Optimal basins of attraction in randomly sparse neural network models

1989 J. Phys. A: Math. Gen. 22 1969

(http://iopscience.iop.org/0305-4470/22/12/002)

J. Phys. A: Math. Gen. 22 (1989) 1969-1974. Printed in the UK

Optimal basins of attraction in randomly sparse neural network models

E Gardner

Department of Physics, Edinburgh University, King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, UK

Received 9 February 1989

Abstract. The size of the basin of attraction for randomly sparse neural networks with optimal interactions is calculated. For all values of the storage ratio, $\alpha = p/C < 2$, where p is the number of random uncorrelated patterns and C is the connectivity, the basin of attraction is finite, while for $\alpha < 0.42$, the basin of attraction is (almost) 100%.

In the past few years, there has been much interest in neural network models (Little 1974, Hopfield 1982, Amit *et al* 1985, 1987, Gardner 1987a). Recently, an analytic approach was developed in order to calculate properties of interactions which can optimise the storage capacity and the size of the basins of attraction (Gardner 1987a, 1988, Gardner and Derrida 1988).

In this approach, a parameter $K > 0$ is introduced, and the constraints,

$$\xi_i^\mu \sum_j J_{ij} \xi_j^\mu > K \sqrt{\left\{ \sum_j J_{ij}^2 \right\}} \tag{1}$$

must be satisfied for each pattern μ and each site i where $\{\xi_i^\mu\}$ are a set of N-bit Ising spin configurations ($i = 1, \ldots N$, $\mu = 1, \ldots, P$) which one wants to store, and J_{ij} is the interaction strength from the site j to the site i. The self-interaction J_{ii} is defined to be zero and J_{ij} is in general asymmetric. The motivation for introducing the parameter K can be seen from the dynamics used for recovery of information in the network. This will be defined by

$$\begin{aligned} S_i(t+1) &= \operatorname{sgn}(h_i(t)) \\ h_i(t) &= \sum_j J_{ij} S_j(t) \end{aligned} \tag{2}$$

where $S_i(t)$ is the Ising spin configuration of the network at time t. (The iteration will be assumed to be totally parallel although the calculations can also be done for serial update.) If S_i^μ is a noisy version of the pattern ξ_i^μ obtained by flipping a finite fraction of spins at random, the pattern would be recovered in one iteration if,

$$\xi_i^\mu h_i(\{S_j^\mu\}) > 0 \tag{3}$$

for each site i. Noise reduction in one iteration is therefore more likely the larger the value of K. Clearly, parameters other than K could be important in determining the optimal basin of attraction (Krauth *et al* 1988). In this paper, however, only the one parameter space of K will be considered and the optimal attraction basin should be given by the maximal value of K. Interactions of a given value of K can be generated

0305-4470/89/121969+06$02.50

using a generalisation (Krauth and Mézard 1987, Gardner 1988, Forrest 1988, Kepler and Abbot 1988) of the perceptron learning algorithm (Rosenblatt 1962, Minsky and Papert 1969). This algorithm converges to some solution for the interactions of given K provided such solutions exist. Optimal content addressability is therefore obtained by increasing K as far as possible.

The maximum value of the storage ratio α for which random patterns can be stored was calculated (Gardner 1987a, 1988, Gardner and Derrida 1988). For uncorrelated images, the value decreases from $\alpha = 2$ for $K = 0$ (Cover 1965, Venkatesh 1986) and vanishes as K tends to infinity.

In this paper, results for static properties of storage will be generalised to optimal dynamical properties. The basin of attraction at the optimal value of K will be calculated as a function of the storage ratio α for random uncorrelated patterns on randomly dilute networks in the limit that the fraction of non-zero bonds tends to zero.

The random dilute model is defined by independently cutting bonds with probability $1 - C/N$. Exact results will be obtained in the limits C, N tend to infinity where C behaves as $\ln N$. The model is asymmetric since the bonds, J_{ij} and J_{ji}, are chosen independently and the probability that both are non-zero can be neglected in the thermodynamic limit. If the spin configuration $\{S_i(t)\}$ at time t has a finite (macroscopic) overlap $M(t)$ with pattern 1 at time t, i.e.

$$M(t) = N^{-1} \sum_i \xi_i^1 S_i(t) \tag{4}$$

then it will be shown that for parallel dynamics, $M(t+1)$ is given in the thermodynamic limit by

$$M(t+1) = f_K(M(t)) \tag{5}$$

for the dilute model where $t \to t+1$ implies one completely synchronous time step. f_K is a function which depends only on the overlap at time t. The exact solution for the dynamics can therefore be determined by repeated iteration of (5).

The reason that the model is exactly soluble or that f_K depends only on $M(t)$ for C of order $\ln N$ is identical to an argument used previously for the solution of models of random automata (Derrida and Pomeau 1986) and for the asymmetrically diluted Hopfield model (Derrida *et al* 1987). Consider a site i. The calculation of the spin $S_i(t)$ at site i and at time t involves knowing, by equation (2), all spins $S_{j(i)}(t-1)$ for sites j connected to site i. The determination of $S_i(t)$ as a function of the initial condition, $\{S_j(0)\}$ at time $t = 0$, thus involves a tree of ancestor sites starting at site i and branching number C. Provided

$$C^t \ll \sqrt{N} \tag{6}$$

then for a typical site i all sites in its tree will be different, i.e. there are no loops. The spins $S_{j(i)}(t-1)$ at time $t-1$ are therefore uncorrelated provided bonds coming into the site i are independent of those coming into a different site j. This was proved by Derrida and Pomeau (1986) and Derrida *et al* (1987). The dynamics, therefore, is determined completely in terms of the first time-step equation (5) which will now be derived.

We introduce the distribution, $P_K(\Lambda)$,

$$P_K(\Lambda) = \delta\left(\Lambda - \xi_i^1 \sum_j J_{ij} \xi_j^1 \left(\sum_j J_{ij}^2\right)^{-1/2}\right) \tag{7}$$

for a typical realisation of the patterns μ and a typical set of interactions which satisfy (1). If the configuration, $S_j^1(t)$ is defined by flipping spins in ξ_j^1 with a random independent probability, $\frac{1}{2}(1-M(t))$, then using equation (2),

$$f_K(M(t)) = \int d\Lambda P_K(\Lambda) 2 \operatorname{erf}\left(\frac{M(t)\Lambda}{\sqrt{1-M(t)^2}}\right) \tag{8}$$

where

$$\operatorname{erf}(x) = \frac{1}{\sqrt{2\pi}} \int_0^x dy \exp\left(\frac{-y^2}{2}\right). \tag{9}$$

$P_K(\Lambda)$ will now be derived using the maximum entropy methods of Gardner (1988). $P_K(\Lambda)$ is given by

$$P_K(\Lambda) = p_K(\Lambda)/Z \tag{10}$$

where

$$p_K(\Lambda) = N^{-1} \int \prod_{ij} dJ_{ij} \left[\sum_i \delta\left(\Lambda - \frac{1}{\sqrt{C}} \sum_j \xi_i^1 J_{ij} \xi_j^1\right)\right] \prod_{i,\mu} \theta\left(\sum_j J_{ij} \xi_i^\mu \xi_j^\mu - K\sqrt{C}\right)$$
$$\times \prod_i \delta\left(\sum_j J_{ij}^2 - C\right) \tag{11}$$

and

$$Z = \int \prod_{ij} dJ_{ij} \prod_{i,\mu} \theta\left(\sum_j J_{ij} \xi_i^\mu \xi_j^\mu - K\sqrt{C}\right) \prod_i \delta\left(\sum_j J_{ij}^2 - C\right). \tag{12}$$

In (11) and (12) the integrations are over the set of interactions $\{J_{ij}\}$ where the bond between sites i and j has not been cut and the spherical constraint

$$\sum_j J_{ij}^2 = C \tag{13}$$

has been imposed. $P_K(\Lambda)$ depends on the particular realisation of the random patterns. Since it is extensive, we will assume that it is self-averaging and the distribution for a typical realisation of the random patterns is therefore given by

$$\bar{P}_K(\Lambda) = \left\langle \lim_{n\to 0} Z^{n-1} p_K(\Lambda) \right\rangle \tag{14}$$

where $\langle\ \rangle$ represents an average over the distribution of patterns. From equations (11)-(14) we then obtain

$$\bar{P}_K(\Lambda) = \left\langle \lim_{n\to 0} \int \prod_{\alpha,j} dJ_{ij}^\alpha \delta\left(\Lambda - \frac{1}{\sqrt{C}} \sum_j \xi_i^1 J_{ij}^1 \xi_j^1\right) \prod_{\mu,\alpha} \theta\left(\frac{1}{\sqrt{C}} \sum_j \xi_i^\mu J_{ij}^\alpha \xi_j^\mu - K\right)\right.$$
$$\left. \times \prod_\alpha \delta\left(\sum_j (J_{ij}^\alpha)^2 - C\right)\right\rangle \tag{15}$$

where α is a replica index which runs from 1 up to n, and the index j runs over the C sites connected to site i. The calculations follow those of Gardner (1988) and are done by introducing integral representations for the θ and δ functions which appear in (15). This allows the averaging over different sites $j \neq i$ and different patterns $\mu \neq 1$ to be factorised so that the averaging can be done. In the large-C limit, the integral

can be done using the saddle-point method. The order parameters are assumed to be repica symmetric at the saddle point. The stability of this solution was given by Gardner (1988). One then finds that $\bar{P}_K(\Lambda)$ is given by

$$\bar{P}_K(\Lambda)=\frac{1}{\sqrt{(1-q)}}\,\theta(\Lambda-K)\int_{-\infty}^{\infty}\mathrm{d}t\,\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-t^2}{2}\right)\exp\left(\frac{-(t\sqrt{q}+\Lambda)^2}{2(1-q)}\right)$$
$$\times\left[\int_{(K+t\sqrt{q})/\sqrt{1-q}}^{\infty}\mathrm{d}\lambda\,\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\lambda^2}{2}\right)\right]^{-1} \tag{16}$$

where q is given by the maximum of the function

$$G(q)=\alpha\int\mathrm{d}t\,\frac{1}{2\pi}\exp\left(\frac{-t^2}{2}\right)\ln\int_{(K+t\sqrt{q})/\sqrt{1-q}}^{\infty}\mathrm{d}\lambda\,\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-\lambda^2}{2}\right)$$
$$+\frac{1}{2}\ln(1-q)+\frac{1}{2}\,\frac{q}{(1-q)} \tag{17}$$

where

$$\alpha=p/C \tag{18}$$

and p is the number of patterns. The physical interpretation of the 'Edwards-Anderson' order parameter,

$$q=q^{\alpha\beta}=\sum_j\frac{J_{ij}^{\alpha}J_{ij}^{\beta}}{C}\qquad\alpha\neq\beta \tag{19}$$

is that it is equal to the cosine of the solid angle of the cone of solutions to (1). The upper storage capacity is determined by taking the limit q tends to 1 or the limit that the solid angle of the cone tends to zero. Gardner (1987a, 1988) showed that the upper storage capacity for given K is given in this limit by

$$\alpha=\left[\int_{-K}^{\infty}\mathrm{d}t\,\frac{1}{\sqrt{2\pi}}(t+K)^2\exp\left(-\frac{t^2}{2}\right)\right]^{-1}. \tag{20}$$

In this limit there are two contributions to $\bar{P}_K(\Lambda)$, a δ function at $\Lambda=K$ from the part of the t integral with $t>-K$, and a Gaussian for $\Lambda>K$ from $t<-K$. $\bar{P}_K(\Lambda)$ is given by

$$\bar{P}_K(\Lambda)=\left(\frac{1}{2}+\mathrm{erf}(K)\right)\delta(\Lambda-K)+\theta(\Lambda-K)\exp\left(-\frac{\Lambda^2}{2}\right)\frac{1}{\sqrt{2\pi}} \tag{21}$$

and equation (8) becomes

$$M(t+1)=f_K(M(t))$$
$$=\left(\frac{1}{2}+\mathrm{erf}(K)\right)\mathrm{erf}\left(M(t)K\frac{1}{\sqrt{1-M(t)^2}}\right)$$
$$+\int_K^{\infty}\mathrm{d}\Lambda\exp\left(-\frac{\Lambda^2}{2}\right)\frac{1}{\sqrt{2\pi}}\,2\,\mathrm{erf}\left(\frac{M(t)\Lambda}{\sqrt{1-M(t)^2}}\right). \tag{22}$$

The behaviour of the parallel dynamics defined in equation (2) is therefore determined by equation (22) provided the initial configuration has a macroscopic overlap with one pattern only. K can be eliminated using equation (20) giving the maximal basins of attraction for fixed α. Equation (20) can be solved numerically and in figure 1 the fixed points of $M=f_K(M)$ are plotted as a function of α.

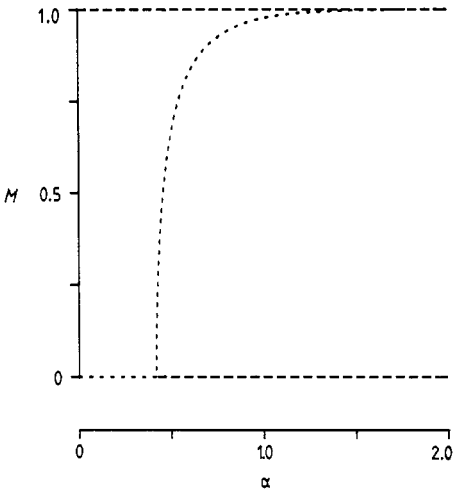

Figure 1. The stable fixed point (– – –) and unstable fixed point (M_L, ----) of $M = f_K(M)$ (equation (5)) plotted as a function of α.

For $\alpha < 2$, there is a stable fixed point at $M = 1$ with a finite basin of attraction; for an initial overlap, $M(0) > M_L$, the pattern is recovered. The unstable fixed point, $M_L \to 1$ only in the limit that $\alpha \to 2$, implying that for any value of α below 2, a pattern can be made stable with respect to a finite fraction of spin flips. For $M(0) < M_L$, the pattern iterates onto the fixed point at $M = 0$. It is possible to show that the $M = 0$ fixed point of (22) corresponds to a long cycle in configuration space. This can be done by deriving the equation for the overlap between configurations at different time steps in a similar way to the derivation of (22). If the configuration iterates onto the $M = 0$ fixed point, the overlap between configurations at finite time differences tends to zero. For values of $\alpha < \alpha_B = 0.42$, the unstable fixed point, $M_L = 0$ and the basin of attraction is (almost) 100%; this means that configurations with a macroscopic overlap with a pattern (a non-zero value of M however small) iterate onto the pattern. For $\alpha_B < \alpha < 2$ the unstable fixed point M_L increases and the size of the basin of attraction decreases.

For the fully connected model, equation (22) gives correct results for the first time step of parallel iteration. Figure 1 therefore gives information about whether the initial configuration moves towards or away from the input pattern at the first time step of parallel iteration. Comparison with numerical results (Forrest 1988) for the symmetric fully connected model shows that M_1, the size of the basin of attraction as determined by the first time step (M_1 is the initial value of M above which the configuration moves towards the pattern at the first time step), is typically larger than M_0, the size of the domain of attraction determined by complete iteration to stability. For example,

$$M_0 = \begin{cases} 0.4 & \text{at } \alpha = 0.25 \\ 0.75 & \text{at } \alpha = 0.5 \end{cases}$$

after complete iteration while

$$M_1 = \begin{cases} 0 & \text{at } \alpha = 0.25 \\ 0.67 & \text{at } \alpha = 0.5 \end{cases}$$

at one time step. A similar effect occurs in the Hopfield-Little model (Little 1974, Hopfield 1982) where the critical storage ratio α_c increases from 0.14 for infinite times to 0.63 at the first time step (Gardner *et al* 1987).

Basins of attraction are not improved by going to higher-order interactions (Gardner 1987b). The above calculations can also be done for interactions of order $q > 2$. The dynamics of (2) is replaced by

$$S_i(t+1) = \operatorname{sgn}\left(\sum_{i_2 < \ldots < i_q} J_{i,i_2,\ldots,i_q} S_{i_2}(t) \ldots S_{i_q}(t) \right) \tag{23}$$

where $J_{i,\ldots,i_q}$ is the q-body interaction from sites $i_2, \ldots, i_q$ into site i. If C is the total number of connections into a site, then equation (20) is unaltered and so the upper storage capacity is independent of the order of the interaction provided α is scaled in terms of the connectivity, C. The basins of attraction, however, become smaller as q increases since $M(t)$ on the right-hand side of (22) is replaced by $(M(t))^{(q-1)}$ using the same argument as was used in the derivation of (8).

It would be interesting to extend the results of this paper to the storage of correlated patterns. Although the storage capacity is larger (Gardner 1987a, 1988, Gardner and Derrida 1988) the basins of attraction should be limited in directions associated with this correlation.

Acknowledgments

I thank B Derrida for useful discussions. Financial support was from the Science and Engineering Research Council.

Note added. This paper has been prepared from an unfinished, hand-corrected typescript and Elizabeth may not have intended it to be published in this form exactly. It should be noted that since Elizabeth drafted the paper a calculation of $\bar{P}_K(\Lambda)$ for the fully connected model has been published by Kepler and Abbott (1988). The final typescript was produced by Martin Evans and Nick Stroud.

References

Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev. Lett.* **55** 1530
—— 1987 *Ann. Phys., NY* **173** 30
Cover T M 1965 *IEEE Trans. Electron. Comput.* **EC-14** 326
Derrida B and Pomeau Y 1986 *Europhys. Lett.* **1** 45
Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
Forrest B M 1988 *J. Phys. A: Math. Gen.* **21** 245
Gardner E 1987a *Europhys. Lett.* **4** 481
—— 1987b *J. Phys. A: Math. Gen.* **20** 3453
—— 1988 *J. Phys. A: Math. Gen.* **21** 257
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
Gardner E, Derrida B and Mottishaw P 1987 *J. Physique* **48** 741
Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745
Krauth W, Nadal J P and Mézard M 1988 *J. Phys. A: Math. Gen.* **21** 2995
Little W A 1974 *Math. Biosci.* **19** 101
Minsky M and Papert S 1969 *Perceptrons: An Introduction to Computational Geometry* (Cambridge, MA: MIT Press)
Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan)
Venkatesh S 1986 *Neural Networks for Computing* (*AIP Conf. Proc.* **151**) ed J S Denker (New York: American Institute of Physics) p 440