The interaction space of neural networks with sign-constrained synapses

1989 J. Phys. A: Math. Gen. 22 4687

(http://iopscience.iop.org/0305-4470/22/21/030)

J. Phys. A: Math. Gen. **22** (1989) 4687-4693. Printed in the UK

The interaction space of neural networks with sign-constrained synapses

Daniel J Amit†‡, C Campbell§ and K Y M Wong†

† Department of Physics, Imperial College, London SW7 2BZ, UK

§ Department of Applied Physics, Kingston Polytechnic, Kingston-on-Thames KT1 3EE, UK

Received 24 May 1989

Abstract. We investigate the optimal storage capacity of attractor neural networks with sign-constrained weights, which are prescribed *a priori*. The storage capacity is calculated by considering the fractional volume of weights which can store a set of random patterns as attractors, for a given stability parameter. It is found that this volume is independent of the particular distribution of signs (gauge invariance) and that the storage capacity of such constrained networks is exactly one half that of the unconstrained network with the corresponding value of the stability parameter.

1. Introduction

This paper is a sequel to a previous study [1] in which it was shown that a perceptron-like learning algorithm can be defined for a situation in which the perceptron weights have a fixed set of signs. That algorithm was then shown to converge, provided that a solution for a somewhat stronger set of inequalities exists. Since in a neural network the inputs to every neuron depend, in general, on a different, i.e. independent, set of synaptic efficacies, the learning theorem for one perceptron implies learning for the network. See, e.g., [2]. The same holds for the existence of the required solution.

To be specific, for a set of p patterns $\xi_i^\mu (= \pm 1)$ $(i = 1, \ldots, N, \mu = 1, \ldots, p)$ to be stored in a network of N neurons, the interaction matrix, J_{ij}, must be such as to satisfy the inequalities

$$\xi_i^\mu \sum_{j \neq i}^{N} J_{ij} \xi_j^\mu > 0 \tag{1}$$

at each neuron i for every pattern μ. The same patterns enter at every site j, but different sites i are affected by different sets of couplings. Hence, it is sufficient to discuss the existence of a solution of the set of inequalities at a single site, for the full set of patterns. This, of course, is the perceptron problem [3], which has been formulated in terms of a set of weights A_i and a set of p normalised vectors $\phi_i^\mu (= \pm N^{-1/2})$ [1]. In terms of these variables one is searching for a solution of the p

‡ On leave from the Racah Institute of Physics, Hebrew University, Jerusalem, Israel.

0305-4470/89/214687+07$02.50

linear inequalities

$$\sum_{i=1}^{N} A_i \phi_i^\mu > 0. \tag{2}$$

The convergence of the perceptron learning algorithm [3], and of its modified versions in [1], is ensured if there exists a normalised set of coefficients A_i^* which satisfies the stronger inequalities

$$\sum_{i=1}^{N} A_i^* \phi_i^\mu > \delta > 0. \tag{3}$$

The parameter δ, on the right-hand side of (3), plays the role of a stability (or basin of attraction) parameter in a neural network [2, 4, 5]. Thus, the determination of the existence of a solution to inequalities (3) for a given value of δ provides a double message. First, it ensures that the algorithm which depended on it converges, albeit to a solution of the weaker inequalities (2). Second, it establishes the maximal number of random patterns that can be stored with this particular stability parameter. To reach a solution of the more strongly imprinted patterns by the perceptron algorithm would demand, of course, the existence of a solution satisfying a yet stronger set of inequalities [2].

In [1] we have shown that one could devise algorithms for weight correction which respect a prescribed distribution of signs on the weights. Such algorithms have been shown to converge under the same formal assumptions as were required for the free perceptron. In other words, suppose that a set of signs $g_i = \pm 1$ is prescribed; then the target is to learn a set of weights A_i which satisfy inequalities (2) but with the constraint $A_i g_i > 0$ satisfied at every stage of the learning process. The necessary condition has been shown to be the existence of a solution A_i^* which satisfies (3), with $A_i g_i > 0$, for all i.

Following the ideas of Gardner [2], we proceed to compute the relative volume in interaction space in which both the inequalities (3) and the sign restrictions are satisfied. We find that

(a) the volume is independent of the particular realisation of the g_i; this is the local gauge invariance discussed in [1];

(b) the number of random patterns which can be stored with a given stability parameter K—the minimal magnitude of the local field at a site—by an optimal choice of the sign constrained weights is always one half that of the number which can be stored in the unconstrained network†. This implies in particular, that the learning algorithm of [1] will converge for any set of signs g_i with a stability parameter $K - \delta$, for any loading level $\alpha = p/N$, up to one half of the level for which the learning in the unconstrained network converges.

2. The computation of the volume and the saddle-point equations

Following Gardner [2], we proceed to compute the relative volume of the part of the N-dimensional space of weights of a single perceptron, A_i, which

—are normalised;

† After the completion of this work we have received a preprint by G A Kohring [6], arriving at very similar conclusions.

—embed a set of p random patterns ϕ_i^μ with stability parameter K;
—have a set of signs prescribed by $g_i = \pm 1$.

These conditions are summarised, respectively, by the relations

$$\sum_{i=1}^{N} A_i^2 = N \tag{4}$$

$$\sum_{i=1}^{N} A_i \phi_i^\mu > K \qquad \text{for all } \mu \tag{5}$$

$$A_i g_i > 0 \qquad \text{for } i = 1, \ldots, N. \tag{6}$$

Note that the normalisation on the weights chosen in (4) differs by a factor of N from that used in the discussion of the perceptron.

This volume can be written as an integral over the full space of weights of the form

$$V(\{G_i\}) = \int_{-\infty}^{\infty} \prod_{i=1}^{N} \mathrm{d}A_i \delta\left(\sum_{i=1}^{N} A_i^2 - N \right) \prod_{\mu=1}^{p} \Theta\left(\sum_{i=1}^{N} A_i \phi_i^\mu - K \right) \prod_{i=1}^{N} \Theta(A_i g_i) \tag{7}$$

where each of the three factors in the integrand represents the corresponding constraint (4), (5) and (6).

The computation proceeds as in [2], i.e. one concentrates on the computation of $\ln(V)$, which is a quantity of order N. This quantity is averaged over the quenched distribution of the random patterns, ϕ, in the expectation that the fluctuations of $\ln(V)$ from sample to sample will be negligible. The average of this quantity is then carried out by the replica technique.

It is already at this point that one can establish the local gauge invariance of the theory for random patterns. Consider the average over the patterns of any function of V, equation (7). If the sign of g_i, for any i, is reversed, the effect within V can be compensated by a change in the sign of the corresponding ϕ_i. But since the ϕ are averaged over both signs of each of their components, the averaged quantity is unchanged. Note that we have not averaged over the g_i, but compared the same quantity for two specific realisations of the signs. The quantity of interest can therefore be written as

$$S = \lim \frac{1}{nN} \left\langle \left\langle \int_0^{\infty} \prod_{i=1}^{N} \prod_{\rho=1}^{n} \mathrm{d}A_i^\rho \delta\left(\sum_i (A_i^\rho)^2 - N \right) \prod_{\mu=1}^{p} \Theta\left(\sum_{i=1}^{N} A_i \phi_i^\mu - K \right) \right\rangle \right\rangle_\phi \tag{8}$$

where the limit takes N to infinity and n to zero. Note that the lower limit on the A_i^ρ integrations has been set to zero. This expresses the gauge invariance explained above, which allows all $g_i = 1$.

Taking the theory to be replica symmetric, one performs all the integrals as well as take the limits in N and n. The result is expressed in terms of three parameters E, F and q, which enter much in the same way as in the original calculation of Gardner. Namely, S can be written as

$$S(E, F, q, K) = \alpha G_0(q, K) + G_1(E, F, q) \tag{9}$$

where the order parameter q is the replica symmetric off-diagonal term of the correlation matrix of different possible solutions, i.e.

$$q^{\rho\sigma} = \frac{1}{N} \sum_{i=1}^{N} A_i^\rho A_i^\sigma \tag{10}$$

which measures the size of the solution volume (it is assumed to be independent of ρ and σ for $\rho \neq \sigma$). In particular, as $q \to 1$ this volume shrinks to zero, identifying the storage capacity. The parameters E and F are Lagrange parameters enforcing the normalisation, (4) and (10), respectively. The parameter $\alpha = p/N$, as usual. The value of the function S, for a given value of K, is the minimal value of the right-hand side of (9), under the variation of the other three parameters. All this is in complete parallel with [2]. In fact the only difference turns out to be in the expression for G_1, which is

$$G_1(E, F, q) = \tfrac{1}{2}qF + \int_{-\infty}^{\infty} \frac{dz}{\sqrt{2\pi}} \exp(-z^2/2) \times \ln\left(\int_0^{\infty} \frac{dA}{\sqrt{2\pi}} \exp([-\tfrac{1}{2}(2E+F)A^2 - \sqrt{F}\, zA + E] \right). \tag{11}$$

The only difference with Gardner's result [2] is that the internal integral ranges here from 0 rather than from $-\infty$.

It is now straightforward to write down the limit implied in (8). The result is

$$S = \tfrac{1}{2}qF + \alpha \int_{-\infty}^{\infty} \frac{dt}{\sqrt{\pi}} \exp(-t^2/2) \ln\left(\int_{\Lambda(q,t)}^{\infty} \frac{d\lambda}{\sqrt{2\pi}} \exp(-\lambda^2/2) \right) + \tfrac{1}{2} \ln\left(\frac{2\pi}{2E+F} \right) + \frac{F}{2(2E+F)} + E + \int_{-\infty}^{\infty} \frac{dz}{\sqrt{2\pi}} \exp(-z^2/2) \ln\left(\int_{\Omega(E,F,z)}^{\infty} \frac{d\omega}{\sqrt{2\pi}} \exp(-\omega^2/2) \right) \tag{12}$$

where

$$\Lambda(q, t) = \frac{K + \sqrt{q}\, t}{\sqrt{1-q}} \qquad \Omega(E, F) = \frac{\sqrt{F}}{\sqrt{2E+F}}.$$

The saddle-point equations are obtained from (12) by equating the derivatives of S with respect to E, F and q to zero. Instead of writing these equations in the general situation we shall proceed directly to the asymptotic region, in which the storage capacity is determined. That region is in the neighbourhood of $q = 1$.

3. The saddle-point equations near saturation

We shall concentrate first on the equations which determine F and E in terms of q. Then the final equation will be studied by comparison with the final Gardner equation. We shall assume that as $q \to 1$, $F/(2E+F)$ diverges, as in [2] and verify that the solution satisfies this condition. In this limit, if $z > 0$, we can write

$$\int_{\Omega(E,F,z)}^{\infty} \frac{d\omega}{\sqrt{2\pi}} \exp(-\omega^2/2) \equiv \tfrac{1}{2}[1 - \operatorname{erf}(z\Omega/\sqrt{2})] \approx \frac{1}{2\sqrt{\pi}\,\Omega z} \exp(-z^2\Omega^2/2)$$

where erf( ) is the error function. For $z < 0$ the integral approaches unity up to terms which are exponentially small.

The part of S which depends on E and F is approximately

$$S = \tfrac{1}{2}qF - \tfrac{1}{4}\ln F - \tfrac{1}{4}\ln(2E+F) + \frac{F}{2E+F} + E.$$

Consequently the saddle-point equations for these variables are

$$E + F = (2E+F)^2 \qquad q - \frac{1}{2F} = \frac{F}{2(2E+F)^2}.$$

Near $q=1$ these equations can be solved to give E and F as a power expansion in $1-q$. The solution is

$$E = \frac{-1+4(1-q)}{4(1-q)^2} \qquad F = \frac{1-3(1-q)}{2(1-q)^2}.$$

The condition $\Omega = F/(2E+F) \to \infty$ as $q \to 1$ is indeed verified.

4. The critical storage level

The final step is to introduce the expressions for E and F in terms of q into S, to write the saddle-point equation for q and to find the highest value of $\alpha(K)$ for which this equation has a non-zero solution. When E and F are substituted in S one finds to leading order in $1-q$

$$S \approx \frac{1}{4(1-q)} + \alpha \int_{-\infty}^{\infty} \frac{dt}{\sqrt{\pi}} \exp(-t^2/2) \ln\left(\int_{\lambda(q,t)}^{\infty} \frac{d\lambda}{\sqrt{2\pi}} \exp(-\lambda^2/2) \right).$$

This expression can be compared with (20) of [2] which, to leading order in $1-q$, is

$$G(q) \approx \frac{1}{2(1-q)} + \alpha \int_{-\infty}^{\infty} \frac{dt}{\sqrt{\pi}} \exp(-t^2/2) \ln\left(\int_{\lambda(q,t)}^{\infty} \frac{d\lambda}{\sqrt{2\pi}} \exp(-\lambda^2/2) \right).$$

Hence the correspondence of the saddle-point equations is complete if α in the present calculation is substituted by $\alpha/2$ of the unconstrained network. The critical value of α, $\alpha_c(K)$ is one half the critical storage of the unconstrained network. In particular, for $K=0$ the constrained network can store N patterns, compared with $2N$ for the unconstrained one.

5. Conclusion

The above discussion demonstrates that when synaptic signs are prescribed and quenched synaptic coefficients can be found to store half as many random patterns as would be possible for an unconstrained network, with the same local stability parameter. The finiteness of the volume in the space of coupling constants guarantees that in the limit of a large network such a set of couplings can be found with probability one for any set of random patterns, below saturation.

In our previous study [1] the discussion was restricted to the learning of random patterns with zero stability parameter. Nevertheless, the search in weight space for regions which store random patterns with a finite stability parameter is directly relevant, since the maximal stability parameter for a given number of random patterns controls

the speed of convergence of the algorithm at zero stability [6]. As has been shown by Gardner [2], the extension of the learning algorithm to the case of a finite stability parameter presents no difficulty.

Since the storage capacity in the constrained network is half that of the unconstrained case, one may be led to speculate that the solution for the coupling matrix in the constrained case is attained by pruning half of the synapses, namely those which are 'of the wrong sign'. This would imply that half of the weights will be zero. We can test this conclusion by considering the distribution of the values of any given weight, e.g. A_1, as given by

$$P(A_0)=\left\langle\!\!\left\langle \frac{\Pi_j \int \mathrm{d}A_i \Theta(g_i A_i)\delta(\Sigma_j A_j^2 - N)\,\Pi_\mu\,\Theta(\Sigma_j A_j \phi_j^\mu - K)\delta(A_0 - A_1)}{\Pi_j \int \mathrm{d}A_i \Theta(g_i A_i)\delta(\Sigma_j A_j^2 - N)\,\Pi_\mu\,\Theta(\Sigma_j A_j \phi_j^\mu - K)} \right\rangle\!\!\right\rangle .$$

Using the replica method we arrive at

$$P(A_0)=\Theta(g_1 A_0)\sqrt{\frac{(2E+F)^2}{4\pi(E+F)}}\exp\left(-\frac{(2E+F)^2}{4(E+F)}A_0^2\right)$$
$$\times \int \frac{\mathrm{d}z}{\sqrt{2\pi}}\exp(-z^2/2)\left(\int_{\bar{\Omega}(E,F,z)}^{\infty}\frac{\mathrm{d}\omega}{\sqrt{2\pi}}\exp(-\omega^2/2)\right)^{-1}$$

where

$$\bar{\Omega}(E,F,z)=\sqrt{\frac{F}{2(E+F)}}\left(z-\sqrt{\frac{F(2E+F)}{2(E+F)}}\,g_1 A_0\right).$$

In particular, near saturation,

$$P(A_0)=\Theta(g_1 A_0)\exp(-A_0^2/\sqrt{4\pi})$$

which is a Gaussian of variance 2, truncated on the side of the wrong signs.

One can therefore conclude that no 'wrong sign' weights of the unconstrained network are actually pruned. Instead, the constrained network stabilises the patterns by finding a new solution which is uncorrelated with the unconstrained ones. The low incidence of zero weights is confirmed by numerical experiments.

For a purely ferromagnetic (purely excitatory) network the system is equivalent to a ferromagnetic spin glass. The truncated Gaussian distribution of weights still allows for a multi-valley energy landscape, in which the N patterns can be stored, as in any other distribution of signs.

It has also been shown that the volume in weight space which can store a given number of random patterns is independent of the particular distribution of signs, i.e. the gauge invariance invoked in [1]. The question of the eventual dynamics and the concomitant basins of attraction has been left open. This question becomes of special concern when one realises that sets of random patterns can even be stored with a purely ferromagnetic (purely excitatory) matrix.

It turns out that *the invariance extends also to the dynamics.* On the formal level one observes that starting from any initial configuration of the network the successive configuration will be completely determined by the distribution of local fields generated by the initial configuration, via the given set of couplings [4, 5]. The probability distribution of these local fields, for a typical set of synaptic coefficients, can be computed directly [4, 5] and depends on the stored patterns in just the same way as the volume does (e.g. equation (7)). In other words, it is invariant under the change of the sign distribution.

Acknowledgments

DJA is indebted to the SERC for a fellowship which has made his stay at Imperial College possible and to Professor David Sherrington for hospitality. The work of KYMW has been supported by a grant from the SERC.

References

[1] Amit D J, Wong K Y M and Campbell C 1989 *J. Phys. A: Math. Gen.* **22** 2039
[2] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[3] Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan)
Minsky M L and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT Press)
[4] Krauth W, Nadal J-P and Mezard M 1989 *J. Phys. A: Math. Gen.* **21** 2995
[5] Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
[6] Kohring G A 1989 Coexistence of global and local attractors in neural network *Preprint* Bonn University