A perceptron with a skeletal weight-space

1994 J. Phys. A: Math. Gen. 27 23

(http://iopscience.iop.org/0305-4470/27/1/003)

J. Phys. A: Math. Gen. 27 (1994) 23–39. Printed in the UK

A perceptron with a skeletal weight-space

R W Penney and D Sherrington

Department of Physics, Theoretical Physics, Oxford University, 1 Keble Road, Oxford OX1 3NP, UK

Received 28 September 1993

Abstract. A perceptron whose space of interactions can interpolate between spherically constrained and binary-valued synapses is introduced and investigated as an associative neural-network memory. For maximally stable storage, and where the weight-space remains connected, the critical storage capacity, α_c, is found to be reduced by a factor determined solely by the geometry of the weight space, and is shown to interpolate, within the replica-symmetric approximation, between $\alpha_c = 2$ (in the Gardner-model limit) and $\alpha_c = 4/\pi$. Various comparisons of the synaptic weights with those of the binary perceptron show that such differences as remain between this weight space and that of the true binary perceptron are crucial to obtaining $\alpha_c \geqslant 4/\pi$. Although these differences limit the use of such models in realizing optimal binary networks, they may yet provide worthwhile binary systems by simple weight clipping. Simulation results are presented in support of the theoretical analyses.

1. Introduction

The paradigm of neural networks, the perceptron (Minsky and Papert 1969) can exist in a great many varieties but has perhaps two extreme and contrasting forms, which have been the focus of a great deal of investigation. At one extreme, properties of, and training algorithms for, perceptrons whose controlling variables (predominantly the synaptic weights $\{J_i\}$) can each take essentially any real value, are quite well documented (e.g. Gardner 1988, Anlauf and Biehl 1989). In contrast, systems whose synapses are constrained to be discrete variables, themselves also the subject of much study, show greater pathological behaviour than the former, most dramatically in the complexity of their training procedure (e.g. Gutfreund and Stein 1990, Pérez Vicente *et al* 1992). A method of interpolating between these two poles is therefore likely to be instructive in better defining the challenges of the binary perceptron problem. In this paper we suggest a simple model that takes a small step in this direction.

The perceptron problem may be defined as follows: given a set of αN input patterns ('questions'), $\{\xi_i^\mu\}$, with associated outputs ('answers'), $\{\eta^\mu \in \{\pm 1\}\}$, the goal of training is to find a choice of synapses $\{J_i\}$ and a threshold ϕ which satisfy

$$\eta^\mu = \operatorname{sgn}\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} J_i \xi_i^\mu + \phi\Big) \qquad \forall \mu \in \{1, \cdots \alpha N\}. \tag{1.1}$$

In the present context we will assume, for simplicity, the questions and answers to be independent, unbiased stochastic variables, such that $\phi = 0$ may be assumed, and take $\langle \xi_i^{\mu 2} \rangle_\xi = 1$ ($\langle \xi_i^\mu \rangle_\xi = 0$). The system size N will be taken as large, with α and finite cumulant averages of $\{J_i\}$ and $\{\xi_i^\mu\}$ all as being of order N^0 as $N \to \infty$. The indeterminacy

0305-4470/94/010023+17$07.50

of the answers to input questions means that solutions to the problem can only exist for limited loading, α, and thus a critical storage capacity may be defined. This also means that such a perceptron more closely resembles a heteroassociative memory than a cybernetic network.

Given that (1.1), with $\phi = 0$, is invariant under the scaling $J_i \to \lambda J_i \ \forall i$, the perceptron vector, $\boldsymbol{J} = (J_1, \cdots J_N)$ may be normalized without affecting the tractability of (1.1). This produces a spherical weight space if no additional *a priori* constraints on $\boldsymbol{J}$ are present. A much more restrictive selection of $\boldsymbol{J}$ is implied by the binary choice $J_i \in \{\pm 1\} \ \forall i$, although both these schemes are capable of some success with the goal of (1.1) (Gardner 1988, Krauth and Mézard 1989). However, in common with the general intractability of integer-programming tasks relative to unconstrained optimization, those algorithms proposed for training the binary perceptron (e.g. Krauth and Opper 1989, Amaldi and Nicolis 1989, Köhler 1990) are generally both more complicated and less successful than those known for the spherical system (e.g. Minsky and Papert 1969, Krauth and Mézard 1987, Anlauf and Biehl 1989). The disconnectedness of the binary model's hypercubic weight space is at the root of this difficulty given that optimal solutions to (1.1) do not lie in a a single region of the weight space and that iterative local enhancement of an approximation to the perceptron vector is more difficult to effect than in a continuous weight space.

Consider therefore a perceptron whose space of couplings is defined by

$$\mathcal{W}_N(B) = \left\{ (J_1, \cdots J_N) \in \mathbb{R}^N : \sum_i J_i^2 = N, \ |J_i| \leqslant B \ \forall i \right\} \tag{1.2}$$

so that $\boldsymbol{J}$ lies on the surface of a sphere, but with each J_i being locally constrained by a bound B, such as to always permit binary-valued $\boldsymbol{J}$. Clearly as $B \to \infty$ (and strictly, for $B > \sqrt{N}$), $\mathcal{W}_N$ approaches an entire spherical surface but as $B \to 1^+$, $\mathcal{W}_N$ contains disconnected regions centred on $\boldsymbol{J} = (\pm 1, \pm 1, \cdots)$. In this paper we will examine such networks in storing patterns according to the maximum stability rule, for which the aligning fields, $\Lambda^\mu = N^{-\frac{1}{2}} \sum_j \eta_i^\mu J_i \xi_i^\mu$, of all patterns must exceed a positive threshold κ. (This represents a slightly more demanding requirement than the original task, but one that facilitates the correct response to degraded input patterns.) It will be advantageous to define a set of candidate solution vectors by

$$\mathcal{C}_N(\{\xi^\mu\}) = \left\{ (J_1, \cdots J_N) \in \mathbb{R}^N : \sum_i J_i^2 = N, \ \frac{1}{\sqrt{N}} \sum_i \eta_i^\mu J_i \xi_i^\mu > \kappa \ \forall \mu \right\} \tag{1.3}$$

so that $\mathcal{S} \equiv \mathcal{C} \cap \mathcal{W}$ represents those networks which successfully learn the responses to the input questions. In the following sections, we study the critical storage capacity $\alpha_c(\kappa)$ beyond which no solutions to (1.1) exist on average, as a function of B and seek insight into the distribution of $\boldsymbol{J}$ within $\mathcal{S}$ with a view to characterizing the approach of this perceptron to the binary model.

2. Investigation of the weight space

One may identify a number of general properties of the present perceptron problem for general N. Firstly, as is well known, the candidate solution space $\mathcal{C}_N$ is convex and connected, being the intersection of convex and connected regions defined by the planes $\eta^\mu \boldsymbol{J} \cdot \boldsymbol{\xi}^\mu = \kappa/\sqrt{N}$ †. Regarding $\mathcal{W}$ itself, clearly the points $\boldsymbol{J} = \{\pm 1, \pm 1, \cdots\}$ always lie

† The scalar product of two N-vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined as $\frac{1}{N} \sum_l a_l b_l$.

within $\mathcal{W}$ if $\mathcal{W} \neq \emptyset \Leftarrow B \geqslant 1$. $\mathcal{W}$ will be connected if there exist paths connecting these vertices. A geodesic linking two points a and b in a spherical space may be defined by

$$c = \gamma \cdot \frac{1}{2}(a - b) + \sqrt{\frac{2 - \gamma^2(1 - a \cdot b)}{1 + a \cdot b}} \cdot \frac{1}{2}(a + b) \tag{2.1}$$

where the position vectors are relative to the origin of the sphere in its embedding Euclidian space. This geodesic represents the shortest arc linking a and b, and is such that $\gamma = \pm 1$ represent the end-points and $\gamma = 0$ is mid-way between a and b. Considering two representative adjacent corners, such as $(+1, +1, +1, \cdots)$ and $(-1, +1, +1, \cdots)$, it may be seen that the magnitudes of the synaptic weights along the geodesic are given by the two coefficients in (2.1). By seeking the maximum values of these coefficients along this path, one may see that no synapse will violate the confinement constraint $|J_i| < B$ if

$$\frac{1}{1 - \frac{1}{N}} \leqslant B^2. \tag{2.2}$$

This immediately shows that $\mathcal{W}_N$ will be disconnected in $N = 2$ if $B < \sqrt{N}$, as is geometrically clear, but that as $N \to \infty$, $\mathcal{W}_N$ is disconnected only if $(B - 1) = O(N^{-1})$. Hence, comparing $\mathcal{W}_N$ with figure 1, the 'edges' linking the corner regions of weight space will always be accessible if $(B - 1) = O(N^0)$ as $N \to \infty$. Figure 1 also makes it apparent that the solution space of the perceptron problem for $J \in \mathcal{W}$ need not be connected for arbitrary positive κ.

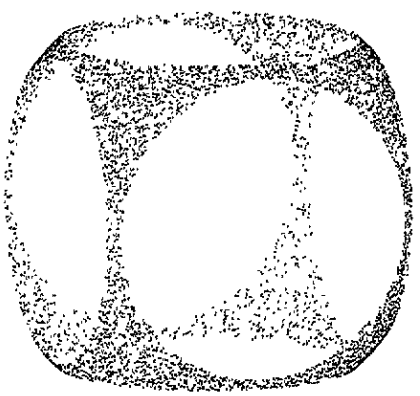

Figure 1. A representation of the weight space $\mathcal{W}_N$ for $N = 3$, as an aid to visualization.

For large dimensionality N, one may calculate straightforwardly the fractional volume of the sphere $\sum_i J_i^2 = N$ that lies within $\mathcal{W}_N$. Denoting this ratio by F, one finds that this quantity is always exponentially small, i.e. $-\ln F = O(N)$, indicating that the bounds $|J_i| \leqslant B$ sharply reduce the accessible weight-space volume. Confining attention to the asymptotics of this fraction $-N^{-1} \ln F \sim \sqrt{2/\pi} \exp(-\frac{1}{2}B^2)/B$ as $B \to \infty$, while for $B \to 1$ we have $-N^{-1} \ln F \sim -\ln(B - 1)$. Therefore it would appear that the approach towards the Gardner limit ($B \to \infty$) is much swifter than that towards the binary model ($N(B - 1) \to 0$).

In the interests of conciseness, and in view of the frame-like nature of $\mathcal{W}_N$, the perceptron whose synaptic-weights lie within $\mathcal{W}_N$ will be referred to as 'skeletal'.

3. The saturation condition

In a context where typical properties of perceptrons are desired and therefore, where the patterns ξ^μ are stochastic variables, the bound α_c beyond which solutions to (1.1) do not exist naturally cannot be defined other than statistically and this inevitably makes its definition more flexible and less intuitive. In addition, such conditions as are typically applied to the spherical model and to the binary model appear to be starkly dissimilar. Given that our goal is to interpolate between these extremes, it is germane to attempt to reinterpret the original analyses of these two models in a way that facilitates their generalization to the present problem.

Two aspects of the solution set $\mathcal{S}$ are particularly indicative. Firstly, there is an appropriate measure of its size ($|\mathcal{S}|$); if this is sufficiently large on average, then one would assert that $\alpha < \alpha_c$. Secondly, one might consider the breadth of dispersal of solutions; as solutions become scarcer, one would expect these to become increasingly localized. In Gardner's pioneering work on the spherical model (1987, 1988), it is the second form of indicator that was used to determine α_c, whilst the former test is more relevant to Krauth and Mézard's study (1989) of the binary perceptron. However, in the context of the binary model, although the resulting zero-entropy condition is quite intuitive, it is definitely inequivalent to the condition used by Gardner (technically, $q \to 1$), a result that can be attributed to the disconnectedness of the solution space of the binary problem. It therefore seems more helpful to recast the localization condition into one of vanishing solution volume, given that these two conditions are equivalent in the case of Gardner's work. Nevertheless, there remains an additional complication. For the binary model, the weight space consists of discrete points, whilst that of the spherical model is continuous, and therefore the volume measures appropriate to these two spaces must be different and so are not immediately comparable. Again, in hoping to mediate between these two extremes, some reconciliation must be achieved.

Consider, therefore, the fractional volume V of weight space that lies within the solution space. In general, $\mathcal{W}$ will consist of n_a disjoint connected domains; $\mathcal{W} = \bigcup_a \mathcal{W}^a$, $\mathcal{W}^a \cap \mathcal{W}^b = \emptyset \ \forall a \neq b$. Defining $\mathcal{S}^a = \mathcal{C} \cap \mathcal{W}^a$, one has

$$V \equiv \frac{|\mathcal{S}|}{|\mathcal{W}|} = \sum_{a=1}^{n_a} \frac{|\mathcal{S}^a|}{|\mathcal{W}^a|} \frac{|\mathcal{W}^a|}{|\mathcal{W}|} = \frac{|\mathcal{W}^1|}{|\mathcal{W}|} \cdot n'_a \left\langle \frac{|\mathcal{S}^{a'}|}{|\mathcal{W}^{a'}|} \right\rangle_{a'} \tag{3.1}$$

assuming that the weight space is sufficiently homogeneous for all domains to have the same volume. n'_a denotes the number of weight-space domains that contain a point in the solution set $\mathcal{S}$ so that $|\mathcal{S}^{a'}| \neq 0$; $\langle\ \rangle_{a'}$ represents an average over this subset of domains. In the thermodynamic limit $N \to \infty$, because for $\alpha < 1$ the αN conditions $\Lambda^\mu = \kappa$ section $\mathcal{W}_N$ into $2^{\alpha N}$ regions no more than one of which can satisfy $\Lambda^\mu \geqslant \kappa \ \forall \mu$, it is expected that $\ln V \sim N$ and that this quantity be self-averaging in the patterns, ξ^μ, i.e. it should depend only on the stochastic properties of these variables. It is thus reasonable to assume that α_c should be associated with a severe contraction of V and a loss of self-averaging. It is in such terms that α_c is usually defined analytically. In view of the relation of $\ln V$ to microcanonical statistical thermodynamics, we will refer to $\ln V$ as the 'entropy' of solutions to the perceptron problem within $\mathcal{W}$.

The separation of V into factors, as in (3.1), allows the criticality conditions of the spherical and binary models to be viewed together. For the former, the weight space is connected, so $n'_a = 1$ and $|\mathcal{W}^1|/|\mathcal{W}| = 1$ and saturation is signalled by a divergence of $N^{-1} \ln\{|\mathcal{S}^1|/|\mathcal{W}^1|\} \to -\infty$. In the binary network, $|\mathcal{W}^1|/|\mathcal{W}| = 2^{-N}$ and $|\mathcal{S}^{a'}|/|\mathcal{W}^{a'}| = 1$,

so that $N^{-1}\ln n'_a \to 0$ marks the onset of saturation. (With such definitions of criticality, it is entirely natural that additional pathologies (such as replica symmetry breaking) should be exhibited for $\alpha > \alpha_c$.) It is probable that for networks intermediate between these two limits, a less trivial interplay of the factors of $\ln V$ will determine the limit of existence of solutions.

However, as indicated in the previous section, if $(B-1) = O(N^0)$ then $\mathcal{W}_N$ is connected, so it is to be expected that, in terms of $\ln V$, such a system will show more similarity with the spherical model than the binary. It would therefore appear likely that it is in the region $(B-1) \sim N^{-1}$ that cross-over from one extreme to the other occurs. The ensuing analyses and simulations offer some support of this hypothesis, and suggest that the disconnectedness of the binary perceptron's weight space is fundamentally significant to its intractability, far more so than even a strong preference for $|J_i| \simeq 1$.

4. Analysis of the storage capacity

Using Elizabeth Gardner's, now standard, techniques of replica mean-field theory, one may determine the critical storage capacity, α_c, for the weight space $\mathcal{W}_N$, as $N \to \infty$ by calculating $\langle \ln V \rangle_\xi$. The calculation parallels the original work of Gardner (1988) very closely, so only brief comments on the analysis are appropriate. Restricting attention to $(B-1) = O(N^0)$, given that saturation will be signalled by a divergence of $\ln V$ one may ignore all non-divergent contributions to the entropy. This allows determination of the singular parts of $\ln V$ and of the distribution of synaptic weights, $p(J)$, to be combined, and eventually invites simplifying asymptotic analysis. Following Gardner, one may write

$$p(J) = \lim_{n\to 0}\left\langle \prod_{b=1}^{n}\prod_{i=1}^{N}\int_{-B}^{B} \mathrm{d}J_i^b\, \delta\Big(\sum_i J_i^{b2} - N\Big) \prod_{\mu=1}^{\alpha N} \theta\big(\Lambda^{\mu,b} - \kappa\big)\, \delta(J - J_j^a)\right\rangle_\xi \tag{4.1}$$

where the indices j and a represent an arbitrary choice of synapse and replica. By use of Fourier decompositions of unity, $p(J)$ is reduced to a field theory in replica-dependent order parameters, to which the method of steepest descent may be applied, a procedure that becomes exact as $N \to \infty$. Assuming a replica-symmetric saddle-point, $p(J)$ may be given in terms of an extremization problem, as follows:

$$p(J) = \lim_{n\to 0}\prod_{b=1}^{n}\int_{-B}^{B} \mathrm{d}J^b\, \exp\{-\varepsilon J^{b2}\}\, \exp\Big\{\hat{q}\sum_{b<c} J^b J^c\Big\}\, \delta(J - J^a) \tag{4.2}$$

in which ε and $\hat{q}$ are chosen, along with q, so as to extremize the function

$$G(q, \varepsilon, \hat{q}) = \big(n\varepsilon + \tfrac{1}{2}n(1-n)q\hat{q} + \alpha G_0(q) + G_1(\varepsilon, \hat{q})\big) \tag{4.3}$$

in the limit $n \to 0$, where

$$\begin{aligned} G_0(q) &= \ln\left[\prod_{b=1}^{n}\int \frac{\mathrm{d}y^b \mathrm{d}z^b}{2\pi} \exp\{\mathrm{i}y^b z^b - \tfrac{1}{2}z^{b2}\}\theta(y^b - \kappa)\, \exp\Big\{-q\sum_{b<c} z^b z^c\Big\}\right] \\ G_1(q) &= \ln\left[\prod_{b=1}^{n}\int_{-B}^{B} \mathrm{d}J^b\, \exp\{-\varepsilon J^{b2}\}\, \exp\Big\{\hat{q}\sum_{b<c} J^b J^c\Big\}\right]. \end{aligned} \tag{4.4}$$

The entropy of solutions is then determined by G_{extr}, to within additive constants independent of α. The order parameter $q = \frac{1}{N}\sum_i \langle J_i \rangle_S^2$ reflects the broadness of dispersal of solutions throughout $\mathcal{W}_N$, with $q \to 1$ indicating that solutions to (1.1) lie in a single small domain.

The discussion of the previous section suggests that if the weight space is connected, criticality is signalled by $\ln V \to -\infty$. In view of the trend of the spherical model, and in the absence of any obvious alternative means of obtaining a singularity in $G_{\text{extr}}(\alpha)$, we assume that $q \to 1$. It is worth emphasising that such an assumption for the binary perceptron is known to give an erroneously high storage capacity of $\alpha_c(\kappa = 0) = 4/\pi$ (Gardner and Derrida 1988, Krauth and Mézard 1989), and saturation occurs for $\alpha = 0.833$, where $q \simeq 0.56$.

By analogy with the work of Gardner, we assume that the limit $1 - q \equiv \delta \to 0$ is accompanied by divergences of the remaining order parameters, $\hat{q}$ and ε, and make the following *ansätze*:

$$\hat{q} \sim \mu/\delta^2 \qquad (\hat{q} + 2\varepsilon) \sim \nu/\delta \Rightarrow \varepsilon \sim -\tfrac{1}{2}\mu/\delta^2. \tag{4.5}$$

The saddle-point conditions appropriate to extremizing G then become

$$\mu = \alpha_c \int_{-\kappa}^{\infty} \mathrm{D}x\,(\kappa + x)^2 \tag{4.6}$$

$$\sqrt{\mu} - \frac{\sqrt{\mu}}{\nu} = 2\int_{B\nu/\sqrt{\mu}}^{\infty} \mathrm{D}x \left(Bx - \frac{x^2\sqrt{\mu}}{\nu}\right) \tag{4.7}$$

$$1 - \frac{\mu}{\nu^2} = 2\int_{B\nu/\sqrt{\mu}}^{\infty} \mathrm{D}x \left(B^2 - \frac{x^2\mu}{\nu^2}\right). \tag{4.8}$$

(The standard shorthand $\mathrm{D}x = \exp(-\frac{1}{2}x^2)\mathrm{d}x/\sqrt{2\pi}$ has been employed.) (4.8) may then be solved for the ratio μ/ν^2, with (4.7) being used to find μ, and hence α_c from (4.6). Comparison of (4.6) with the Gardner formula for the spherical model (specifically, $1 = \alpha_c \int_{-\kappa}^{\infty} \mathrm{D}x\,(\kappa + x)^2$) shows that the ratio $\mu = (\alpha_c(B)/\alpha_c(B \to \infty))$ is independent of κ, hence for a given pattern stability the storage capacity of the perceptron is reduced by solely a geometrical factor. The storage capacity is shown explicitly in figure 2 as a function of the bound B and for $\kappa = 0$. It is seen that for $B \simeq 2$, α_c is already very close to its unrestricted asymptote $\alpha_c = 2$ (where $\mu = \nu = 1$), but that as $B \to 1$, $\alpha_c > 0.833$; this disagreement with the binary-model limit indicates that the limits $B \to 1$ and $N \to \infty$ do not commute. Moreover, in this limit

$$\frac{\mu}{\nu^2} \sim \frac{2}{9\pi(B-1)^2} \tag{4.9}$$

which implies that

$$\alpha_c \to \frac{2}{\pi}\left\{\int_{-\kappa}^{\infty} \mathrm{D}x\,(\kappa + x)^2\right\}^{-1} \tag{4.10}$$

so that $\alpha_c(\kappa = 0) \to 4/\pi$ as $B \to 1$ (cf Gardner and Derrida 1988). The possibility that this superficial inconsistency is the result of a failure of the replica-symmetric *ansatz* is made less attractive by the evident connectedness of the solution space when the maxised stability

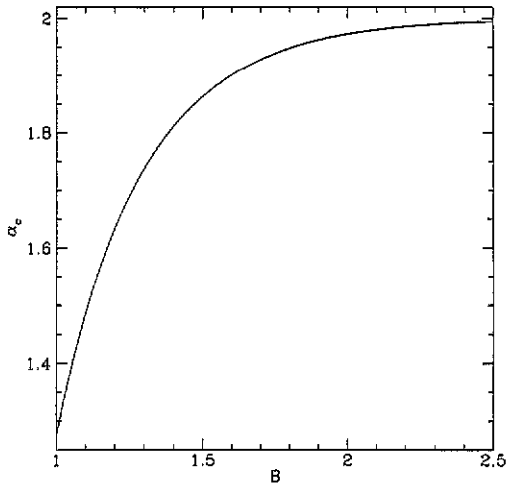

Figure 2. The storage capacity limit α_c versus the synaptic bound B for $\kappa = 0$.

κ is positive. (We note in passing that there is no unambiguous relationship between a disconnected solution space and replica symmetry breaking (O'Kane and Monasson 1993).)

Evaluation of $p(J)$ itself gives the following form of saturation:

$$p(J) = H\big(B\nu/\sqrt{\mu}\big)\big\{\delta(J+B)+\delta(J-B)\big\} + \frac{\nu}{\sqrt{2\pi\mu}}\exp\left\{-\frac{J^2\nu^2}{2\mu}\right\}\cdot\theta\big(B-|J|\big) \quad (4.11)$$

where $H(x) = \int_x^\infty \mathrm{D}y$. This should be compared and contrasted with the form appropriate for the spherical model, which has

$$p_{\mathrm{sph}}(J) = \frac{1}{\sqrt{2\pi}}\exp(-\tfrac{1}{2}J^2) \quad (4.12)$$

(as derived by Bouten *et al* 1990). It is appealing that the restricted freedom of the synaptic weights should result in compression of the extremities of the distribution into δ-functions at $J = \pm B$, and that a Gaussian profile should remain between these limits. However, although the expression (4.12) is valid for any loading within saturation, (4.11) applies only at saturation. A smoother distribution emerges as α is reduced, owing to the simplification of the perceptron problem, and also the associated reduction in importance of the restriction $|J_i| \leqslant B$.

In the limit $B \to 1$, $p(J)$ becomes almost flat between $J = \pm B$, with $p(J=0) \simeq \frac{3}{2}(B-1)$, indicating only a slow decay of this amplitude. Therefore, although synaptic weights $J = \pm 1$ become strongly favoured, near saturation the network makes use of the remaining freedom in $\mathcal{W}$, and will contain synaptic weights that span the entire interval $(-B, B)$. This feature would seem essential to exceeding the capacity of the true binary model, namely $\alpha_c = 0.833$. (Illustrative curves of $p(J)$ are given in figure 3.)

The distribution of pattern stabilities may also be calculated by similar methods, following Kepler and Abbott (1988), and on saturation has the simple form

$$\rho(\Lambda) = \theta(\Lambda-\kappa)\frac{\mathrm{e}^{-\Lambda^2/2}}{\sqrt{2\pi}} + \delta(\Lambda-\kappa)\,H(-\kappa) \quad (4.13)$$

familiar from the spherical model.

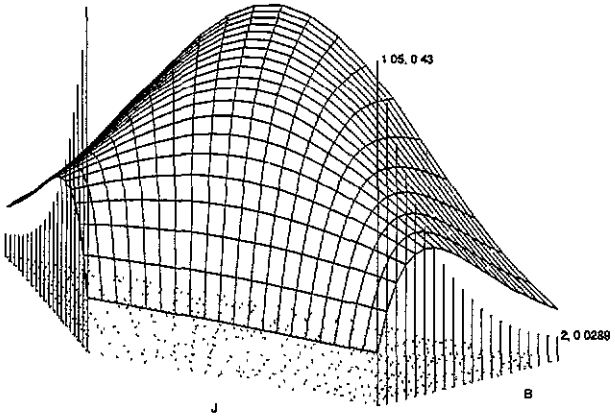

Figure 3. The distribution of synaptic weights for various bounds B.

5. Conversion to binary synapses

Although the properties of the weight space itself and the form of the maximum storage ratio, $\alpha_c(\kappa)$, both show that the skeletal model remains significantly different from the true binary perceptron (while $(B-1)=O(N^0)$), the similarity suggested by $p(J)$ to a binary weight space is almost overwhelming. It is therefore worthwhile seeking comparisons between the present system and the binary perceptron other than the storage capacity or weight-space geometry. There are perhaps two characteristics that seem particularly natural to explore. Firstly there would be properties of a binary perceptron formed by taking a trained skeletal model and clipping its synapses (thereby replacing J_i by $\mathrm{sgn}(J_i)$) in terms of the resulting stabilities of the patterns, ξ^μ. Alternatively one might seek to compare the choice of synaptic weights of a skeletal model and a binary model when both systems are trained independently on the same questions.

The distribution of pattern stabilities of the clipped skeletal model may be calculated closely following Penney and Sherrington (1993a), starting from

$$\rho_{\mathrm{clp}}(\Lambda)=\lim_{n\to 0}\Bigg\langle\prod_{b=1}^{n}\prod_{i=1}^{N}\int_{-B}^{B}\mathrm{d}J_i^b\,\delta\Big(\sum_{i=1}^{N}J_i^{b2}-N\Big)\prod_{\mu=1}^{\alpha N}\theta(\Lambda^{\mu,b}-\kappa)$$
$$\times\,\delta_{\mathrm{Kr}}\Big(\sqrt{N}\Lambda-\eta^\nu\sum_i\mathrm{sgn}(J_i^a)\xi_i^\nu\Big)\Bigg\rangle_\xi. \tag{5.1}$$

On saturation, this distribution may be reduced to the form

$$\rho_{\mathrm{clp}}(\Lambda)=\frac{\exp\left(-\frac{1}{2}\Lambda^2\right)}{\sqrt{2\pi}}H\left(\frac{\kappa-s\Lambda}{\sqrt{1-s^2}}\right)$$
$$+\frac{\exp\left(-\frac{1}{2}(t\kappa-\Lambda)^2/u\right)}{u\sqrt{2\pi}}H\left(-\frac{\kappa u+(s-t)(t\kappa-\Lambda)}{\sqrt{1-s^2}}\right) \tag{5.2}$$

in which

$$s=\lim_{q\to 1}\Big\langle\overline{|J_j|}\Big\rangle_\xi=2B\cdot H\big(B\nu/\sqrt{\mu}\big)+\sqrt{\frac{2}{\pi}}\frac{\sqrt{\mu}}{\nu}\left\{1-\exp\left(-\frac{B^2\nu^2}{2\mu}\right)\right\} \tag{5.3}$$

$$t=\lim_{q\to 1}\frac{1}{1-q}\Big\langle\overline{\big(J_j-\overline{J_j}\big)\big(\mathrm{sgn}(J_j)-\overline{\mathrm{sgn}(J_j)}\big)}\Big\rangle_\xi=\sqrt{\frac{2}{\pi\mu}} \tag{5.4}$$

$$u=\sqrt{1-2st+t^2}. \tag{5.5}$$

Curves illustrative of this $\rho_{clp}(\Lambda)$ are given in figure 4. The distribution, (5.2), can simply be shown to reduce, in the limit $B \to \infty$, to the analogous expression for the spherical model (Penney and Sherrington 1993a). However, taking the limit $B \to 1$ and using the asymptotic characters of μ and ν, we find

$$\rho_{clp}(\Lambda) \sim \theta(\Lambda - \kappa)\frac{e^{-\Lambda^2/2}}{\sqrt{2\pi}} + \delta(\Lambda - \kappa)\,H(-\kappa) \tag{5.6}$$

which is seen to be of precisely the same form as the distribution of stabilities of the underlying skeletal model (4.13). Despite this, beneath the superficial equivalence of $\rho_{clp}(\Lambda)$ and $\rho(\Lambda)$, there remains an important distinction. Whilst, asymptotically, both distributions contain step functions ($\theta(\Lambda - \kappa)$), because this constituent of $\rho_{clp}(\Lambda)$ involves a limit of an error function rather than being a true Heaviside function, the fraction of unstable patterns ($\int_{-\infty}^{0} \rho_{clp}(\Lambda)\,d\Lambda$) is always of order N^0. Therefore, although $\rho(\Lambda)$ of (4.13) represents a situation in which all patterns have stabilities exceeding κ (for a typical set of questions and answers), $\rho_{clp}(\Lambda)$ implicitly has an extensive number of questions which are wrongly answered, provided $(B-1) = O(N^0)$, even though this fraction becomes arbitrarily small as $B \to 1$. (The decay of $\int_{-\infty}^{0} \rho_{clp}(\Lambda)\,d\Lambda$ is fairly rapid, being proportional to $\exp\left(-\frac{1}{2}\kappa^2/(B-1)\right)$.) The work of Krauth and Mézard (1989) centres on binary perceptron which *strictly* stabilises the patterns ξ^μ, and as a result of greater difficulties involved in this problem, has a lower storage capacity for a given κ.

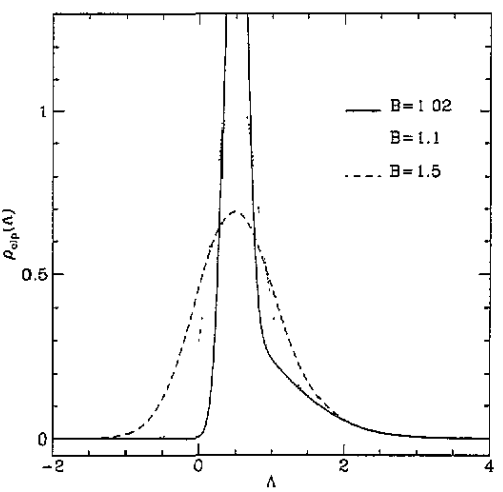

Figure 4. Aligning field distributions of the binary perceptron derived from the skeletal model by synaptic-weight clipping, for $\kappa = 0.5$.

It would seem from $\rho_{clp}(\Lambda)$ that the binary perceptron formed by weight clipping a skeletal model with $B \sim 1$, falls only slightly short of solving the perceptron problem accessible to its parent network. The distribution of synaptic efficiencies of this parent are also very nearly binary valued; moreover this network avoids the ambiguity of the true (thermodynamic) binary perceptron (where $q \not\to 1$ on saturation). Hence, it might be expected that in the limit $B \to 1$ the skeletal model is that unambiguously defined network whose synapses are most similar to those of the multiplicity of solutions to a given perceptron problem for $\{J_i\}$ having binary values.

An analytic means of comparing the synapses of spherical-synapse and binary-synapse solutions to a given perceptron problem was introduced by Penney and Sherrington (1993b),

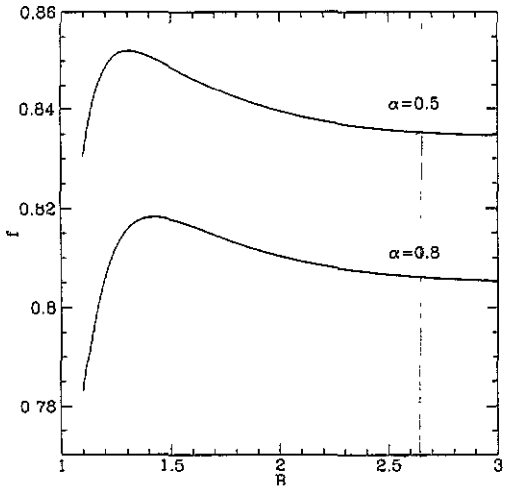

Figure 5. The fraction of synapses in the binary perceptron f that are correctly predicted by clipping those of a skeletal model, as a function of the bound B.

in analogy with the work of Wong *et al* (1992). The procedure yields the fraction of binary-valued synapses that are correctly predicted by clipping the synapses of the continuous network, as compared with the true binary solution for the given pattern set and hence the same loading, α. The generalization of the methods from the spherical model to the skeletal model is straightforward, so only the resulting algebraic conditions will be given. The fraction of correctly predicted synapses is given by $f = \frac{1}{2}(1 + p)$, where

$$p = 2\int \mathrm{D}x\, \tanh(x\sqrt{\hat{q}})\, H\left(\frac{-x\hat{s}}{\sqrt{\hat{q} - \hat{s}^2}}\right) \tag{5.7}$$

and s is determined self-consistently with $\hat{s}$ according to

$$\begin{aligned} s = {} & 2B\int \mathrm{D}x\, \tanh(x\sqrt{\hat{q}})\, H\left(\frac{B\sqrt{\hat{q}}\nu/\sqrt{\mu} - x\hat{s}}{\sqrt{\hat{q} - \hat{s}^2}}\right) \\ & + \frac{\sqrt{\mu}}{\nu}\left\{\hat{s}\int \mathrm{D}x\, \operatorname{sech}^2(x\sqrt{\hat{q}})\, \left(1 - 2H\left(\frac{B\sqrt{\hat{q}}\nu/\sqrt{\mu} - x\hat{s}}{\sqrt{\hat{q} - \hat{s}^2}}\right)\right)\right. \\ & \left. - \sqrt{\frac{2}{\pi}}\exp\left\{-\frac{B^2\nu^2}{2\mu}\right\}\int \mathrm{D}x\, \tanh(x\sqrt{\hat{q} - \hat{s}^2} + B\hat{s}\nu/\sqrt{\mu})\right\} \end{aligned} \tag{5.8}$$

and

$$\begin{aligned} \hat{s} = {} & \frac{\alpha}{\sqrt{\mu}}\int \mathrm{D}x \ln\left(\frac{\kappa^B + x\sqrt{q}}{\sqrt{1-q}}\right)\left[\left(\frac{s}{q}(1 - x^2) - \frac{\kappa^S x}{\sqrt{q}}\right) H\left(-\frac{\kappa^S\sqrt{q} + sx}{\sqrt{q - s^2}}\right)\right. \\ & \left. - \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(\kappa^S\sqrt{q} + sx)^2}{2(q - s^2)}\right\}\cdot\frac{x\sqrt{q - s^2}}{q}\right]. \end{aligned} \tag{5.9}$$

In (5.9), κ^B and κ^S refer to the, generally distinct, maximum stabilities achievable for the given pattern set by the binary and skeletal networks respectively. Solution of these equations for various loadings leads to the curves depicted in figure 5. It is seen that as B proceeds towards unity, the fraction of synapses in the true binary perceptron that are correctly predicted by clipping those of the skeletal model, is not monotonically increasing. This would suggest that although the binary perceptron and the skeletal model with $B \to 1$ show many superficial similarities, they are rather less similar in their internal structure.

Given that the maximum synaptic similarity between the skeletal and binary perceptron occurs at finite $(B-1)$ (presuming this quantity to be $O(N^0)$), we focus in figure 6 on the maximum value of f and the associated synaptic bound B, as an assessment of the utility of the former model in providing a starting point towards the construction of the optimal binary perceptron. Here f_{max} is compared with the corresponding fraction achieved by the spherical model, and also with the upper bound derived by taking $\hat{s} \to \sqrt{\hat{q}}$ in (5.7).

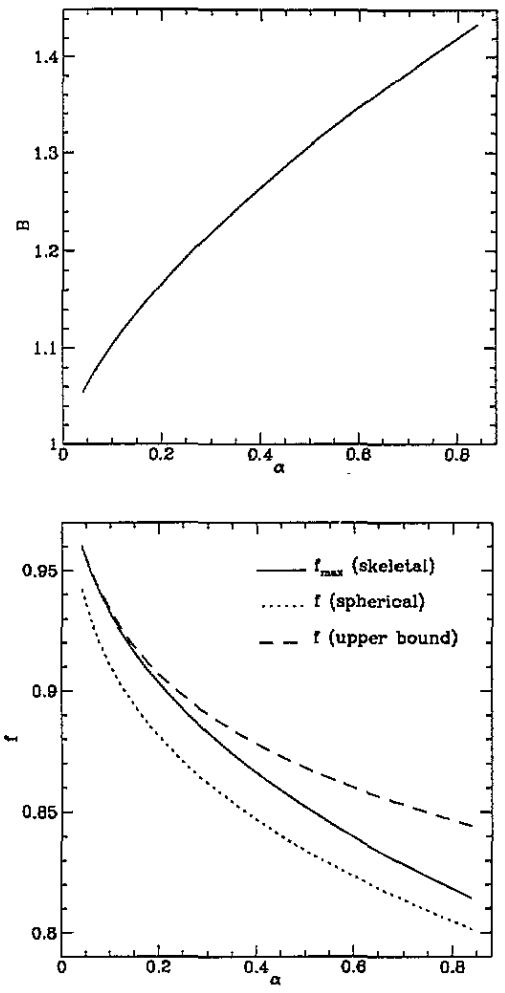

Figure 6. The synaptic bound that maximises the fraction of correctly predicted binary synapses (top) and the resulting fraction (bottom) as a function of the pattern loading α.

From these comparisons it would appear that the skeletal model has a slightly schizophrenic relationship with the binary perceptron. On the one hand, constructing a binary network by training in the continuous skeletal space for $B \simeq 1$, and then clipping the resulting synapses, produces a network that performs nearly perfectly as a binary-synapse decision unit. Yet, by contrast, the skeletal model with $B \simeq 1$ is less successful in predicting the actual synapses of the true binary solution than a more readily trained system having larger B, or indeed the spherical model itself.

6. Numerical simulations

Direct evidence in support of the theoretical analyses above clearly depends on being able to realize a network within the confines of $\mathcal{W}_N(B)$. In the absence of the local constraint $J_i \leqslant B$, a global spherical constraint could be ignored during an iterative training procedure, whereafter the perceptron vector could simply be normalized. Alternatively, a network with discrete synapses can be subjected to exact enumeration methods (Krauth and Opper 1989) or other techniques better suited to discrete rather than continuous optimization (e.g. Köhler 1990). Between these two extremes the geometry of the weight space is more awkward. Although a more methodical and efficient means of training the present model would be desirable (perhaps using methods of constrained optimization (e.g. Fletcher 1987)), we have employed as a first step a form of simulated annealing (Kirkpatrick *et al* 1983) in order to try to validate the theoretical assertions. It is known from the work of Horner (1992) that simulated annealing is largely unsuccessful when applied to the binary perceptron problem for $N > 50$, but the apparent simplicity of the solution space, $\mathcal{S}$, for the skeletal model, in contrast to the binary network, suggests that the annealing method is less likely to suffer impenetrable ergodicity breaking.

In essence, the method of simulated annealing proceeds by constructing a fictitious stochastic dynamics for the annealed variables, dependent on a temperature T_{an}, and an energy function on the space of the dynamical variables. Together these control the extent to which undesirable system configurations are suppressed; for large T_{an}, the dynamics allows wide exploration of the configuration space even if the energy landscape is rough, but as the annealing temperature is gradually reduced, the dynamics increasingly gravitates around low-energy areas. The choice of dynamics is central to any practical application of the method; the reconfigurations considered at each step of the algorithm should be sufficiently similar to recent states so as to capitalize on the successes achieved thus far and also be readily generated, yet should differ enough to allow broad exploration of the range of accessible configurations. In training the skeletal model for a specified set of patterns $\{\eta^\mu; \xi_i^\mu\}$, the objective function to be minimized is taken as minus the lower bound on pattern stability $-\kappa$, and a dynamics that always maintains J within $\mathcal{W}_N$ is used. It would, of course, be possible to consider the constraints implicit in $\mathcal{W}_N$ as soft, but only at the expense of specifying their strength relative to the importance of increasing κ.

For the binary-synapse network, the simplest dynamics would be a single-site process, whereby individual synapses are addressed and flipped according to a stochastic rule dependent on the annealing temperature. The constraints of $\mathcal{W}_N(B)$ do not allow single synapses to be considered in such a fashion, but the interests of simplicity would seem to favour an algorithm which focuses primarily on individual synapses rather than larger groups of couplings. One scheme would be to stochastically modify a particular synapse $J_m \to J_m'$, and thereafter simply scale all other weights equally so as to maintain $\sum_i J_i^2 = N$. This is essentially the approach that has been adopted. Due account must be taken of the constraints $|J_i| \leqslant B$, which frustrate this process in a number of ways. Not only must J_m' respect this bound, but the rescaling of the other synapses required by the proposed move $J_m \to J_m'$ must not violate these conditions either, thereby imposing a lower limit on $|J_m'|$ determined by the largest weight amongst the other synapses. Further, if no separation is made between synapses having $|J_i| = B$ and those with $|J_i| < B$, the dynamics can quickly become jammed as a significant number of couplings become saturated. Therefore, a separation between saturated and non-saturated couplings is made, with only those weights with $|J_i| < B$ being scaled following the change in the designated synapse. In this manner, relatively concise limits on the adjustment $J_m \to J_m'$ are defined, and a move $J \to J'$ can

be effected. This proposed change is then accepted or rejected according to the Metropolis algorithm (Metropolis *et al* 1953), with $-\kappa$ playing the role of the Hamiltonian. After considering such moves a number of times for each synapse in turn, corresponding to thermal equilibriation at the given annealing temperature, this temperature is reduced slightly and the cycle repeated. Every reconfiguration accepted is examined for its optimality, and after the temperature has been sufficiently reduced, the best configuration yet found is taken to represent the optimal solution to the perceptron problem within $\mathcal{W}$. (Given that as the true optimum solution is localized, the distance separating the current weight configuration and this optimum will decrease, it seems advantageous to reduce the step size of the tentative move $J_m \to J'_m$ with the annealing temperature. This is effected by updating J_m using a Gaussian random variable with standard deviation equal to the annealing temperature. The energy cost of this reconfiguration is scaled so as to be of order N^0 and $T^0_{\rm an}$.)

There are a number of aspects of the theory that we have tried to confirm. There is the assertion that, for a given stability κ, the maximum storage ratio is reduced relative to the Gardner limit, by a factor dependent only on B. Furthermore, there is the dependence of this ratio on B itself. Simulation can be used to provide measurements of κ for a given α, thereby determining $\kappa(\alpha_c)$, which may be used to determine an experimental value of μ from (4.6). Although for $N \to \infty$ such a quantity is expected to be independent of the precise choice of the pattern elements, for modest system sizes self-averaging is not strong, so that a comparison with the theory is effected only after averaging the results over various choices of patterns. In all cases $N = 50$, and averages are taken over 100 choices of patterns having Gaussian-distributed elements. The standard deviations of the calculated quantities over the choice of patterns are used to produce error-bars for the depicted mean quantities, making no allowance for any systematic failings of the training algorithm, which would be difficult to quantify. The experimental data are presented in figure 7.

It is seen that there is a plausible agreement of the trend of $\mu(B)$ with that predicted by theory, particularly for larger values of the bound B. Towards $B = 1$ and $\alpha = \alpha_c$, the increasing complexity of the learning problem and the heightened importance of finite-size effects, owing to the approach towards a dissected weight space, mean that convincing validation of the theory becomes grossly expensive in terms of computer time. However, the extrapolation of the trend of the simulation results towards $B = 1$ is perhaps more suggestive of $2/\pi \sim 0.64$, appropriate for the theory with $(B - 1) = O(N^0)$ rather than a value near 0.5, as for the true binary perceptron. As regards the constancy of $\mu(\alpha)$, simulations offer some support of this conjecture, but indicate that, at least for the method of training employed, significant fluctuations away from the theoretical value are to be expected for small systems. However, given the apparent lack of systematic disagreement between theory and experiment except towards $\alpha = \alpha_c$, the relevance of such practical considerations would seem apparent.

The observation that the fractional volume of the spherical space that lies within the skeletal weight space $\mathcal{W}_N$, namely F, is exponentially small in N does not preclude the possibility that $\mathcal{W}_N$ will actually contain the optimal spherical perceptron vector for a given set of patterns. For finite-size systems, this is likely to lead to significant fluctuations of the experimental value of κ, and hence μ, associated with such purely statistical effects. As system size N increases, these deviations are expected to decrease, and in the thermodynamic limit should be quashed entirely.

Regarding the analytic comparison of the skeletal model with the true binary perceptron, in view of the unexpected non-monotonicity in B of the fraction of binary synapses predictable by weight clipping f, numerical evaluation of this overlap was considered desirable. However, this task necessarily requires that both species of optimal network can

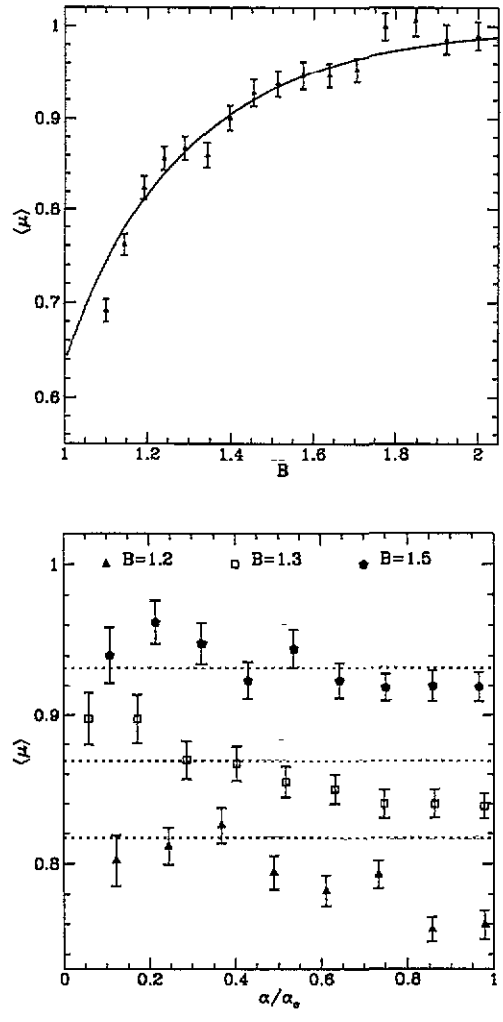

Figure 7. Results of numerical evaluation of μ, using simulated annealing. The top graph shows $\mu(B)$ (at $\alpha = 0.5$), along with the theoretical prediction. The bottom graph is of $\mu(\alpha)$ for $B = 1.2$, 1.3, 1.5 (from below), which theory predicts to be independent of α as $N \to \infty$.

be constructed before their synaptic weights can be compared, and for system sizes such that finite-size effects can be considered small. The most reliable method of seeking the optimal binary perceptron is laborious enumeration of all 2^N states of the N synapses in search of that configuration that maximises κ for the imposed loading, α. For system sizes where this is feasible, typically $N < 20$, the skeletal model near $B = 1$ (where the peculiarities of f are most prevalent) is prone to resemble the binary perceptron itself simply due to finite-size effects or other largely practical difficulties. It is therefore desirable to be able to train both systems for rather larger numbers of neurons.

Given the attractions of genetic algorithms in optimizing the binary perceptron owing to the dispersion of good networks over the corners of their hypercubic weight space, we have employed such a scheme based on Köhler's algorithm (1990), which is itself seen to give performance in close agreement with the theory of Krauth and Mézard (1989) at least up to $N \sim 50$, where training by exact enumeration would be totally impractical. We

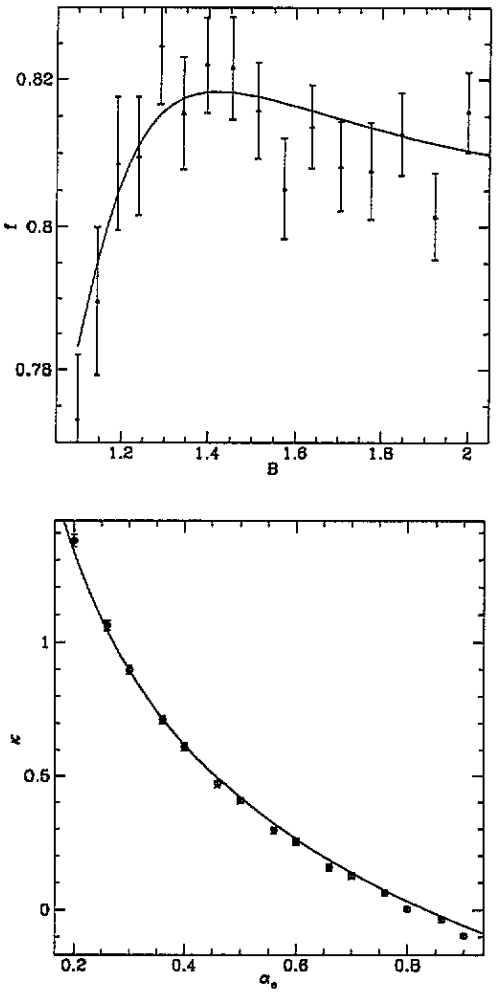

Figure 8. The top graph shows simulation results for the fraction of weights in the binary perceptron correctly predicted by clipping those of the skeletal model f for $\alpha = 0.8$ and $N = 50$. In the lower graph the performance of the genetic algorithm used to train the binary model is indicated by comparing the resulting stability–capacity data points with the theory of Krauth and Mézard.

have thus compared the synapses of the skeletal model with those of a near-optimal binary perceptron when both networks store identical patterns, obtaining results shown in figure 8. (As an indication of the fidelity of the genetic algorithm, the stability κ it achieves for a given storage ratio α is also depicted.) A definite non-monotonic trend is observed, and one that shows fair agreement with the theoretical predictions, although fluctuations appear to be still more important than in the measurements of μ.

7. Conclusions

A perceptron model intermediate between the spherical model, in which the synaptic weights are limited only by a normalization constraint, and a binary-synapse system has been

introduced. Investigation of this system functioning as a heteroassociative memory has suggested that the disconnectedness of the binary model's weight space is central to the differences between this and the canonical spherical model, with a mere preference for synaptic weights near $J = \pm 1$ being very much a secondary consideration. It has been indicated that the skeletal model may closely resemble the analogous binary perceptron even while its weight space remains continuous and connected, but that this exterior similarity belies rather different internal configurations of their synapses. It would therefore appear that whilst the skeletal model may be useful in realizing a functional binary-synapse network, it does not greatly assist the search for the true maximally stable binary perceptron as compared with what can be accomplished using the spherical model. A simple form of simulated annealing has been shown to achieve some success with the task of training the skeletal model.

The results of these investigations would suggest that binary networks derived by training entirely within a continuous weight space (cf Pérez Vicente *et al* 1992) may differ in their properties from those constructed by optimizing in a discrete space, perhaps after pre-training in a continuous space (cf Penney and Sherrington 1993b). Despite this, these differences are likely to be strongly dependent on the practicalities of the training procedures employed, and are therefore not readily closely specified *a priori*.

That a viable binary perceptron may be constructed by clipping the synapses of a skeletal model with $B \simeq 1$ (as suggested by the forms of $p(J)$ (4.11) and $\rho_{\mathrm{clp}}(\Lambda)$ (5.2)) might encourage the development of more efficient algorithms for training the skeletal model than the simplistic scheme used in this work, as it is possible that in some applications of binary-synapse networks, the imperfections of the clipped skeletal model would be excusable. When this is not the case, it would seem that this model is most readily cannibalised into the binary perceptron when $(B - 1)$ is finite. In view of the the investigations of Penney and Sherrington (1993b), it would seem likely that any adjustment to the clipped model needed to improve its performance can be reliably targeted using information provided by the continuous-synapse parent, in the form of the magnitudes of its synaptic weights. Even if exact enumeration of all states of a fraction synapses in a binary network was required, by pre-training with a skeletal model a large fraction of these weights could be predicted with comparative certainty. Thus the combination of pre-training using a skeletal model and either exact enumeration or some form of genetic optimization, might allow rather larger binary networks to be well trained than could be achieved by other means, with a great reduction of the learning task having been achieved by optimization in a continuous space. However, whether the skeletal model is to be preferred to the spherical model as a starting point of such an enumeration is crucially dependent on the efficiency with which the skeletal model may be trained, and on the system size.

Acknowledgments

We would like to thank Ole Winther for the suggestion of the relevance of an inhomogeneous continuous weight space to binary-synapse perceptrons, which was inspirational for this work. The financial support of the SERC (under grant number 9130068X) is gratefully acknowledged. RWP would also like to thank Jesus College, Oxford, for its kindly award of a scholarship.

References

Amaldi E and Nicolis S 1989 Stability–capacity diagram of a neural network with Ising bonds *J. Physique* **50** 2333

Anlauf J K and Biehl M 1989 The AdaTron: an adaptive perceptron algorithm *Europhys. Lett.* **10** 687

Bouten M, Engel A, Komoda A and Serneels R 1990 Quenched versus annealed dilution in neural networks *J. Phys. A: Math. Gen.* **23** 4643

Fletcher R 1987 *Practical Methods of Optimization* (New York: Wiley)

Gardner E J 1987 Maximum storage capacity in neural networks *Europhys. Lett.* 4 481

—— 1988 The space of interactions in neural network models *J. Phys. A: Math. Gen.* **21** 257

Gardner E J and Derrida B 1988 Optimal storage properties of neural network models *J. Phys. A: Math. Gen.* **21** 271

Gutfreund H and Stein Y 1990 Capacity of neural networks with discrete synaptic couplings *J. Phys. A: Math. Gen.* **23** 2613

Horner H 1992 Dynamics of learning for the binary perceptron problem Z. *Phys.* B **86** 291

Kirkpatrick S, Gelatt C D and Vecchi M P 1983 Optimization by simulated annealing *Science* **220** 671

Köhler H 1990 Adaptive genetic algorithm for the binary perceptron problem *J. Phys. A: Math. Gen.* **23** 1265

Krauth W and Mézard M 1987 Learning algorithms with optimal stability in neural networks *J. Phys. A: Math. Gen.* **20** L745

—— 1989 Storage capacity of memory with binary couplings *J. Physique* **50** 3057

Krauth W and Opper M 1989 Critical storage capacity of the $J = \pm 1$ neural network *J. Phys. A: Math. Gen.* **22** L519

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A and Teller E 1953 Equation of state calculations by fast computing machines *J. Chem. Phys.* **21** 1087

Minsky M L and Papert S A 1969 *Perceptrons: An introduction to computational geometry* (Cambridge, MA: MIT Press)

O'Kane D and Monassen R 1993 *Private communication*

Penney R W and Sherrington D 1993a Noise-optimal binary-synapse neural networks *J. Phys. A: Math. Gen.* **26** 3995

—— 1993b The weight space of the binary perceptron *J. Phys. A: Math. Gen.* **26** 6173

Pérez Vicente C J, Carrabina J and Valderrana E 1992 Study of a learning algorithm for neural networks with discrete synaptic weights *Network* **3** 165

Wong K Y M, Rau A and Sherrington D 1992 Weight space organisation in optimized neural networks *Europhys. Lett.* **19** 559