For the noise-optimal spherical model (Wong and Sherrington 1990b), in the limit $\beta \to \infty$ the quantity $\beta(1 - q)$ is observed to remain finite, indicating that the angular diameter ($\sim \cos^{-1} q$) of the lowest cost region of weight space decays as a power law in the annealing temperature. It would seem reasonable to expect a similar behaviour to be

seen in the binary model, if the annealing temperature can be reduced to zero without ergodicity breaking. We therefore adopt, as an ansatz, $\lim_{\beta\to\infty}\beta(1-q)=\gamma$. Using this one may apply the method of steepest descents to the $e^{-\beta g}$ integrals in (2.2) etc, having scaled y by $\sqrt{\beta}$, following Wong and Sherrington (1990b). However, it is found necessary to know slightly more about the asymptotic behaviour if the entropy (2.3) is to be calculated. By including the first corrections in β^{-1} to γ, $\hat{q}$ and the integrals in (2.2) as $\beta\to\infty$, one may generalize the methods of Wong and Sherrington to allow calculation of the replica-symmetric entropy, to zeroth order in β^{-1}. Making use of the expansion

$$\int \mathrm{D}x\,\frac{\ln(2\cosh ax)}{a} \sim \sqrt{\frac{2}{\pi}}+\frac{1}{\sqrt{2\pi}}\left\{\frac{1}{12}\left(\frac{\pi}{a}\right)^2\cdots+(-1)^{n-1}\frac{(2n-3)!!\,(2^{2n-1}-1)B_n}{4^{n-1}.(2n)!}\left(\frac{\pi}{a}\right)^{2n}\cdots\right\} \tag{3.1}$$

one may determine the asymptotic character of the entropy to be

$$S_{\mathrm{RS}}\sim\frac{\alpha}{2}\int \mathrm{D}x\,\left\{xy^*\sqrt{\gamma}-\ln f''(y^*)\right\} \tag{3.2}$$

in which

$$f(y)=\tfrac{1}{2}y^2-g(y\sqrt{\gamma}-x) \qquad \text{and, for each } x,\quad y^*=\{y: f(y)=\inf f\}\,.$$

Having eliminated $\hat{q}$ from (2.2), using (3.1), the condition determining q, and hence γ, now becomes

$$\alpha\int \mathrm{D}x\,y^{*2}=\frac{2}{\pi\gamma}\,. \tag{3.3}$$

We have examined (3.2) both analytically (towards small, fixed, m_t) and numerically, in search of choices of γ and m_t which make $S_{\mathrm{RS}}(\beta\to\infty)$ positive, but have found only negative values. Iterative root-searching using Newton's method also failed to converge. The general trend of the $\beta_{\max}$ curves in figure 1 would seem consistent with this lack of success.

Rather than retaining a fixed training overlap as the annealing temperature (β^{-1}) is reduced, the following section shows that if $m_t\to 0$ as $\beta\to\infty$, zero annealing temperature can be attained without ergodicity breaking.

4. The clipped Hebb rule limit

We have tried to show analytically that the limit $m_t\to 0$ reproduces the distribution of pattern stabilities associated with the clipped Hebb rule. Assuming that m_t is sufficiently small, one may approximate the cost function $g(\Lambda)=-\mathrm{erf}(\beta_{\mathrm{tr}}\Lambda)$ by the first term in its Maclaurin expansion, $g(\Lambda)\simeq-(2/\sqrt{\pi})\beta_{\mathrm{tr}}\Lambda$. In order that higher order terms are negligible, self-consistency of this approximation will require $\beta\beta_{\mathrm{tr}}^3\ll 1$, i.e. that the training overlap decreases with annealing temperature, together with $\beta\beta_{\mathrm{tr}}\gg 1$. (The less restrictive conditions on β_{tr} considered in the previous section, are seen to lead to ergodicity-breaking.)

In the low temperature limit, we again expect the following general trends; $q \to 1$, $\beta(1-q) \to \gamma$, $\hat{q} \to \infty$. Once more performing asymptotic analysis on (2.2), making use of the (3.1), one may arrive at the following properties:

$$\beta(1-q) \sim \frac{1}{\sqrt{\alpha}\beta_{\rm tr}} \tag{4.1}$$

$$S_{\rm RS} \sim \frac{1}{\beta}\sqrt{\frac{2}{\alpha}}\frac{\pi^2}{24\beta_{\rm tr}} \tag{4.2}$$

$$\rho(\Lambda) \to \frac{1}{\sqrt{2\pi}}\exp -\frac{1}{2}\left(\Lambda - \sqrt{\frac{2}{\pi\alpha}}\right)^2 . \tag{4.3}$$

We recognise (4.3) as identical with $\rho_{\rm cH}$ (2.4). Therefore it would appear that, provided m_t approaches zero sufficiently quickly as $\beta \to \infty$, the replica-symmetric entropy remains positive, and the the clipped Hebb rule is reproduced by the training cost function. The close agreement between this limiting behaviour and that of the optimal network at $m_t = 0.1$ (and $\beta = 256$) implies that the clipped Hebb rule is already performing well at this noise level. However, it is clear that even though simulated annealing might be successful under such circumstances, being iterative it could scarcely compete with the, directly prescriptive, clipped Hebb rule.

5. Training using real-valued synapses

Having indicated that simulated annealing cannot be relied upon to produce an optimal binary-synapse network, one might wonder whether a viable network can be constructed from a companion network having real-valued weights, algorithms for whose construction are more accessible (e.g., for the noiseless regime, AdaTron (Anlauf and Biehl 1989) etc). The most obvious way of reducing a real-valued synapse to a binary-valued quantity is simply to take $J_{ij} = \mathrm{sgn}(J_{ij}^{\rm real})$, an operation that takes the Hebb rule into the clipped Hebb rule, for example. As an illustration that this means of training a binary system is in no way favourable, we have examined the effect of this operation on a maximally stable network, representing a noise-free training.

We imagine that, for a given neuron, the synaptic weights connecting to this neuron are constrained only by an inconsequential (spherical) normalization constraint, $\sum_j J_{ij}^2 = N$, and that choices of these J's are sought compatible with the maximally stable rule. This learning rule has been widely studied, particularly following the seminal work of Gardner (1987, 1988). If the parameter κ is maximized for a given loading, α, consistent with being able to store the patterns, one obtains a well defined choice of weights, which may then be clipped and used to form a binary-synapse system. For a network constructed in this fashion we calculate the resulting distribution of the stabilities of the patterns originally embedded by the parent system. This distribution may be obtained by a simple generalization of the methods of Kepler and Abbott (1988) and Gardner (1989); an outline of the steps in this derivation is given in appendix B.

$$\rho_{\rm cM}(\Lambda) = \frac{1}{\sqrt{2\pi}}\exp(\tfrac{1}{2}\Lambda^2)\cdot\tfrac{1}{2}\,\mathrm{erfc}\left(\frac{\kappa - \Lambda\sqrt{2/\pi}}{\sqrt{2(1-2/\pi)}}\right)$$
$$+ \frac{1}{\sqrt{2\pi(1-2/\pi)}}\exp\left(-\frac{(\Lambda - \kappa\sqrt{2/\pi})^2}{2(1-2/\pi)}\right)\cdot\tfrac{1}{2}\,\mathrm{erfc}(-\kappa/\sqrt{2})\,. \tag{5.1}$$

(The subscript 'cM' refers to 'clipped MSN'.) Whilst (5.1) is believed correct for all positive values of κ, for the range of this parameter for which it is reasonable to obtain a passable binary model, this expression may be simplified considerably. It would certainly be optimistic to expect this method of training to produce worthwhile results for α close to, or beyond, the capacity limit of this system, namely $\alpha = 0.83$. This means that κ should be chosen greater than about 0.6, which value produces a storage capacity of $\alpha = 0.84$ for underlying spherical model.

For 'large' positive κ one may simplify both (5.1) and the Gardner formula linking α to κ (giving $\kappa \sim \alpha^{-1/2}$). This yields

$$\rho_{\mathrm{cM}}(\Lambda) \sim \frac{1}{\sqrt{2\pi(1-2/\pi)}} \exp\left(-\frac{(\Lambda - \alpha^{-1/2}\sqrt{2/\pi})^2}{2(1-2/\pi)}\right) \tag{5.2}$$

which expression is seen to resemble that for the clipped Hebb rule (2.4). (For $\kappa \sim 1.0$ this approximate form of ρ_{cM} is already representative of (5.1), although for positive κ its mean is slightly greater than that of the true distribution.) It would appear from this comparison that training a binary network by taking a trained spherical MSN, and clipping its synapses produces little better typical stabilization of the patterns than the use of the clipped Hebb rule, whose implementation is far easier. So, even with noise-free training, this method of realising a binary-network would not seem promising in isolation. Given that on introducing progressively more noise into the training procedure the clipped Hebb rule becomes genuinely optimal, there seems little reason to expect that a clipped spherical MSN would ever excel over the former rule if it does not do so for noiseless training.

6. Conclusion

We have investigated the application of training noise to neural networks having discrete-valued synapses, and found effects not observed in continuous synapse systems. Our results imply, on assuming the novel form of replica-symmetry breaking seen in binary-synapse networks, that a search for a unique noise-optimal network using simulated annealing cannot, for practical purposes, succeed (except under the very restrictive conditions of the clipped Hebb rule limit), even with a static cost-function (representing annealed noise, in the terminology of Wong and Sherrington (1991)). This thermodynamic approach suggests that disjoint regions of weight space exist, containing near-optimal networks, meaning that any arbitrarily started training algorithm is unlikely to reach the optimal network, or converge properly, if applied to a large system.

Our approach is in some ways dual to that of Horner (1992), who examined the dynamics of learning in the binary perceptron, for the maximally stable rule. The dynamic mean-field theory used by Horner involves taking the limit of large system size ($N \to \infty$) before examining the long-time learning behaviour ($t \to \infty$). The use of the Gibbs formalism represents the long-time limit being taken first, whereafter the thermodynamic limit ($N \to \infty$) is taken. Horner shows that, even for the MSN, the annealing method cannot be expected to converge to the result of the Gibbsian method even within those limits where Krauth and Mézard find replica-symmetry to remain intact. (Although there remains some scepticism about the validity of the one-step RSB scheme proposed, evidence in support of the results of Krauth and Mézard is not lacking; e.g. Krauth and Opper (1989), Derrida *et al* (1991).) The present results suggest that diverging timescales can occur even for very small training noise ($m_t \to 1$), and hence that the MSN may represent a very special learning rule.

For the spherical model, although a discontinuity in behaviour is seen between $m_t = 1^-$ and $m_t = 1$, the effects of ergodicity breaking seem less significant (Wong and Sherrington 1991).

These investigations, along with those of Horner, although strictly limited to modelling training by simulated annealing, suggest that learning in discrete-synapse systems is far more difficult to effect without a prescription (such as the clipped Hebb rule) than in continuous-synapse systems. The method of exact ennumeration, as used by Krauth and Opper (1989) to find, numerically, the storage capacity of the binary perceptron, would currently seem to be the most reliable method of optimizing such networks. Clearly this is a highly undesirable method for large systems, as the time required for training grows exponentially with system size. Although our results do not discount the possibility that simulated annealing might find the optimal network, or a close substitute, given enough computer time, the effects of the ergodicity breaking that we have found will be associated with diverging timescales for the annealing schedule. If these diverge exponentially in the system size, N, then the advantages of this approach over exact ennumeration may quickly be lost.

The use of genetic algorithms might ultimately prove valuable. Although the algorithm used by Köhler (1990) showed considerable advantages over an iterative scheme for training an MSN, Köhler implied that his algorithm was far from being ideal. However, given the existence of disparate near-equivalent solutions (suggested by the form of replica-symmetry breaking seen in binary-synapse systems) it would seem that some form of genetic optimization process would offer the most hopeful means of training such systems. The topology of the weight space of a discrete-synapse network is most naturally provided by the dynamics of the learning process. Simulated annealing, in conventional realizations, provides a Hamming metric on the weight space, which means that the near-optimal solutions, which the learning process will aim to choose between, cannot be kept within the field of view of the algorithm if convergence requires searching the locality of each near-optimal solution. The crossing of genotypes allows association of points separated by large Hamming distances, and with an appropriate genome might allow more objective comparison of possible solutions, without requiring the whole of the weight space to be explored, as for the method of exact ennumeration.

Acknowledgments

Helpful discussions of the work of Krauth and Mézard (1989) with Dr Marc Mézard, and of Derrida *et al* (1991) with Dr Bernard Derrida, are gratefully acknowledged. RWP would like to thank Jesus College, Oxford, for the generous award of a scholarship. We thank the SERC for financial support (under grant number 9130068X).

Appendix A

We will discuss some aspects of the derivation of (2.1), with particular concern to the one-step replica symmetry breaking which besets this object at low annealing temperature. A system of N neurons, joined by asymmetric weights, chosen such as to optimally stabilize αN states $S_j = \xi_j^\mu$, will be considered. Choices of the weights, J_{ij}, are influenced by a cost function $\sum_\nu g(\Lambda_i^\nu)$, where $\Lambda_i^\nu = \frac{1}{\sqrt{N}} \sum_j \xi_i^\nu J_{ij} \xi_j^\nu$ is the stability of pattern ν on site i.

Using replicas to perform averages over the choice of stored patterns, following Gardner (1989), one may write $\rho(\Lambda)$ in the form;

$$\rho(\Lambda) = \lim_{n\to 0}\left\langle \operatorname*{Tr}_{\{J_{ij}^b\in\{\pm 1\}\}} \delta_{\mathrm{Kr}}\left(\sqrt{N}\Lambda - \sqrt{N}\Lambda_i^{\mu,a}\right) \exp\left(\sum_{\substack{\nu=1,\ldots,\alpha N\\ b=1,\ldots,n}} -\beta g(\Lambda_i^{\nu,b})\right)\right\rangle_\xi . \tag{A.1}$$

The μ and a in the Kronecker δ refer to an arbitrary choice of pattern and replica. Introducing various Fourier decompositions of unity

$$1 = \sum_{y_\mu^b=-N}^{N} \int_{-\pi/2}^{3\pi/2} \frac{\mathrm{d}z_\mu^b}{2\pi} \exp\left(\mathrm{i}z_\mu^b\left(y_\mu^b - \sum_j \xi_i^\mu J_{ij}\xi_j^\mu\right)\right) \tag{A.2}$$

in order to extract the patterns from within the cost functions, the pattern average may be performed. For large N, the z_μ^b integrals are dominated by the regions $z_\mu^b \sim 0$ and $z_\mu^b \sim \pi$. With suitable translation and scaling of these integrals, the summations over y_μ^b may be converted to integrals, and the Kronecker δ in (A.1) replaced by a Dirac δ-function, provided $\rho(\Lambda)$ is re-interpreted as being a continuous distribution. Introducing some further identity operators, *à la* Gardner, one may reduce (A.1) to an integral representation of the form;

$$\begin{aligned}\rho(\Lambda) = \lim_{n\to 0} \int \prod_{b<c} \frac{\mathrm{d}q^{bc}\,\mathrm{d}\hat{q}^{bc}}{2\pi/N} \exp N\left(\mathrm{i}\sum_{b<c} q^{bc}\hat{q}^{bc} + \alpha G_0(\{q^{bc}\}) + G_1(\{\hat{q}^{bc}\})\right)\\ \times\left\{\int\prod_b \frac{\mathrm{d}y^b\,\mathrm{d}z^b}{2\pi}\exp\left(\mathrm{i}y^bz^b - \beta g(y^b) - \tfrac{1}{2}(z^b)^2\right)\right.\\ \left.\times\exp\left(-\sum_{b<c} q^{bc}z^bz^c\right)\delta(\Lambda - y^a)\right\}\end{aligned} \tag{A.3}$$

in which

$$\begin{aligned}G_0(\{q^{bc}\}) &= \ln\left\{\int\prod_b \frac{\mathrm{d}y^b\,\mathrm{d}z^b}{2\pi}\exp\left(\mathrm{i}y^bz^b - \beta g(y^b) - \tfrac{1}{2}(z^b)^2\right)\cdot\exp\left(-\sum_{b<c} q^{bc}z^bz^c\right)\right\}\\ G_1(\{\hat{q}^{bc}\}) &= \ln\left\{\operatorname*{Tr}_{\{J^b\}}\exp\left(-\mathrm{i}\sum_{b<c}\hat{q}^{bc}J^bJ^c\right)\right\}\end{aligned} \tag{A.4}$$

(cf Krauth and Mézard 1989). Invoking mean-field theory, one may replace the q^{bc} and $\hat{q}^{bc}$ integrations by the values of these parameters at the extremum of the action $G = \left(\mathrm{i}\sum_{b<c} q^{bc}\hat{q}^{bc} + \alpha G_0(\{q^{bc}\}) + G_1(\{\hat{q}^{bc}\})\right)$. In the limit $N\to\infty$ this mean-field theory is exact. It is usual to assume, by way of an ansatz, that these order-parameters are replica symmetric, i.e. $q^{bc} = q$ and $\hat{q}^{bc} = \mathrm{i}\hat{q}$ $\forall b < c$, whereby one may obtain (2.1), (2.2) and (2.3). As a move towards improving this idealization, one may adopt a one-step symmetry breaking,

$$q^{bc} = \begin{cases} q_0 & I(b/m)\neq I(c/m)\\ q_1 & I(b/m) = I(c/m)\end{cases} \quad\text{and}\quad \hat{q}^{bc} = \begin{cases} \mathrm{i}\hat{q}_0 & I(b/m)\neq I(c/m)\\ \mathrm{i}\hat{q}_1 & I(b/m) = I(c/m)\end{cases}$$

where

$$I(x) = \inf\{y \in \mathbb{Z} : y \geqslant x\}.$$

On taking the limit $n \to 0$ this produces a distribution for the mutual overlaps of the various ergodic components given by

$$P(q) = m\delta(q - q_0) + (1 - m)\delta(q - q_1). \tag{A.5}$$

(For more details of this procedure see, e.g., Mézard *et al* 1987.) However, for the maximally stable network (where $g(\Lambda) = \theta(\kappa - \Lambda)$), on seeking the extremum of the one-step broken free-energy functional numerically, Krauth and Mézard found either that the replica-symmetric result was reproduced ($q_1 = q_0$ and $\hat{q}_1 = \hat{q}_0$) or that $q_1 \to 1$ and $\hat{q}_1 \to \infty$. They therefore investigated the effect of assuming $q_1 = 1$ in the free-energy functional, which simplifies considerably in this limit (directly associated with $\hat{q} \to \infty$)

$$G^{(1)}_{\mathrm{RSB}}(q_0, \hat{q}_0, 1, \infty, m, \beta) = (1/m)G_{\mathrm{RS}}(q_0, m^2\hat{q}_0, m\beta). \tag{A.6}$$

The condition determining m, that $G^{(1)}_{\mathrm{RSB}}$ should be extremized, then reduces to the constraint that the replica-symmetric entropy should vanish

$$S_{\mathrm{RS}}(q_0, m^2\hat{q}_0, m\beta) = S_{\mathrm{RS}}(q, \hat{q}, m\beta) = 0 \tag{A.7}$$

and hence that $m\beta = \beta_c =$ constant outside the domain of genuine replica-symmetry. This, together with (A.5), suggests that on reducing the annealing temperature, a unique domain becomes favoured amongst the many disjoint ergodic regions of weight space, although all domains will have finite energy. Thus Krauth and Mézard were led to adopt the condition of vanishing zero-temperature replica-symmetric entropy as that determining the storage capacity, α_c. The maximum of this quantity (obtained for $\kappa = 0$) was found to be 0.833.

These algebraic results of the substitution $q_1 = 1$ can be shown to be identical for all cost functions $g(\Lambda)$, with the same effective temperature, $m\beta$, also appearing in (2.1). We will assume that the same form of replica symmetry breaking holds for a general cost function, not solely the θ-function used in Krauth and Mézard (1989). The numerical difficulties of checking this assumption would be considerable, and for a cost function of the form used in this paper, unlikely to readily produce convincing evidence.

Appendix B

An overview of our derivation of (5.1) will be given. In view of the less pathological nature of networks having connected weight spaces, we have adopted a micro-canonical optimization procedure, in which only that region of the weight space which stabilizes the selected patterns needs be considered. The distribution $\rho_{\mathrm{cM}}(\Lambda)$ may then be obtained as an average over this region, but on saturating the parent network, thereby shrinking this region essentially to a point, ρ_{cM} is expected to become the distribution of stabilities produced by forming a binary network from a unique parent. Our starting point is

$$\rho_{\mathrm{cM}}(\Lambda) = \lim_{n\to 0}\left\{ \prod_{b=1}^{n} \int \{\mathrm{d}J_j^b\}\, \delta\left(\sum_{j=1}^{N} J_j^{b2} - N\right) \right.$$
$$\left. \times \prod_{\substack{\mu=1,\ldots,\alpha N \\ b=1,\ldots,n}} \theta(\Lambda_i^{\mu,b} - \kappa)\, \delta_{\mathrm{Kr}}\left(\sqrt{N}\Lambda - \xi_i^\nu \sum_j \mathrm{sgn}(J_j^a)\xi_j^\nu\right)\right\}_\xi . \tag{B.1}$$

Following Gardner (1987, 1988) we introduce identity operators, in the form $1 = \prod_{\mu,b} \int \mathrm{d}y_\mu^b \, \delta(y_\mu^b - \Lambda_i^{\mu,b})$, taking Fourier representations of the δ-functions. Use of an analogous procedure for the single quantity $\xi_i^\nu \sum_j \mathrm{sgn}(J_j^a)\xi_j^\nu$ seems to be simplified by taking an entire set of identity operators, $1 = \prod_{\mu,b} \sum_{u_\mu^b} \delta_{\mathrm{Kr}}\big(u_\mu^b - \xi_i^\mu \sum_j \mathrm{sgn}(J_j^b)\xi_j^\mu\big)$. Having extricated the patterns, ξ_i^μ, from within the θ-functions and the Kronecker δ in (B.1), one may perform the pattern average. (Analogously to the calculation of $\rho(\Lambda)$, one may replace the summation over u_μ^b by an integral, as N becomes large.) Thus three quantities unfamiliar from previous calculations emerge, namely

$$r^{bc} = \frac{1}{N}\sum_j \mathrm{sgn}(J_j^b)\,\mathrm{sgn}(J_j^c) \qquad s^b = \frac{1}{N}\sum_j |J_j^b| \qquad t^{bc} = \frac{1}{N}\sum_j J_j^b \,\mathrm{sgn}(J_j^c) \tag{B.2}$$

in addition to the standard $q^{bc} = (1/N)\sum_j J_j^b J_j^c$. Yet further partitions of unity are introduced to cater to these terms, and replica-symmetric mean-field theory is ultimately invoked in the limit $N \to \infty$. Thus one obtains $\rho_{\mathrm{GM}}(\Lambda)$ in the form

$$\rho_{\mathrm{GM}}(\Lambda) = \lim_{n\to 0} \prod_{b=1}^{n} \int \mathrm{d}y^b \frac{\mathrm{d}z^b}{2\pi}\mathrm{d}u^b \frac{\mathrm{d}v^b}{2\pi} \exp\big(\mathrm{i}y^b z^b - \tfrac{1}{2}(z^b)^2 + \mathrm{i}u^b v^b - \tfrac{1}{2}(v^b)^2 - z^b v^b s\big)\,\theta(y^b - \kappa)$$
$$\times \exp\Big(-\sum_{b<c} z^b z^c q - \sum_{b<c} v^b v^c r - \sum_{b\neq c} z^b v^b t\Big)\,\delta(\Lambda - u^a) \tag{B.3}$$

$$= \int \mathrm{D}x\,\mathrm{D}k \frac{\mathrm{e}^{\mathrm{i}k\Lambda}}{\sqrt{2\pi}} \left[\mathrm{erfc}\left(\frac{x\sqrt{q} + \kappa + \mathrm{i}ks}{\sqrt{2(1-q)}}\right)\right]\left[\mathrm{erfc}\left(\frac{x\sqrt{q} + \kappa + \mathrm{i}kt}{\sqrt{2(1-q)}}\right)\right]^{-l} \tag{B.4}$$

in which $\{q, r, s, t\}$ are chosen, along with $\{\varepsilon, \hat{q}, \hat{r}, \hat{s}, \hat{t}\}$, so as to extremize, in the limit $n \to 0$, the free-energy functional

$$G(\varepsilon, s, \hat{s}, q, \hat{q}, r, \hat{r}, t, \hat{t}) = \big(n\varepsilon - ns\hat{s} + \tfrac{1}{2}n(1-n)q\hat{q} + \tfrac{1}{2}n(1-n)r\hat{r} + n(1-n)t\hat{t}$$
$$\alpha G_0(q, r, s, t) + G_1(\varepsilon, \hat{q}, \hat{r}, \hat{s}, \hat{t})\big) \tag{B.5}$$

within which

$$G_0(q, r, s, t) = \ln\left\{\prod_b \int \mathrm{d}y^b \frac{\mathrm{d}z^b}{2\pi}\mathrm{d}u^b \frac{\mathrm{d}v^b}{2\pi} \exp\big(\mathrm{i}y^b z^b - \tfrac{1}{2}(z^b)^2 + \mathrm{i}u^b v^b - \tfrac{1}{2}(v^b)^2 - z^b v^b s\big)\right.$$
$$\left.\times\,\theta(y^b - \kappa)\exp\Big(-\sum_{b<c} z^b z^c q - \sum_{b<c} v^b v^c r - \sum_{b\neq c} z^b v^b t\Big)\right\} \tag{B.6}$$

$$G_1(\varepsilon, \hat{q}, \hat{r}, \hat{s}, \hat{t}) = \ln\left\{\prod_b \int \mathrm{d}J^b \exp\big(-\varepsilon J^{b2} + \hat{s}|J^b|\big)\right.$$
$$\left.\times \exp\Big(\sum_{b<c} J^b J^c \hat{q} + \sum_{b<c} \mathrm{sgn}(J^b)\,\mathrm{sgn}(J^c)\hat{r} + \sum_{b\neq c} J^b \,\mathrm{sgn}(J^c)\hat{t}\Big)\right\}. \tag{B.7}$$

Gardner's calculation of the capacity of the spherical model (Gardner 1988) involved a very similar form of free-energy functional, except that all the parameters $\{s, \hat{s}, r, \hat{r}, t, \hat{t}\}$ were absent. However, on simplifying G_0, one finds that this object is actually independent of $\{r, s, t\}$, and is therefore identical to the corresponding function for Gardner's model. This

simplification is entirely natural, given that these order parameters had no bearing on the thermodynamics of the spherical model simply storing patterns. As a consequence of this, one may deduce from the saddle-point conditions $\partial G/\partial r = \partial G/\partial s = \partial G/\partial t = 0$ that the conjugate order parameters $\{\hat{r}, \hat{s}, \hat{t}\}$ are all zero. Therefore, it is only necessary to evaluate G_1 to first order in the latter quantities, so that the values of $\{r, s, t\}$, the latter two of which are needed in $\rho_{cM}(\Lambda)$, may be obtained via the conditions $\left(\partial G/\partial \hat{r}\right)\big|_{\hat{r}=0} = \left(\partial G/\partial \hat{s}\right)\big|_{\hat{s}=0} = \left(\partial G/\partial \hat{t}\right)\big|_{\hat{t}=0} = 0$. With this reduction, G_1 may be evaluated thereby again revealing close similarity with Gardner's analysis, but in addition giving expressions for the new order parameters

$$r = \langle \overline{\mathrm{sgn}(J_j)}^2 \rangle_\xi = \int \mathrm{D}x \left[\mathrm{erf}\left(\frac{x\sqrt{q}}{\sqrt{2(1-q)}} \right) \right]^2$$

$$s = \langle \overline{|J_j|} \rangle_\xi = \sqrt{\frac{2}{\pi}} \tag{B.8}$$

$$t = \langle \overline{J_j} \cdot \overline{\mathrm{sgn}(J_j)} \rangle_\xi = q\sqrt{\frac{2}{\pi}}$$

(The notation $\langle \overline{f} \rangle_\xi$ denotes an average of f over accessible regions of the spherical weight space, followed by a disorder average, over the choice of stored patterns.) As we will be interested in saturating the spherical model which underlies the derived binary model, we take the limit $q = \langle \overline{J_j}^2 \rangle_\xi \to 1$, in which limit the couplings become closely-defined (without any of the problems that plague the true binary model in such a limit). This limit reproduces the Gardner formula linking κ and α_c;

$$\alpha_c^{-1} = \int_{-\kappa}^{\infty} \mathrm{D}t \, (\kappa + t)^2 \tag{B.9}$$

as would have been expected. In addition, the x-integration in (B.4) can be split into three segments:

(i) the section from $x = -\infty$ to $x = -(\kappa + \mathrm{i}ks)$ produces the first term in (5.1) on inserting asymptotic forms for the error functions;
(ii) the range $x = -(\kappa + \mathrm{i}kt)$ to $x = +\infty$ analogously yields the second term in (5.1);
(iii) the remaining section $x \in \left(-(\kappa + \mathrm{i}ks), -(\kappa + \mathrm{i}kt)\right)$ produces a contribution that vanishes in the limit $q \to 1$.

Thus one obtains (5.1). (It is our assumption that the x-integration contour may be distorted as necessary without crossing any poles of the integrand.)

References

Anlauf J K and Biehl M 1989 *Europhys. Lett.* **10** 687
de Almeida J R L and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983
Derrida B, Griffiths R B and Prügel-Bennett A 1991 *J. Phys. A: Math. Gen.* **24** 4907
Gardner E J 1987 *Europhys. Lett.* **4** 481
—— 1988 *J. Phys. A: Math. Gen.* **21** 257
—— 1989 *J. Phys. A: Math. Gen.* **22** 1969
Gardner E J and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
Györgyi G 1990 *Phys. Rev. Lett.* **64** 2957

Horner H 1992 *Z. Phys.* B **86** 291
Holmstrom L and Koistinen P 1992 *IEEE Trans. Neural Networks* **3** 24
Köhler H 1990 *J. Phys. A: Math. Gen.* **23** 1265
Kepler T and Abbott L 1988 *J. Physique* **49** 1657
Krauth K and Mézard M 1989 *J. Physique* **50** 3057
Krauth W and Opper M 1989 *J. Phys. A: Math. Gen.* **22** L519
Murray A F 1991 *Electronics Lett.* **27** 1546
Mézard M, Parisi G and Virasoro M 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
Parisi G 1980 *J. Phys. A: Math. Gen.* **13** 1101
Press W H, Flannery B P, Teukolsky S A and Vetterling W T 1988 *Numerical Recipes in C* (Cambridge: Cambridge University Press)
Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
van Hemmen J L 1987 *Phys. Rev.* A **36** 1959
Wong K Y M and Sherrington D 1990a *J. Phys. A: Math. Gen.* **23** L175
—— 1990b *J. Phys. A: Math. Gen.* **23** 4659
—— 1991 Retrieval behaviours and basin structures of noise-optimal neural networks *Oxford University Preprint* OUTP-91-36S