The weight-space of the binary perceptron

1993 J. Phys. A: Math. Gen. 26 6173

(http://iopscience.iop.org/0305-4470/26/22/018)

J. Phys. A: Math. Gen. 26 (1993) 6173–6185. Printed in the UK

The weight-space of the binary perceptron

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

Received 22 July 1993

Abstract. With a view to finding features of the weight-space of the binary perceptron that might be instructive for training binary-synapse neural networks, the maximally-stable perceptron having binary-valued weights is compared with continuous-weight perceptrons, for universal choices of stored patterns. The fraction of synaptic-weights correctly predicted by clipping the synapses of the continuous network is calculated in the thermodynamic limit and compared with simulation results for smaller systems. Numerical experiments show good agreement with theory but, in addition, indicate that those binary synapses likely to be wrongly predicted by weight-clipping are predominantly those which are weakest in the continuous-synapse perceptron. Although not rescuing training time from growing exponentially in the system size, our results suggest ways of significantly accelerating the search for successful, albeit possibly imperfect, neural networks with discrete-valued couplings.

1. Introduction

Although, generically, neural networks have many attractive properties, such as an ability to learn behaviour from examples and to function as associative memories (see e.g., Müller and Reinhardt 1990), unsurprisingly, their actual implementation is generally not facile. However, the species of network, and particularly the nature of its synapses, often greatly influences the ease with which such a network may be trained to fulfil a particular task. Networks whose synaptic weights may take a continuum of values are typically far easier to train than those with discrete-valued synapses, in spite of the practical attractions of the latter in terms of simplicity and robustness. Training procedures for continuous-synapse systems may profitably invoke gradient-descent methods; back-propagation (Rumelhart *et al* 1986) being an extreme example or the AdaTron algorithm (Anlauf and Biehl 1989) is a more sophisticated scheme. For discrete-synapse networks such an approach is clearly inadmissible, and even simulated annealing is apt to fail to find good choices of weights in finite time (Horner 1992).

In order to try to identify general features of binary-synapse networks, particularly relative to similar networks with continuous synapses, we have examined the prototypical task of associative memorization of random patterns by a network of $N+1$ formal neurons, $S_i \in \{\pm 1\}$, $i \in \{0 \ldots N\}$. The neurons are linked by synapses, J_{ij}, and have a discrete-time dynamics $S_i(t+1) = \mathrm{sgn}(N^{-1/2}\sum_j J_{ij}S_j(t))$. A set of αN binary patterns, $\xi_i^\mu \in \{\pm 1\}$, $\mu \in \{1 \ldots \alpha N\}$, are intended to be attractors of these dynamics, representing the memories, or concepts, learned. Requiring that these patterns are fixed points of the neuronal dynamics, thereby demanding that the aligning fields, $\Lambda_i^\mu = N^{-1/2}\sum_j \xi_i^\mu J_{ij}\xi_j^\mu$, are all positive, generally leads to finite basins of attractions for the patterns only if the Λs exceed a non-zero positive threshold, κ. Imposing such requirements allows attention to be confined to synapses feeding only a single neuron, which reduces the training problem to

0305-4470/93/226173+13$07.50

a perceptron architecture (i.e. a single output neuron connected to N input units). Given that training a binary perceptron with noiseless data, corresponding to the maximally stable rule ($\Lambda_i^\mu > \kappa > 0\ \forall i, \mu$), generally appears to be more demanding than learning optimal behaviour for noise-corrupted information (see Penney and Sherrington 1993), it is the former task that will be addressed here. The practically more interesting problem of networks learning a rule from examples (see Watkin *et al* 1993 for a recent review) is expected to suffer from similar impediments to the idealized problem which we shall be considering. Further, we will assume the most extreme form of discrete synapses, namely those limited to values ± 1 only.

A number of training schemes for binary-synapse networks have been proposed. Amongst these, the simplest is the clipped Hebb rule (van Hemmen 1987), genetic algorithms have shown some value (Köhler 1990, Fontanari and Meir (1991) have even used a genetic algorithm to evolve an iterative learning algorithm), and corruption of a network with continuous weights has been suggested (Pérez Vicente *et al* 1992, Penney and Sherrington 1993). In this paper we will re-examine the latter method and explore how much information about successful binary-synapse networks might be inferred from a system with real-valued synapses, whose direct construction is likely to be far more readily accomplished than the system of ultimate interest.

For a specific choice of patterns, ξ_i^μ, we imagine a network with continuous-valued weights ($J_{ij}^{\text{cts}} \in \mathbb{R}$) to be trained according to one of three popular schemes: the maximally stable, pseudo-inverse and Hebb rules. Algorithms that produce, or rapidly approach, realizations of these rules are well known, but we are not aware of any analogous methods applicable to binary synapses, so it would be interesting to know how useful the existing algorithms might be in training binary-synapse networks. Given that the most obvious method of reducing a continuous synapse to a binary-valued one (J_{ij}^{bin}) would seem to be clipping the former, producing $J_{ij}^{\text{bin}} = \text{sgn}(J_{ij}^{\text{cts}})$, it would be instructive to know what proportion of synapses could be correctly predicted by such a method and, moreover, whether those synapses liable to be incorrectly predicted could be simply identified, thereafter to be subjected to attention. Below, we will first address theoretical predictions of the fraction of synapses correctly predicted by weight-clipping in large perceptrons, and thereafter discuss simulations which consider the second question in addition to corroborating the theory.

2. The mutual overlap of binary and spherical networks

For a given choice of patterns and training rules, the fraction of binary-valued synapses (feeding a particular neuron, i) that can be correctly deduced by weight-clipping the companion continuous network is given by

$$f = \frac{1}{2}\left\{1 + \frac{1}{N}\sum_j J_{ij}^{\text{bin}}\,\text{sgn}(J_{ij}^{\text{cts}})\right\} \tag{2.1}$$

in which J_{ij}^{bin} and J_{ij}^{cts} represent networks successfully trained on the patterns, according to their respective rules. If the networks concerned are large ($N \gg 1$), then it is expected that it is not the precise details of the patterns, ξ_j^μ, that are significant, rather their stochastic properties, such as $\langle \xi_j^\mu \rangle_\xi$ and $\langle \xi_j^\mu \xi_k^\nu \rangle_\xi$. In acknowledging the possibility that the training procedures need not define networks uniquely (an effect accentuated by taking N large), it would seem reasonable to consider averaging f over choices of good networks of each

species, and over the selection of patterns (which should be a benign operation). It is such an average quantity that we have calculated, and on which we focus. The patterns are taken to be independent, unbiased random variables of unit variance; $\langle \xi_j^\mu \rangle_\xi = 0$, $\langle \xi_j^\mu \xi_k^\nu \rangle_\xi = \delta_{jk}\delta^{\mu\nu}$. These cumulants are common to the distributions $p(\xi) = \frac{1}{2}(\delta(\xi-1)+\delta(\xi+1))$ (true binary patterns) and $p(\xi) = \exp(-\frac{1}{2}\xi^2)/\sqrt{2\pi}$ ('Gaussian' patterns), a point which has implications for numerical simulations of the theory.

Purely as an analytical device, it is usual in discussing large continuous-synapse networks, following Gardner (1988), to impose a normalization constraint on these synapses, so that their weight-space becomes spherical, and thus has finite volume. The continuous-valued weights meeting such a constraint will be denoted W_{ij}, and choosing the norm such that $\sum_j W_{ij}^2 = N\ \forall i$ ensures that the hyper-cubic weight-space of the binary network is a subset of the spherical space. Hereafter the continuous network will be referred to as 'spherical', and the labels 'J_{ij}' will be reserved for the binary synapses.

By applying replica mean-field theory to the joint weight-space of the two types of network, f and a number of other instructive quantities may be calculated. The calculation, sketched in the appendix, gives expressions for the following objects (in the limit $N \to \infty$, and after suppressing the output index, i):

$$p = \frac{1}{N}\sum_j \overline{J_j \operatorname{sgn}(W_j)} \qquad \text{and} \qquad s = \frac{1}{N}\sum_j \overline{J_j}\,\overline{W_j} \tag{2.2}$$

which reflect the similarities between the two classes of network that we have targeted,

$$q = \frac{1}{N}\sum_j \overline{J_j}\,\overline{J_j} \qquad \text{and} \qquad Q = \frac{1}{N}\sum_j \overline{W_j}\,\overline{W_j} \tag{2.3}$$

indicating the dispersal of networks over the two weight-spaces, and

$$l = \frac{1}{N}\sum_j \overline{W_j}\,\overline{\operatorname{sgn}(W_j)} \qquad m = \frac{1}{N}\sum_j \overline{|W_j|} \qquad \text{and} \qquad r = \frac{1}{N}\sum_j \overline{\operatorname{sgn}(W_j)}\,\overline{\operatorname{sgn}(W_j)} \tag{2.4}$$

giving more information about the spherical network. ($\overline{h}$ denotes a weighted average of h over all choices of the optimal networks of each species.) The order parameters q and Q are familiar from earlier works (Gardner 1988, Krauth and Mézard 1989), and are central to the other conditions determining p, s, l, m and r. The fraction of binary synapses directly deducible by clipping the continuous synapses is given by $f = \frac{1}{2}(1+p)$. The detailed expressions determining these order parameters are unilluminating, and are relegated to the appendix.

For each of three classes of spherical network, we will explore how f varies with the pattern loading, α, for $\alpha < 0.83$, this being the capacity limit of the maximally-stable binary perceptron (Krauth and Mézard 1989). The three learning rules considered will be as follows.

(i) The maximally stable network (MSN), for which $\Lambda_i^{\mu,\mathrm{sph}} > \kappa^{\mathrm{sph}} > 0$ where κ^{sph} is chosen to be as large as possible for the loading, α, subject to a spherical constraint on the synapses. There is no requirement $\kappa^{\mathrm{sph}} = \kappa^{\mathrm{bin}}$ (where κ^{bin} is the analogous greatest lower bound on the stabilities achievable with binary synapses); these thresholds are independently determined by the pattern loading being such as to saturate each network.

(ii) The pseudo-inverse network (PIN) characterized by all patterns having the same stability: $\Lambda_i^{\mu,\mathrm{sph}} = \kappa^{\mathrm{sph}}$. Again, there is no implication that $\kappa^{\mathrm{sph}} = \kappa^{\mathrm{bin}}$ or, indeed, that $\kappa_{\mathrm{MSN}} = \kappa_{\mathrm{PIN}}$.

(iii) The Hebb rule has a Gaussian distribution of pattern stabilities centred on $\Lambda^{\mathrm{sph}} = \alpha^{-1/2}$.

Each of these learning rules allow the quantity f to be determined readily, using appropriate performance measures given in the appendix; profiles of $f(\alpha)$ are shown in figure 1. It is seen that, although the spherical MSN correctly predicts the largest fraction of binary synapses compared with the other rules, simple weight-clipping does not produce near-optimal binary networks unless the loading, α, is small. However, it is noteworthy that upwards of 80% of synapses could be correctly determined using a spherical MSN. The pattern stability field distribution for the clipped spherical MSN ($\rho_{\mathrm{cM}}(\Lambda)$) was given in Penney and Sherrington 1993 (equation (5.1)), and using this one may calculate the fraction of patterns that are unstably stored by the derived binary model in the absence of any further training;

$$\gamma = \int_{-\infty}^{0} \rho_{\mathrm{cM}}(\Lambda)\,\mathrm{d}\Lambda \sim \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{2}{\alpha(\pi-2)}\right)\sqrt{\frac{\alpha(\pi-2)}{2}} \qquad \text{as } \alpha \to 0. \tag{2.5}$$

For $\alpha = 0.8$ $\gamma = 0.17$, but by $\alpha = 0.4$ the fraction of unstable patterns has fallen to 5%.

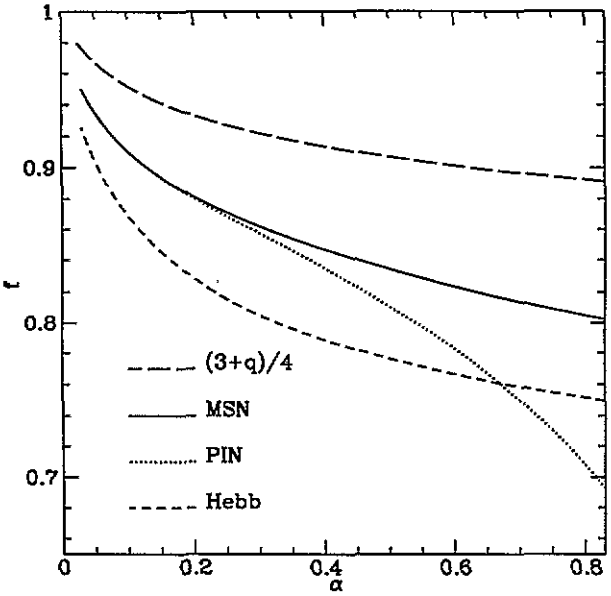

Figure 1. The fraction of binary synapses (f) in an MSN that are correctly predicted by weight-clipping three types of spherical networks (maximally stable, pseudo-inverse and Hebb), compared with the upper bound $\frac{1}{4}(3+q)$, as a function of pattern loading, α.

The distribution of optimal binary networks over their own weight-space itself provides an upper bound on the average number of synapses that could be correctly predicted by any algorithm that produces a single continuous-synapse network. That there are optimal

binary perceptrons that differ amongst themselves in a fraction of $\frac{1}{2}(1-q)$ synapses implies that an arbitrary binary network must differ in at least a fraction $\frac{1}{4}(1-q)$ of synapses from some optimal networks. Therefore, on averaging over all optimal networks, the number of bits correctly predictable by a single network is bounded above by $\frac{1}{4}(3+q)$†. In contrast to the spherical model, on saturation of a discrete-synapse network the order parameter q does not generally reach unity (Krauth and Mézard 1989, Gutfreund and Stein 1990), indicating wide dispersal of these networks; e.g., for $\alpha \simeq 0.8$, $q \simeq 0.6$. Although any algorithm that correctly predicted all the binary synapses of a particular optimal network would be a highly attractive scheme, that other optimal networks may not be close means that its average overlap with all optimal networks would not be so impressive. In the absence of a mechanism (and probably quite a sophisticated one) in the spherical-model learning rule which favours its being useful for training a binary-synapse perceptron by weight-clipping, it is reasonable to assume that the overlaps, f, depicted in figure 1 would represent the overlap of any of the optimal binary networks with the given spherical model. This assertion is lent weight by numerical simulations.

A further question that can be addressed using the order parameters (2.2) and (2.3) is whether the centre of gravity of the optimal binary networks is parallel to that of the optimal spherical networks, i.e. whether $\overline{W_j} = \lambda \overline{J_j}\ \forall j$, for some choice of λ? Equivalently, one could ask whether the minimum of $N^{-1}\sum_j(\overline{W_j} - \lambda\overline{J_j})^2$, with respect to λ, is zero. This quantity is minimized for the choice $\lambda = s/q$, and from the positivity of each term in the summation one has the identity

$$\begin{aligned} &\frac{1}{N}\sum_j\{\overline{W_j} - \overline{J_j}s/q\}^2 \geqslant 0 \\ &\Rightarrow Q - s^2/q \geqslant 0. \end{aligned} \tag{2.6}$$

Thus, whether or not the two species of optimal networks have centres of gravity that coincide may be determined using the order parameters calculated. It would appear that this inequality is satisfied strictly. Thus, mapping all networks into the spherical-model's weight-space, the optimal binary networks are seen to be distributed *asymmetrically* about the domain of optimal spherical models, which is perhaps contrary to intuitive expectations.

3. Insight from numerical experiments

The discreteness of the binary-model's weight-space is a mixed blessing; unlike a continuous space, the possibility of searching all points for the best possible network is open, whilst this discreteness is at the root of the difficulty of constructing algorithms for training binary perceptrons. Learning by exact enumeration of all 2^N states of the binary synapses is seemingly the only method known to produce optimal networks, despite the exponential divergence of training time with system size. In seeking to examine the properties of optimal perceptrons, as addressed by the analysis, we are forced into exact enumeration, and hence are limited to small system sizes relative to those accessible to algorithms such as the AdaTron.

† More rigorously, one may consider bounding $p = N^{-1}\sum_j \overline{J_j}\,\mathrm{sgn}(W_j)$ by the maximum of this quantity with respect to $\{W_j\}$. The supremum occurs for $\mathrm{sgn}(W_j) = \mathrm{sgn}(\overline{J_j})$, but from the identity $N^{-1}\sum_j(\overline{J_j} - \mathrm{sgn}(\overline{J_j}))^2 \geqslant 0$, one obtains $p < \frac{1}{2}(1+q)$, hence $f < \frac{1}{4}(3+q)$.

The theoretical predictions of the previous section suggest that existing learning rules for continuous-synapse networks will always wrongly predict a finite fraction of synapses in a binary network. Given that the lower bound on this fraction is finite, if exact enumeration of the untrustworthy synapses was undertaken, this part of the training process would still scale exponentially with the number of connections. However, if only perhaps 40% of couplings required attention, learning time would scale as $e^{0.4\lambda N}$ rather than $e^{\lambda N}$ for the whole system. For practical applications the increase in accessible system size thus afforded might well be significant.

In performing simulations we have sought to test the hypothesis that those spherical-synapses that are weakest are those that least-reliably predict binary-valued weights. It seems unlikely that a very weak synapse in the spherical model could most often be replaced, in converting it to a binary value, by a stronger synapse having only the same sign. Further, those synapses that are strong in the spherical model are likely to be highly significant in stabilizing patterns, and therefore not sensibly converted to one of opposing sign, as well as being potentially weaker. In order to examine these ideas we have followed the following strategy during the enumeration of the states of the binary synapses:

- For a given, random, choice of patterns, $\{\xi_i^\mu\}$, a spherical-synapse network is trained on these patterns using either Hebb's rule or the AdaTron algorithm. The starting configuration of the binary synapses is derived by clipping these spherical weights. (A random starting configuration is also considered.)
- The synapses are prioritized according to their magnitude in the continuous-synapse network, and relabelled so that $W_j \leqslant W_{j+1}$.
- The stability of the least stable pattern (according to the initial state of the binary perceptron) is noted.
- The 2^1 states of weight J_0 are explored and if either of these two states produces better minimum stability, this stability is noted and the relevant choice of synapses is recorded.
- The 2^2 combinations of J_0 and J_1 are tested with similar procedures.
- $\vdots$
- All 2^N combinations of all synapses are examined in search of the maximum lower bound on pattern stability, again noting the configuration that produces it.

Thus after the full enumeration, the accuracy of the initial weight-configuration is manifested. Further, the evolution of the minimum stability during the enumeration indicates to what extent the use of the spherical network might allow subsequent training to be confined to only selected binary synapses. If the instantaneous best minimum stability rapidly approaches its asymptotic value then enumeration of only a subset of all synapses would seem necessary; a more gradual improvement would signal that states of a larger fraction of synapses would need exploring before the optimal network could be approached, or attained. By continuing enumeration to completion, we may identify the optimal network and compare it with the spherical model.

In affecting such a comparison an important feature of the theoretical approach must be addressed. Given that our analysis relies on the taking of the thermodynamic limit, $N \to \infty$ (formally representing very large systems), and that in this limit only low-order cumulant averages of the pattern elements, ξ_i^μ, have significance, for finite-size systems the higher-order cumulants of the ξ_i^μ will influence the suitability of the theory. As far as the analysis is concerned, the ξ_i^μ behave as Gaussian random variables (hence with only two non-trivial cumulants), even though they strictly can assume only binary values if they are to represent neuronal states. Given the modest number of synapses that can be enumerated without inordinate consumption of computer time, it is unsurprising that theory and simulations do not agree well when true binary patterns are used. However,

in common with Krauth and Opper (1989) we find that the use of Gaussian patterns (for which $p(\xi_i^\mu) = \exp(-\frac{1}{2}\xi_i^{\mu 2})/\sqrt{2\pi}$) brings theory and experiment into far more convincing correspondence in their average behaviours. It is interesting that on average, binary patterns behave more favourably than Gaussian patterns in terms of producing larger overlap with the spherical model. Large fluctuations about the average behaviour beset both forms of pattern however; this is another symptom of the meagre number of synapses.

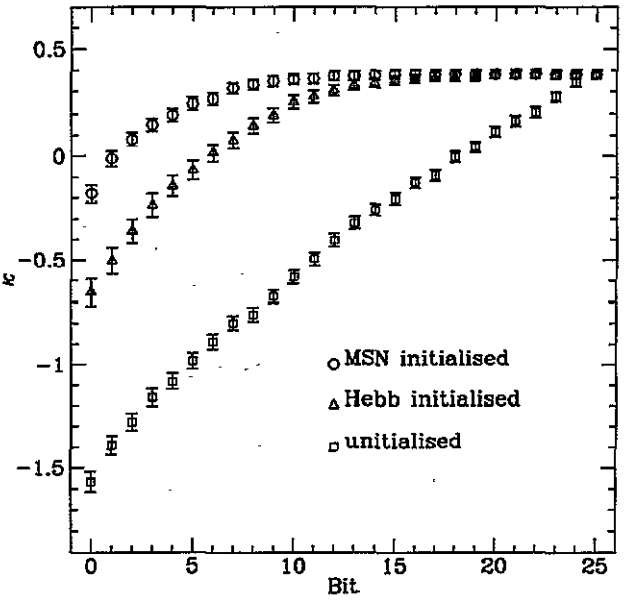

Figure 2. The evolution of the minimum pattern stability found, κ, with the number of synapses enumerated, 'bit'. Only binary patterns are considered, for $N = 25$, $P = 13 \Rightarrow \alpha = 0.52$.

Figure 2 shows the evolution of the maximum lower bound on pattern stability during the enumeration and indicates the clear advantage of pre-training the binary network using its spherical companion. The figure suggests that those synapses that are strongest in the spherical model do appear to be reliable in predicting binary-valued synapses. As a further test of this hypothesis, we also calculate the average magnitude of those spherical weights that correctly predict their binary cousins ($\langle|W_t|\rangle$) relative to the magnitude of the remaining weights that make erroneous predictions ($\langle|W_f|\rangle$). This ratio can be expressed in terms of accessible quantities (2.2), (2.4);

$$\frac{\langle|W_t|\rangle}{\langle|W_f|\rangle} = \frac{m+s}{m-s}\frac{1-f}{f} \tag{3.1}$$

and is depicted in figure 3. Both simulations and theory lend some support to the hypothesis, and again the behaviour for true binary patterns appears more favourable than for Gaussian patterns, which themselves show better correspondence with the theory, as might be expected.

The optimal binary networks identified during the enumeration procedure allow the results of the previous section to be tested; comparison of theory and experiment is given in figure 4.

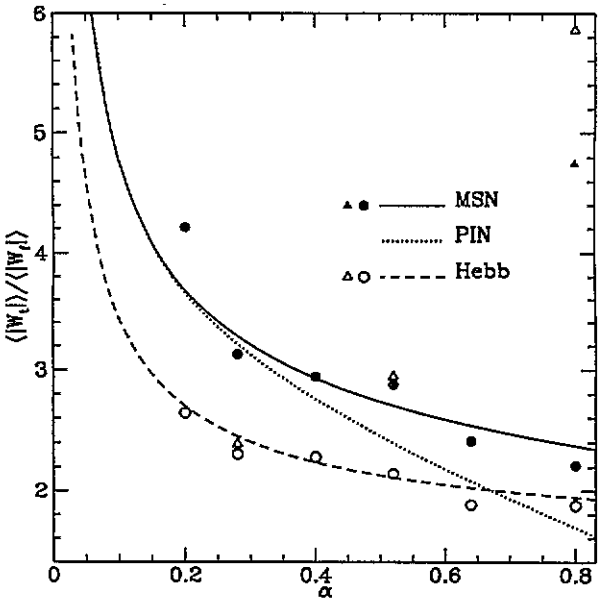

Figure 3. The ratio of the magnitude of synapses in the spherical model which correctly predict the corresponding binary synapses ($\langle|W_i|\rangle$), to those weights which make incorrect predictions ($\langle|W_f|\rangle$). Simulation results for $N = 25$ are shown as circles (Gaussian patterns) and triangles (binary patterns).

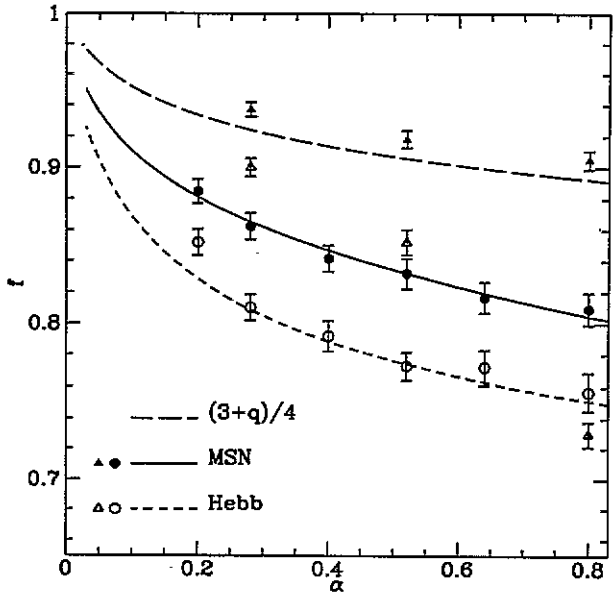

Figure 4. Comparison of the theoretical prediction of the fraction of correctly predicted binary synapses, f, with simulation results for $N = 25$. Both binary patterns (triangles) and Gaussian patterns (circles) are shown, with error bars obtained from an ensemble of patterns. The theoretical curves are as found in figure 1.

In all cases, simulation results represent averages over approximately 100 choices of patterns, and error bars, where shown, are standard deviations of the mean values represented by the plotted points. The standard deviations of the relevant quantities themselves are a factor approximately 10 larger, and that these are generally large for the system sizes examined confirms the strong finite-size effects manifested by the binary perceptron (Amaldi and Nicolis 1989, Krauth and Opper 1989, Derrida *et al* 1991).

4. Conclusion

Rather than examining the weight-space of the binary perceptron in isolation (e.g. Fontanari and Köberle 1990), we have compared this with the weight-space of a companion network with continuous-valued weights, but storing the same patterns as the binary network. Limits on the utility of a network with continuous-synapse values in training one with binary synapses have thus been indicated. Accessible algorithms applicable to the real-valued synapses have been shown to be useful for directly predicting a significant fraction of the discrete synapses (thereby rating the starting point in the training scheme of Pérez Vicente *et al* (1992)) and, moreover, to provide guidance as to which weights are less faithfully predicted. Simulation results have suggested that training by exact enumeration of the binary weights may be facilitated by examining the corresponding real weights, thereby allowing larger system sizes to be trained by this method than without any such information.

Acknowledgments

We would like to thank Ole Winther and Andreas Wendemuth for interesting discussions. RWP is grateful to Jesus College, Oxford, for its generous award of a scholarship. Financial support from the Science and Engineering Research Council (under grant number 9130068X) is appreciated.

Note added in proof. Since completing this work we have learned that similar methods of training the binary perceptron have independently been under investigation by Jacek Iwanski of Limburgs Universitair Centrum.

Appendix

An overview of the weight-space calculation leading to expressions for the quantities in (2.3), (2.4) and (2.2) will be given. Most of the methods of such calculations are widely used, particularly following the work of Elizabeth Gardner (1988).

It is typical of a mean-field theory that the calculation of a system's free energy naturally highlights the order parameters of physical significance, without necessitating prior insight. Given that we will be addressing large, range-free (or infinite-dimensional), but disordered, models, replica mean-field theory is the obvious methodology to employ. Thus, by examining the total free energy of the union of a binary- and spherical-synapse network, we expect to be able to find physically meaningful quantities that reflect the similarities of these two components. Use of a simple trick ensures that the quantity of central interest, namely the fraction of binary synapses correctly predicted by weight-clipping, emerges along with other relevant information.

For each weight-space we impose a cost-function, $E(\{J\}) = \sum_\mu g(\Lambda^\mu)$ with a global annealing temperature, β^{-1}, and examine the entire set of perceptrons according to a Boltzmann weighting, $\exp(-\beta(E^{\text{bin}}+E^{\text{sph}}))$. On taking the temperature to zero, all networks which fail to optimize the cost function are excluded from consideration, leaving features typical of all optimal networks. Assuming the associated free energy to be self-averaging in the stored patterns, we invoke the replica method, and proceed to calculate integer moments of the partition function, subsequently analytically continuing our expressions to small replica number, exploiting the identity $\langle \ln Z\rangle = \lim_{n\to 0}(\langle Z^n\rangle - 1)/n$. Our starting point is

$$\langle Z^n\rangle_\xi = \left\langle \prod_{b=1}^{n} \mathrm{Tr}_{\{J_j^b\}} \exp\left(-\beta \sum_\mu g^{\text{bin}}(\Lambda_i^{\mu,b,bin})\right) \times \int \{\mathrm{d}W_j^b\}\delta\left(\sum_{j=1}^{N} W_j^{b2} - N\right) \exp\left(-\beta \sum_\mu g^{\text{sph}}(\Lambda_i^{\mu,b,\text{sph}})\right)\right\rangle_\xi \tag{A.1}$$

in which $\{J_j^b\}$ are the binary synapses in the replicated system and $\{W_j^b\}$ represent the weights of the replicated spherical model. The aligning fields of the patterns are defined as follows

$$\Lambda_i^{\mu,b,\text{bin}} = \frac{1}{\sqrt{N}} \sum_j \xi_i^\mu J_j^b \xi_j^\mu \qquad \Lambda_i^{\mu,b,\text{sph}} = \frac{1}{\sqrt{N}} \sum_j \xi_i^\mu W_j^b \xi_j^\mu. \tag{A.2}$$

Following Gardner (1988) we introduce Fourier representations of unity in order to facilitate the pattern average by extracting the ξ_j^μ from within the cost functions:

$$1 = \sum_{y_\mu^b=-N}^{N} \delta_{Kr}\left(y_\mu^b - \sum_j \xi_i^\mu J_j^b \xi_j^\mu\right) = \int \mathrm{d}v_\mu^b\, \delta\left(v_\mu^b - \tfrac{1}{\sqrt{N}} \sum_j \xi_i^\mu W_j^b \xi_j^\mu\right). \tag{A.3}$$

By arbitrarily introducing another similar identity,

$$1 = \sum_{t_\mu^b=-N}^{N} \delta_{Kr}\left(t_\mu^b - \sum_j \xi_i^\mu \,\mathrm{sgn}(W_j^b)\xi_j^\mu\right) \tag{A.4}$$

the eventual emergence of the quantity $f = \frac{1}{2}\{1 + N^{-1}\sum_j \langle J_j \overline{W_j}\rangle_\xi\}$ is assured, where $\overline{}$ denotes a Boltzmann-weighted average. Performing the pattern average, invoking $\langle \xi_j^\mu\rangle = 0$ and $\langle \xi_j^\mu \xi_k^\nu\rangle = \delta_{jk}\delta^{\mu\nu}$, produces a factor of the form

$$\prod_\mu \exp\left\{-\frac{1}{2}\sum_j \left(\sum_b z_\mu^b J_j^b + x_\mu^b W_j^b/\sqrt{N} + u_\mu^b \,\mathrm{sgn}(W_j^b)\right)^2\right\} \tag{A.5}$$

which, on expanding the square, produces precursors of the quantities q, Q, l, m, r, s and p. (Contributions from higher-order cumulants vanish in the large-N limit.) The limit $N \to \infty$ allows the summations over t_μ^b and y_μ^b to be converted into integrations, with an associated rescaling of u_μ^b and z_μ^b. Further partitions of unity are introduced *à la Gardner*, and in the limit $N \to \infty$, these various integrations may be evaluated using the method of steepest descents. At the saddle point, we assume replica-symmetric values of all order

parameters, an assumption that seems justified for the ranges of α of interest (Gardner 1988, Krauth and Mézard 1989). By this stage the thermodynamics depend on the extremization, in the limit $n \to 0$, of a cumbersome free-energy functional;

$$G(\varepsilon, m, \hat{m}, q, \hat{q}, Q, \hat{Q}, r, \hat{r}, l, \hat{l}, s, \hat{s}, p, \hat{p})$$
$$= (n(\varepsilon - m\hat{m}) + \tfrac{1}{2}n(1-n)(q\hat{q} + Q\hat{Q} + r\hat{r}) + n(n-1)l\hat{l} - n^2(s\hat{s} + p\hat{p})$$
$$+ \alpha G_0(m, q, Q, r, l, s, p) + G_1(\varepsilon, \hat{m}, \hat{q}, \hat{Q}, \hat{r}, \hat{l}, \hat{s}, \hat{p})) \tag{A.6}$$

with respect to all parameters, and in which

$$G_0 = \ln\left\{ \prod_b \int \mathrm{d}y^b \frac{\mathrm{d}z^b}{2\pi} \exp(\mathrm{i}y^b z^b - \tfrac{1}{2}(z^b)^2 - \beta g^{\mathrm{bin}}(y^b)) \right.$$
$$\times \int \mathrm{d}v^b \frac{\mathrm{d}x^b}{2\pi} \exp(\mathrm{i}v^b x^b - \tfrac{1}{2}(x^b)^2 - \beta g^{\mathrm{sph}}(v^b)) \int \mathrm{d}t^b \frac{\mathrm{d}u^b}{2\pi} \exp(\mathrm{i}t^b u^b - \tfrac{1}{2}(u^b)^2)$$
$$\times \exp\left(-\sum_b x^b u^b m - \sum_{b<c} (z^b z^c q + x^b x^c Q + u^b u^c r) \right)$$
$$\left. \times \exp\left(-\sum_{b \neq c} x^b u^c l - \sum_{bc} (z^b x^c s + z^b u^c p) \right) \right\} \tag{A.7}$$

and

$$G_1 = \ln\left\{ \prod_b \mathrm{Tr}_{J^b} \int \mathrm{d}W^b \exp(-\varepsilon W^{b2} + \hat{m}|W^b|) \right.$$
$$\times \exp\left(\sum_{b<c} (J^b J^c \hat{q} + W^b W^c \hat{Q} + \mathrm{sgn}(W^b)\,\mathrm{sgn}(W^c)\hat{r}) \right)$$
$$\left. \times \exp\left(\sum_{b \neq c} W^b \,\mathrm{sgn}(W^c)\hat{l} + \sum_{bc} (J^b W^c \hat{s} + J^b \,\mathrm{sgn}(W^c)\hat{p}) \right) \right\}. \tag{A.8}$$

Given that the order parameters s, $\hat{s}$, p and $\hat{p}$ enter (A.6) only at order n^2, it is necessary to be careful to retain terms up to this order while performing the limit $n \to 0$; usually neural network or spin-glass problems require only the term in n^1 of their free-energy functionals. (In a similar study Wong *et al* (1992) follow an alternative strategy.) On simplifying G_0 one finds that this function is independent of m, r, l and p, which means that their conjugate parameters, $\hat{m}$, $\hat{r}$, $\hat{l}$ and $\hat{p}$, vanish at the saddle point. This observation allows G_1 to be simplified as only terms up to first order in these conjugate variables are needed. Thus it emerges that the joint free energy $(-(N\beta)^{-1} \lim_{n\to 0}(Z^n - 1)/n)$ is just the sum of the free energies of the two components, as would be expected. These constituents themselves are determined by extrema of two functionals familiar from earlier works:

$$n^{-1}G^{\mathrm{sph}} = \varepsilon + \tfrac{1}{2}Q\hat{Q} + \frac{\hat{Q}}{2(\hat{Q} + 2\varepsilon)} - \tfrac{1}{2}\ln(\hat{Q} + 2\varepsilon)$$
$$+ \alpha \int \mathrm{D}x \ln\left[\int \mathrm{D}y \exp\left(-\beta g^{\mathrm{sph}}\left(y\sqrt{1-Q} - x\sqrt{Q} \right) \right) \right] \tag{A.9}$$

(Gardner 1988) and

$$n^{-1}G^{\text{bin}} = \tfrac{1}{2}\hat{q}(1-q) + \int \mathrm{D}x \ln(2\cosh x\sqrt{\hat{q}}) + \alpha \int \mathrm{D}x \ln\left[\int \mathrm{D}y \exp\left(-\beta g^{\text{bin}}\left(y\sqrt{1-q} - x\sqrt{q}\right)\right)\right] \tag{A.10}$$

(Krauth and Mézard 1989), in which the shorthand, $\mathrm{D}x = \mathrm{d}x \exp(-\frac{1}{2}x^2)/\sqrt{2\pi}$, has been used. Thereafter, the remaining order parameters are given as follows

$$l = Q\sqrt{\frac{2}{\pi}} \qquad m = \sqrt{\frac{2}{\pi}} \qquad r = \int \mathrm{D}x \left\{2\int_0^{\sqrt{Q/(1-Q)}} \mathrm{D}y\right\}^2 \tag{A.11}$$

$$s = \frac{\hat{s}}{\hat{Q} + 2\varepsilon} \int \mathrm{D}x \operatorname{sech}^2\left(x\sqrt{\hat{q}}\right) \tag{A.12}$$

with

$$\hat{s} = \alpha \frac{\partial}{\partial s} \int \mathrm{D}x \, \mathrm{D}k \ln\left[\int \mathrm{D}u \exp\left(-\beta g^{\text{bin}}\left(u\sqrt{1-q} - x\sqrt{q}\right)\right)\right] \times \ln\left[\int \mathrm{D}v \exp\left(-\beta g^{\text{sph}}\left(v\sqrt{1-Q} + k\sqrt{Q - s^2/q} - sx/\sqrt{q}\right)\right)\right] \tag{A.13}$$

and

$$p = \int \mathrm{D}y \, \mathrm{D}k \tanh\left(k\sqrt{\hat{q} - \hat{s}^2/\hat{Q}} + y\hat{s}/\sqrt{\hat{Q}}\right) \cdot 2\int_0^{y\sqrt{Q/(1-Q)}} \mathrm{D}x. \tag{A.14}$$

So far, we have not specified the forms of the cost functions g^{bin} and g^{sph}. Restricting ourselves to the binary MSN, we will always have $g^{\text{bin}}(\Lambda) = \theta(\kappa^{\text{bin}} - \Lambda)$. On taking the zero-temperature limit, $\beta \to \infty$, this requires that all patterns have a stability at least as large as κ^{bin}. For the spherical model, three learning rules may be considered readily within the formalism developed:

(i) taking $g^{\text{sph}}(\Lambda) = \theta(\kappa^{\text{sph}} - \Lambda)$ leads to a spherical maximally stable network;

(ii) the choice $g^{\text{sph}}(\Lambda) = \frac{1}{2}(\Lambda - \kappa^{\text{sph}})^2$ produces a pseudo-inverse network, for which all patterns have the same stability;

(iii) using $g^{\text{sph}}(\Lambda) = \Lambda$ gives rise to a Hebbian spherical model, equivalent to the prescription $W_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu / \sqrt{\alpha N}$.

On selecting forms for the cost functions, the limit $\beta \to \infty$ is taken (thereby eliminating non-optimal networks). The spherical model can be saturated by taking the limit $Q \to 1$, in which limit $\hat{Q}$ diverges, with $\hat{s}$ also diverging such that the ratio $\hat{s}^2/\hat{Q}$ remains finite. The binary model is saturated using the zero-entropy condition ($\lim_{\beta\to\infty} n^{-1}G^{\text{bin}}_{\text{extr}}(\alpha, \kappa, \beta) = 0$), following Krauth and Mézard (1989). It is assumed in taking the zero-temperature limit that the quantity $Q - s^2/q$ remains finite. The remaining order parameters, s, $\hat{s}^2/\hat{Q}$ and p, follow from straightforward numerical solution of the saddle-point conditions, and the values found are consistent with the assumption $Q - s^2/q > 0$. The significance of this observation is discussed in the main text.

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

Derrida B, Griffiths R B and Prügel-Bennett A 1991 Finite-size effects and bounds for perceptron models *J. Phys. A: Math. Gen.* **24** 4907

Fontanari J F and Köberle R 1990 Landscape statistics of the binary perceptron *J. Physique* **51** 1403

Fontanari J F and Meir R 1991 Evolving a learning algorithm for the binary perceptron *Network* **2** 353

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

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

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

Krauth W and Mézard M 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

Müller B and Reinhardt J 1990 *Neural Networks, An Introduction* (Berlin: Springer)

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

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

Rumelhart D E, Hinton G E and Williams R J 1986 Learning representations by back-propagating errors *Nature* **323** 533

van Hemmen J L 1987 Nonlinear neural networks near saturation *Phys. Rev.* A **36** 1959

Watkin T L H, Rau A and Biehl M 1993 The statistical mechanics of learning a rule *Rev. Mod. Phys.* **65** 499

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