Multivalley structure of attractor neural networks

1997 J. Phys. A: Math. Gen. 30 7945

(http://iopscience.iop.org/0305-4470/30/22/028)

J. Phys. A: Math. Gen. **30** (1997) 7945–7951. Printed in the UK PII: S0305-4470(97)84284-9

Multivalley structure of attractor neural networks

C Rodrigues Neto and J F Fontanari

Instituto de Física de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560-970 São Carlos SP, Brazil

Received 13 May 1997, in final form 18 August 1997

Abstract. We numerically investigate the structure of the basins of attraction and the nature of the spurious attractors of the pseudo-inverse and the optimal weights attractor neural networks. We show that the number of attractors in the optimal weights model increases as the margin parameter κ increases, and that the basins of attraction of the stored patterns are not significantly enlarged by κ. Moreover, the number of attractors is smaller than in the pseudo-inverse model.

1. Introduction

Despite the importance of the pseudo-inverse [1, 2] and the optimal weights [3] attractor neural networks to the theoretical modelling of associative memory systems, little is known about the structure of their basins of attractions and the nature of their spurious attractor states. As pointed out by Hopfield [4], attractor neural networks can function as associative memory devices provided that the synaptic weights J_{ij} are specified so that a given set of P binary patterns $\boldsymbol{\xi}^l = (\xi_1^l, \ldots, \xi_N^l)$, $l = 1, \ldots, P$ become the attractors of the neural dynamics

$$S_i(t+1) = \operatorname{sign}\left(\sum_j J_{ij} S_j(t) \right) \qquad i = 1, \ldots, N. \tag{1}$$

Here $S_i(t) = \pm 1$ is the state of neuron i at time t. It is usually assumed that the components ξ_i^l are randomly chosen as ± 1 with equal probability, and that the number of patterns scales linearly with N, i.e. $P = \alpha N$. The retrieval of a stored pattern, say $\boldsymbol{\xi}^l$, will depend on its proximity to the initial state $\boldsymbol{S}(t = 0)$. However, since in general the number of spurious attractors grows exponentially with N, a randomly chosen initial state will almost certainly flow to one of these attractors rather than to one of the stored patterns. Hence the importance of understanding the nature of the spurious attractors in models of associative memory.

In this paper we numerically investigate the multivalley structure of the configuration space of the pseudo-inverse and optimal weights neural network models for both sequential (the neurons are updated one after the other in a fixed order) and parallel (all neurons are updated simultaneously) dynamics.

Denoting by Ω_s the number of initial states that fall onto the sth attractor, we define the weight W_s of this attractor by

$$W_s = \Omega_s / 2^N \tag{2}$$

so that $\sum_s W_s = 1$. To characterize the structure of the basins of attraction we evaluate the quantity

$$Y = \sum_s W_s^2 \tag{3}$$

0305-4470/97/227945+07$19.50

which gives the probability that two randomly chosen initial states fall onto the same attractor [5]. If $Y \to 0$ as $N \to \infty$ then all the weights become smaller and smaller, while if Y remains non-zero there are some big basins of attraction that fill a finite fraction of the configuration space. We note that the minimum of $0 < Y \leqslant 1$ is obtained for the uniform case $W_s = W \forall s$, while the maximum is obtained for $W_r = 1$ and $W_s = 0 \forall s \neq r$. We calculate also the number of fixed points and cycles, denoted by $\mathcal{N}_f$ and $\mathcal{N}_c$, respectively, of equation (1). To evaluate the basins of attraction of the stored patterns we calculate the number of states that fall onto them

$$X_\xi = \sum_{l=1}^{P} \Omega_{\xi^l} \tag{4}$$

so that X_ξ / P gives the average size of the basin of attraction of a stored pattern.

For each realization of the set of stored patterns, the quantities defined above are computed exactly using an efficient algorithm for performing the exhaustive search in the configuration space [6]. The results are then averaged over 100 realizations. Since the computer time needed for the complete enumeration of the states $\boldsymbol{S}(t+1)$ for each of the 2^N states $\boldsymbol{S}(t)$ grows exponentially with N our analysis is restricted to $N \leqslant 24$.

The remainder of the paper is organized as follows. In section 2 we review briefly the main results about the pseudo-inverse and optimal weights attractor neural networks, and present the prescriptions used for obtaining their synaptic weights. The results of our simulations are presented and discussed in section 3, while section 4 is devoted to our concluding remarks.

2. The models

The weights of the pseudo-inverse attractor neural network are given by

$$J_{ij} = \frac{1}{N} \sum_{kl} \xi_i^k \xi_j^l (\mathcal{C}^{-1})_{kl} \tag{5}$$

where $\mathcal{C}$ is the correlation matrix whose elements are $\mathcal{C}_{kl} = \frac{1}{N} \sum_i \xi_i^k \xi_i^l$ [1, 2]. This prescription for the weights guarantees the perfect storage of a set of N linearly independent patterns. Hence its storage capacity, defined as the ratio between the maximal number of random patterns it can store and the number of neurons, is $\alpha_c = 1$. Thanks to the analytical relation between J_{ij} and the set of stored patterns given in equation (5), several analytical investigations of the pseudo-inverse model have been carried out. In particular, the thermodynamics of a variant of this model, where the diagonal terms are set to zero, was studied by Kanter and Sompolinsky [7], while the number of metastable states for the original model as well as for the mentioned variant was calculated by Kuhlmann and Anlauf [8]. An alternative formulation of the pseudo-inverse model is obtained by calculating the minimal norm solution of the following set of P linear equations [9]

$$\Delta_i^l = 1 \qquad l = 1, \ldots, P \tag{6}$$

for each $i = 1, \ldots, N$, where

$$\Delta_i^l = \frac{1}{\sqrt{N}} \xi_i^l \sum_{j \neq i} J_{ij} \xi_j^l \tag{7}$$

is termed the stability of the component ξ_i^l. We note that, in contrast to equation (5), this prescription yields an asymmetric matrix of weights. Since it does not specify the diagonal terms, we impose the additional condition $J_{ii} = 0$. Moreover, as for fixed i there are

$N-1$ unknowns and P equations the above prescription allows the perfect storage of $N-1$ linearly independent patterns, so that $\alpha_c = (N-1)/N = 1 - 1/N$.

The pseudo-inverse model, however, is not optimal in the sense that its storage capacity is not maximal. In fact, Gardner has shown that for unbiased random patterns the maximal storage capacity is $\alpha_c = 2$ [3]. The statistical properties of the ensemble of weights of this optimal neural network can be investigated analytically by calculating the volume of the space of weight vectors $\boldsymbol{J}_i = (J_{i1}, \ldots, J_{ii-1}, J_{ii+1}, \ldots, J_{iN})$ that satisfy the inequalities

$$\Delta_i^l \geqslant \kappa \qquad l = 1, \ldots, P \tag{8}$$

for each $i = 1, \ldots, N$. The margin parameter $\kappa \geqslant 0$ is introduced in order to ensure that the stored patterns $\boldsymbol{\xi}^l$ possess finite basins of attraction. The storage capacity decreases with increasing κ and, in particular, $\alpha_c = 2$ for $\kappa = 0$ [3]. The lack of an explicit, analytical relation between the weights and stored patterns makes the analytical study of the retrieval properties of the optimal neural network very difficult. From the perspective of a numerical investigation, however, this is not a hindrance, since given a set of stored patterns there is a very efficient iterative algorithm that guarantees convergence to an optimal set of weights (if one exists) [10]: starting with a random matrix, repeatedly apply the rule

$$J_{ij} \to J_{ij} + \frac{1}{N}\xi_i^l \xi_j^l (1 - \delta_{ij})\ f(\Delta_i^l)\ |J_i|\ \Theta(\kappa - \Delta_i^l) \tag{9}$$

for each i and l until (8) is satisfied. Here $|J_i| = [\sum_{j \neq i} J_{ij}^2]^{1/2}$ and

$$f(\Delta_i^l) = \kappa + \epsilon - \Delta_i^l + [(\kappa + \epsilon - \Delta_i^l)^2 - \epsilon^2]^{1/2} \tag{10}$$

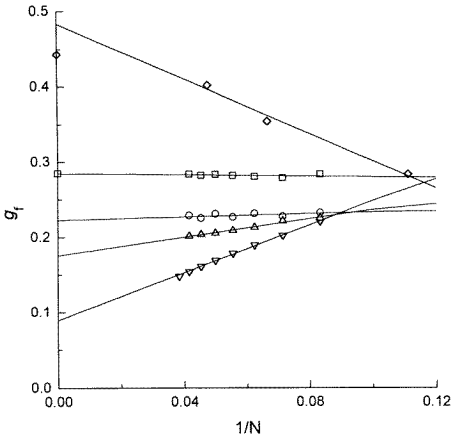
with $\epsilon = 0.01$.

We should mention that in the regime of extreme dilution, where the connectivity C of the network satisfies $C \ll \ln N$ [11], the time evolution of the overlaps between the retrieval states and the stored patterns is given by a simple equation which depends only on the probability distribution of the stabilities Δ_i^l [12, 13]. The diluted versions of the pseudo-inverse and the optimal weights models have been thoroughly studied [13, 9, 14, 15]. In particular, for the optimal weights model, it was found that the basins of attraction of the stored patterns vanish at $\kappa = 0$ for $\alpha \leqslant 2$ [15]. More recently, a new approach introduced by Coolen and Sherrington [16, 17] has allowed the analytical study of the dynamics of the retrieval overlaps for the fully connected Hopfield model. Their method, however, does not seem to be applicable to models for which there is no analytical prescription for writing the synaptic weights in terms of the stored patterns.

3. Analysis of the results

In the following we will refer to the pseudo-inverse prescription (5) as the PGD model [2], and to its variant with $J_{ii} = 0$ as the KS model [7]. An unexpected result of our simulations is the finding that, in the scale of the figures presented in the sequel, the results for the KS and the alternative formulation of the pseudo-inverse (6) were indistinguishable, in spite of the considerable difference in the weight matrices. Hence we will present the results for the KS model only. In order to prevent the vanishing of the argument of the sign function in (1) we have simulated the optimal weights and KS models for even values of N, and the PGD model for odd ones. We will focus mainly on the sequential dynamics; the results for the parallel dynamics will be considered briefly in the end of this section.

In figure 1(a) we present $g_f = \frac{1}{N}\ln\langle\mathcal{N}_f\rangle$ as a function of $1/N$ for $\alpha = \frac{1}{2}$ (optimal weights and KS) and $\alpha = \frac{1}{3}$ (PGD). The largest size is $N = 24$ for the optimal weights and



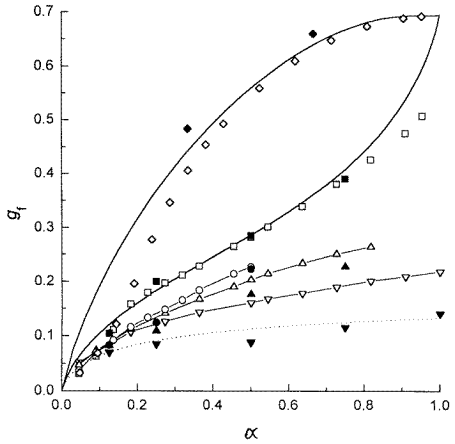


Figure 1. (*a*) Size dependence of the exponent g_f in $\langle \mathcal{N}_f \rangle = \exp(N g_f)$ for the sequential dynamics of the following models: PGD (◇), KS (□), and optimal weights with $\kappa = 0$ (▽), 0.5 (△) and 0.8 (○). Note that $\alpha = \frac{1}{3}$ for PGD and $\alpha = \frac{1}{2}$ for the other models. The data for $1/N = 0$ are the theoretical predictions [8]. (*b*) The exponent g_f for the sequential dynamics as a function of α for $N = 21$ (PGD) and $N = 22$ (other models). The full symbols are the extrapolated values for $1/N \to 0$ and the full lines for the pseudo-inverse models are the theoretical predictions [8]. The broken line is the theoretical prediction for the Hopfield model [18, 19]. The convention is the same as for (*a*).

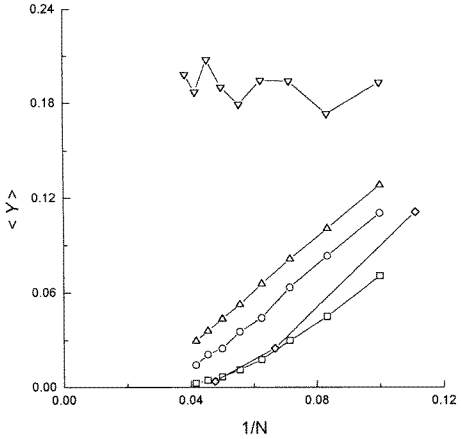


Figure 2. Size dependence of $\langle Y \rangle$ for the sequential dynamics. The parameters and convention are the same as for figure 1(*a*).

KS models, and $N = 21$ for the PGD model. These data indicate an exponential growth of the number of fixed points for all models,

$$\langle \mathcal{N}_f \rangle = c_\alpha \exp(g_\alpha N) \tag{11}$$

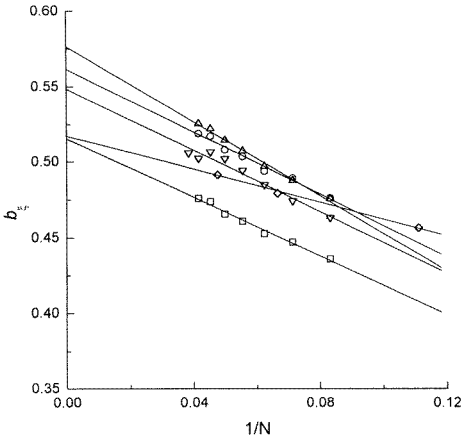
where, for each model, $\ln c_\alpha$ is given by the slope of the straight line and g_α by its intersection with the vertical axis. The analytical prediction of Kuhlmann and Anlauf [8] for $1/N \to 0$ is also shown in that figure. The poor agreement for the PGD model is due to the small

values of P used in the simulations ($P = 3$, 5 and 7), since the analytical calculation is based on the assumption that both P and N go to infinity for any non-zero α. Due to the limitation $N \leqslant 24$, these extrapolations can be carried out for a few values of α only. So, in order to better appreciate the α dependence of g_f, as well as of the other quantities we consider in this work, we have also presented the data for small size networks. Figure 1(b) shows the dependence of g_f with α for $N = 22$ (optimal weights and KS) and $N = 21$ (PGD). The full curves for the pseudo-inverse are the theoretical predictions [8], and the full symbols are the extrapolated values for $1/N \to 0$. As already mentioned, the disagreement for small α was expected; the disagreement for the KS model for $\alpha \approx 1$ is probably due to the fact that $\alpha_c = 1 - 1/N \approx 0.95$ for this model. The number of spurious fixed points is then much smaller in the optimal weights model and grows much slower than for the pseudo-inverse. Besides, for the optimal weights model the number of spurious fixed points does not tend to the saturation value $g_f = \ln 2$ as the storage limit is approached. We note that, in the limit $N \to \infty$, $\alpha_c = 2$, 0.96 and 0.66 for $\kappa = 0$, 0.5 and 0.8, respectively. Interestingly, there is a surprisingly good agreement between the extrapolated data for the optimal weights model with $\kappa = 0$ and the theoretical prediction for the Hopfield model [18, 19]. Finally, we should mention that the calculation of $\langle \ln \mathcal{N}_f \rangle$ yields no significant differences from the results presented above. Similarly, we have calculated $g_c = \frac{1}{N} \ln \langle \mathcal{N}_c \rangle$ for the optimal weights model (since the weight matrix is not symmetric, even the sequential dynamics can lead to cycles). We have found that though the number of cyclic attractors seems to increase exponentially with N, its number is about ten times smaller than the number of fixed points. Moreover, cycles of length two are much more numerous than cycles of larger length.

The dependence of $\langle Y \rangle$ on $1/N$ for fixed α presented in figure 2 shows that there is a major qualitative difference between the optimal weights model with $\kappa = 0$ and the other models. This result is not affected by different choices of the parameter α. According to the physical interpretation of Y, the tendency of $\langle Y \rangle$ to a non-zero value as $N \to \infty$ suggests the existence of few attractors with huge basins of attraction, i.e. $\Omega_s \approx c_s \exp(N \ln 2)$ with $0 < c_s \leqslant 1$. We note that $\langle Y \rangle$ will vanish even if Ω_s increases exponentially with N, say $\Omega_s \approx \exp(N a_s)$, provided that $a_s < \ln 2$. As expected, we have found that for all models $\langle Y \rangle$ decreases monotonically with increasing α, indicating thus the reduction of the size of the basins of attraction of all attractors.

Figures 3(a) and (b) show the dependence of $b_\xi = \frac{1}{N} \ln(\langle X_\xi \rangle / P)$ on N and α, respectively. Although $\langle X_\xi \rangle$ increases exponentially with N, the extrapolation to $N \to \infty$ yields $b_\xi < \ln 2$, so the weights of the stored patterns vanish in that limit. In contrast to the findings for the diluted model [15], the basins of attraction of the stored patterns do not vanish for $\kappa = 0$. As pointed out by Kanter and Sompolinsky [7], the basins of attraction of the stored patterns in the PGD model vanish (i.e. $X_\xi / P = 1$) for $\alpha \geqslant 0.5$ in the thermodynamic limit. The data shown in figure 3(b) indicate this tendency for the PGD model. In particular, the extrapolations for $\alpha = \frac{1}{3}$ and $\frac{2}{3}$ point to a very sharp transition at $\alpha = 0.5$. As for the optimal weights model, the margin parameter κ plays a surprisingly minor role in the determination of the size of the basins of attraction of the stored patterns: for $\alpha > 0.5$ there is practically no differences between the results for $\kappa = 0$ and $\kappa = 0.5$.

We turn now to the analysis of the parallel dynamics. In this case, the models with symmetric weight matrices (PGD and KS) can present cycles of length two, besides the usual fixed points [20]. Of course, the results concerning the number of fixed points, figure 1(b), are not changed, since the fixed points must be the same for both dynamics. Their basins of attraction, however, undergo great changes mainly due to the appearance of the new cyclic attractors. In figure 4 we present g_c as function of α. The positive diagonal term of the



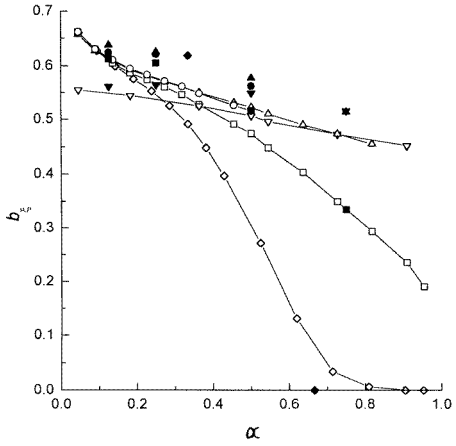


Figure 3. (*a*) Size dependence of the exponent b_ξ in $\langle X_\xi \rangle = P \exp(Nb_\xi)$ for the sequential dynamics. The parameters and convention are the same as for figure 1(*a*). (*b*) The exponent b_ξ for the sequential dynamics as a function of α for $N = 21$ (PGD) and $N = 22$ (other models). The convention is the same as for figure 1(*a*).

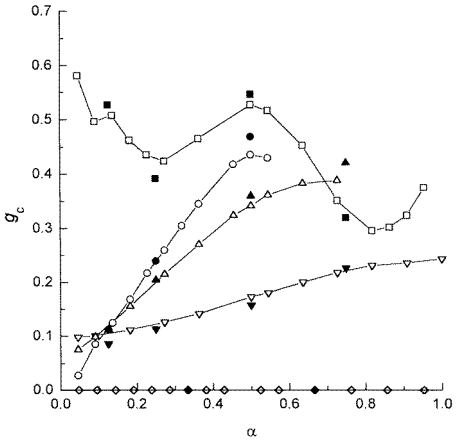


Figure 4. The exponent g_c in $\langle \mathcal{N}_c \rangle = \exp(Ng_c)$ for the parallel dynamics as a function of α for $N = 21$ (PGD) and $N = 22$ (other models). The convention is the same as for figure 1(*a*).

PGD weight matrix ($J_{ii} \approx \alpha$ [7, 8]) prevents the appearance of cycles in this model. For the other models, however, the cycles outweigh the fixed points, except for KS near its storage capacity limit, as can be seen by comparing figures 4 and 1(*b*). Similar to the serial dynamics, we have found that in the optimal weights model the cycles of length two are the predominant ones. The reduction in the size of the basins of attraction of the stored patterns resulting from the appearance of the cycles becomes more pronounced for large α, but there is no noteworthy difference between the behaviour pattern of b_ξ for the parallel and the serial dynamics.

4. Conclusion

The results of the previous section indicate that the number of spurious attractors in the optimal weights model is smaller than in the pseudo-inverse model. The number of attractors increases with κ and, perhaps because of this increase, the basins of attraction of the stored patterns are not greatly enlarged by the margin parameter κ as one would naively expect. In fact, the gain is so small that, in view of the drastic reduction of the storage capacity, it becomes hard to justify the use of $\kappa \neq 0$ for the fully connected neural network. It seems thus that the role of the margin parameter κ is simply to smooth out the neighbourhood of the stored patterns [15], with no great consequences to the size of their basins of attraction.

To conclude, we mention that the extrapolations to $N \to \infty$ of g_f from its N dependence in small networks ($N \leqslant 24$) provide a reliable quantitative base line for analytical calculations of the number of metastable states of the optimal weights model.

Acknowledgments

This work was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). CRN holds a FAPESP fellowship.

References

[1] Kohonen T 1984 *Self-organisation and Associative Memory* (Berlin: Springer)
[2] Personnaz L, Guyon I and Dreyfus G 1986 *Phys. Rev.* A **34** 4217
[3] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[4] Hopfield J J 1982 *Proc. Natl Acad. Sci., USA* **79** 2554
[5] Mezard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[6] Gutfreund H, Reger J D and Young A P 1988 *J. Phys. A: Math. Gen.* **21** 2775
[7] Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
[8] Kuhlmann P and Anlauf J K 1994 *J. Phys. A: Math. Gen.* **27** 5857
[9] Opper M, Kleinz J, Köhler H and Kinzel W 1989 *J. Phys. A: Math. Gen.* **22** L407
[10] Abbott L F and Kepler T B 1989 *J. Phys. A: Math. Gen.* **22** L711
[11] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **2** 337
[12] Kepler T B and Abbot L F 1988 *J. Physique* **49** 1657
[13] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969
[14] Amit D J, Evans M R, Horner H and Wong K Y M 1990 *J. Phys. A: Math. Gen.* **23** 3361
[15] Rodrigues Neto C and Fontanari J F 1996 *J. Phys. A: Math. Gen.* **29** 3041
[16] Coolen A C C and Sherrington D 1993 *Phys. Rev. Lett.* **71** 3886
[17] Coolen A C C and Sherrington D 1994 *Phys. Rev.* E **49** 1921
[18] Gardner E 1986 *J. Phys. A: Math. Gen.* **19** L1047
[19] Treves A and Amit D J 1988 *J. Phys. A: Math. Gen.* **21** 3155
[20] Peretto P 1984 *Biol. Cybern.* **50** 51