Pattern selectivity and binary-synapse neural networks

1993 J. Phys. A: Math. Gen. 26 4479

(http://iopscience.iop.org/0305-4470/26/17/052)

J. Phys. A: Math. Gen. 26 (1993) 4479–4483. Printed in the UK

ADDENDUM

Pattern selectivity and binary-synapse neural networks

R W Penney and D Sherrington

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

Received 27 May 1993

Abstract. Using the Rau, Wong and Sherrington model of the storage of a pair of correlated patterns by an optimized synaptic neural network, we argue that all maximally-stable neural networks will show high selectivity between correlated memorized patterns. The dilute-retrieval phase diagram is calculated specifically for a network with binary-valued synapses, and is shown to be qualitatively identical to that for the spherically-constrained synapses of the original work.

1. Introduction

In attempting to address the relevant question of the response of neural memories to stochastically correlated pieces of information, particularly in terms of their ability to discriminate between them, Rau, Wong and Sherrington (1992, 1993, to be referred to as RWS) proposed a model in which two correlated patterns were stored amidst a large number of background memories. They considered three types of learning rules for the synapses (all achieved by optimization of appropriate performance measures), the Hebb, pseudo-inverse and maximally-stable rules. In all cases the synaptic weights were permitted maximum freedom, subject only to a normalization constraint. The selectivity predicted for the Hebb and pseudo-inverse rules was compatible with expectations, but the maximally stable network (MSN) was predicted, counter-intuitively, to show confused retrieval attractors only for *small* pattern correlation, and within very narrow ranges of total memory loading. In this comment we will indicate that the absence of confusion between the two correlated patterns when strongly correlated, is a feature to be expected of all types of MSN, whatever the nature of their synapses. As a demonstration of this, we will briefly present results for a network with binary-valued synapses, representing a highly local synaptic constraint which contrasts strongly with that used by Rau *et al.*

2. Overview of the model

We consider a large system of N formal neurons, $S_i \in \{\pm 1\}$, $i \in \{1, \ldots, N\}$, sparsely linked by asymmetric synapses, J_{ij}, and having discrete-time dynamics $S_i(t+1) = \mathrm{sgn}\big(C^{-1/2}\sum_j J_{ij}S_j(t)\big)$, where $C\,(\leqslant N)$ is the average connectivity of the network. In order to simplify the retrieval dynamics, we idealise to an extremely diluted network having $\ln N \gg C \gg 1$ (Derrida *et al* 1987). For the maximally-stable rule, we require $\alpha C + 2$ random neuron-states, $\xi_i^\mu \in \{\pm 1\}$ $\mu \in \big\{1, \ldots, (\alpha C + 2)\big\}$, to be stable fixed-points of the update rule, such that all aligning fields, $\Lambda_i^\mu = N^{-1/2}\sum_j \xi_i^\mu J_{ij}\xi_j^\mu$, exceed a positive threshold, κ, understood to be maximized for the given loading, α. The model of RWS considers patterns $\mu = 1, 2$ to be mutually correlated, but uncorrelated with the remaining, mutually uncorrelated, memories. The two special memories are used to partition the C neurons feeding a given site, i, such that $\mathcal{S} = \{j \mid \xi_j^1 = \xi_j^2\}$ and $\mathcal{D} = \{j \mid \xi_j^1 = -\xi_j^2\}$,which quantities relate to the pattern correlation, Q, as follows:

$$Q = \frac{1}{C}\sum_{j=1}^{C}\xi_j^1\xi_j^2 \qquad Q_\pm = \frac{1}{2}(1 \pm Q) = \begin{cases} C^{-1}|\mathcal{S}| & (+) \\ C^{-1}|\mathcal{D}| & (-) \end{cases} \tag{2.1}$$

For each of these two sets of neurons, retrieval order parameters and contributions to the pattern stability fields are defined

$$m_{s/d} = \frac{1}{CQ_\pm}\sum_{j\in\mathcal{S}/\mathcal{D}}\xi_j^1 S_j \qquad \Lambda_{s/d} = \frac{1}{\sqrt{CQ_\pm}}\sum_{j\in\mathcal{S}/\mathcal{D}}\xi_i^1 J_{ij}\xi_j^1\,. \tag{2.2}$$

Confused retrieval of the correlated patterns would be signalled by $m_s > 0$ and $m_d = 0$, whilst perfect retrieval is implied by $m_s = m_d = 1$. The aligning-fields of the two special memories are given by

$$\Lambda^1 = \left[\sqrt{Q_+}\Lambda_s + \sqrt{Q_-}\Lambda_d\right] \qquad \Lambda^2_\pm = \pm\left[\sqrt{Q_+}\Lambda_s - \sqrt{Q_-}\Lambda_d\right] \tag{2.3}$$

according to whether the target neuron is with the set $\mathcal{S}$ (Λ^2_+) or $\mathcal{D}$ (Λ^2_-) of a subsequent neuron. Each of these stabilities, together with those of all background patterns, must exceed κ, for an MSN. In the extremely diluted network that is assumed, the retrieval dynamics, for sequential spin-update, may be written as a flow

$$\frac{\mathrm{d}m_{s/d}}{\mathrm{d}t} = \int \mathrm{d}\Lambda_s\,\mathrm{d}\Lambda_d\,\rho_{s/d}(\Lambda_s, \Lambda_d)\;\mathrm{erf}\left(\frac{\sqrt{Q_+}\Lambda_s m_s + \sqrt{Q_-}\Lambda_d m_d}{\sqrt{2(1 - Q_+m_s^2 - Q_-m_d^2)}}\right) - m_{s/d} \tag{2.4}$$

dependent on pattern aligning field distributions $\rho_{s/d}(\Lambda_s, \Lambda_d)$, expressions for which may be derived using replica mean-field theory after specifying the nature of the synapses.

The flow (2.4) is expected to exist in a number of forms, and transitions between these occur for various network loadings, α, and correlations, Q. The four phase-boundaries thus defined are determined by changes in stability of the non-retrieval fixed-point, $m_s = m_d = 0$, and of fixed-points on the m_s axis. Details of these conditions can be found in Rau *et al* 1993.

3. General results

By considering the flow (2.4) in the strong correlation limit $Q \to 1$ ($Q_+ \to 1$, $Q_- \to 0$), the absence of confused attractors, in maximally stable networks, can be demonstrated. (Here, Q_- is taken as being of order C^0, i.e. the number of sites on which the two special patterns differ is still extensive.) Fixed-points of the flow that lie on the m_s axis must satisfy the condition

$$m = \left\langle \mathrm{erf}\left(\frac{\sqrt{Q_+}m\Lambda_s}{\sqrt{2(1 - Q_+m^2)}}\right)\right\rangle_s \equiv \phi(m) \tag{3.1}$$

in which the short-hand notation $\langle f(\Lambda_s, \Lambda_d)\rangle_{s/d} = \int \mathrm{d}\Lambda_s\,\mathrm{d}\Lambda_d\,\rho_{s/d}(\Lambda_s, \Lambda_d)\,f(\Lambda_s, \Lambda_d)$ has been used. By definition of an MSN, $\rho_s(\Lambda_s, \Lambda_d) = 0 \quad \forall\Lambda_s < \kappa$, implying that in the limit $Q \to 1$, the non-zero solution of (3.1) that is stable with respect to small fluctuations along the m_s axis, m^*, must approach $m^* = 1$. Moreover, the evolution of m_d around this fixed point can be characterized readily. By identifying the lower bound of $\{\sqrt{Q_+}\Lambda_s m_s + \sqrt{Q_-}\Lambda_d m_d\}$ within the domain where $\rho_d(\Lambda_s, \Lambda_d) > 0$, one finds that

$$\frac{\mathrm{d}m_d}{\mathrm{d}t} > m_d\left\{\frac{2}{\sqrt{\pi}}\frac{\kappa}{\sqrt{2(1-Q_+m^{*2})}} - 1\right\} \tag{3.2}$$

implying that this point is manifestly unstable with respect to increasing $|m_d|$ provided $\kappa > \sqrt{\pi/2}\sqrt{1-Q_+m^{*2}}$. Given that $(1-Q_+m^{*2}) \to 0$ as $Q \to 1$, it is seen that this root of the flow cannot be a confused attractor unless $\kappa \to 0$, which is not appropriate for a maximally stable network within its storage capacity limit.

By considering the number of inflection points of the function $\phi(m)$ for $m \in (0, 1)$, one may show that for any distribution $\rho(\Lambda_s)$ such that $\rho(\Lambda_s) = 0 \quad \forall \Lambda_s < \kappa > 0$, there are no more than two non-zero roots of $m = \phi(m)$ within $(0, 1)$. This implies that no confused attractors can exist elsewhere on the m_s axis, since any other finite root is automatically unstable along the m_s direction under the flow. This observation is compatible with the forms of flow predicted to exist towards $Q = 1$ by Rau *et al*. Therefore, it would appear that dilute maximally-stable networks, in generality, entirely avoid confusion for pairs of strongly correlated patterns, and that the looseness of the spherical synapse-constraint $(\sum_j J_{ij}^2 = C \quad \forall i)$ used in RWS is not central to achieving this selectivity.

More generally, for the whole range of Q, examination of the self-consistency of permutations of the phase boundaries pertinent to the model of RWS, together with the inherent properties of the flow (2.4), leads us to suspect that a unique topology of dilute-retrieval phase diagram (in α, Q space) is possible for networks which avoid confusion towards strong correlation (i.e. that topology found by Rau *et al*). Moreover, this procedure suggests that phases exhibiting confused attractors towards weak correlation ($Q \to 0$) are inevitable, despite the surprising contrast with their absence for greater pattern similarity.

4. The binary-synapse network

In support of the implications of the previous section, we briefly consider the RWS model for networks with binary-valued synapses, $J_{ij} \in \{\pm 1\}$. The calculation of $\rho_{s/d}(\Lambda_s, \Lambda_d)$ may be effected as a generalization of the methods of Rau *et al* (1993) and Gardner (1989). However, the peculiarities of the binary model's weight-space, as compared with that of the spherical model, mean that $\rho_{s/d}$ cannot be expressed in as concise a form as was possible for the original model of RWS (cf Penney and Sherrington 1993). In contrast to the combination of δ-functions and Gaussian tails shown by the spherical model, $\rho_{s/d}$ here has the form

$$\rho_{s/d}(\Lambda_s, \Lambda_d) = \theta(\Lambda^1 - \kappa)\,\theta(\Lambda^2_\pm - \kappa) \times \frac{\mathrm{e}^{-\frac{1}{2}\Lambda_s^2}}{\sqrt{2\pi}}\frac{\mathrm{e}^{-\frac{1}{2}\Lambda_d^2}}{\sqrt{2\pi}} \int \mathrm{D}u_s\,\mathrm{D}u_d \left\{\int_{\mathcal{R}_{s/d}} \mathrm{D}y_s\,\mathrm{D}y_d\right\}^{-1} \tag{4.1}$$

$\mathcal{R}_{s/d}$ being defined by

$$\begin{aligned} &\sqrt{Q_+}y_s + \sqrt{Q_-}y_d > \frac{1}{\sqrt{1-q}}\cdot(\kappa - q\Lambda^1) + \sqrt{q}(\sqrt{Q_+}u_s + \sqrt{Q_-}u_d) \\ &\pm(\sqrt{Q_+}y_s - \sqrt{Q_-}y_d) > \frac{1}{\sqrt{1-q}}\cdot(\kappa - q\Lambda^2_\pm) \pm \sqrt{q}(\sqrt{Q_+}u_s - \sqrt{Q_-}u_d) \end{aligned} \tag{4.2}$$

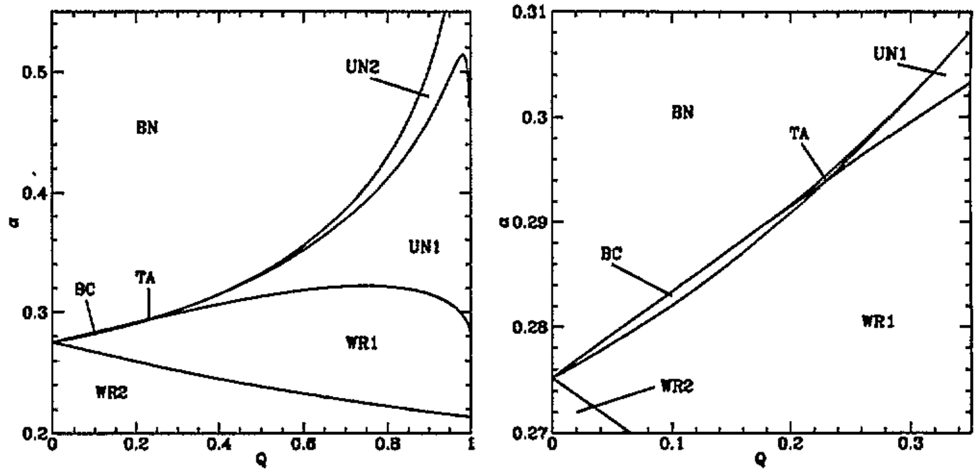

Figure 1. The dilute retrieval phase diagram for the binary-synapse neural network storing two correlated patterns. The various types of retrieval flow are labelled in the notation of Rau *et al*. The confused phases (BC and TA) are show magnified in the graph on the right.

in which q is an order parameter of the mean-field theory, determined by the network loading α, exactly as for the binary perceptron storing uncorrelated patterns (Krauth and Mézard 1989). Using these distributions in place of the original expressions, we may seek solutions of the phase-boundary conditions (Rau *et al* 1993), a somewhat weighty numerical task that leads to a phase diagram shown in figure 1.

It is seen that imposing binary synapses on a maximally stable network does not qualitatively alter its dilute retrieval phase diagram for this model (or indeed for the retrieval of uncorrelated patterns, according to similar comparisons), and regions of parameter space where confused attractors exist are seen to be small and localized near $Q = 0$, as found by RWS.

5. Conclusion

A simple analysis of the maximally stable network storing strongly correlated patterns, in the regime of the model of Rau, Wong and Sherrington, has shown that high pattern selectivity is to be expected of all large MSNs storing two correlated patterns against an extensive background of uncorrelated patterns. This assertion is supported by calculations applied to a binary-synapse network, whose dilute-retrieval phase diagram has been presented.

In addition, we have considered a toy model of pattern selectivity, in which all αC stored patterns are mutually correlated (by biasing them according to $p(\xi_i^\mu) = \frac{1}{2}(1 + m\xi_i^\mu)$). The synapses of this model are subject to a spherical constraint, as for the original model of RWS, and chosen according to the maximally stable rule (cf Gardner 1988). This model shows a qualitatively identical phase diagram, also with tiny confused phases, and an absence of confusion under conditions of strong correlation. Therefore, it would appear that exacting separation of correlated memories is a feature to be expected of all types of MSN, under very general conditions.

The issue of extreme pattern correlation, for which $Q_- \sim C^{-x}$ with $0 < x < 1$, has not been addressed explicitly here, but the analysis of section 2 would suggest that provided positive stability can be granted to all patterns by a network, confused attractors are not to be expected. It is in the abilities of differing species of network to achieve this stabilization that we expect the detailed nature of the synapses to show greatest effect.

(This work is anticipated to form part of the DPhil thesis of RWP, where more details of the methods and analyses, summarized here, may be found.)

Acknowledgments

We would like to thank the many people who have offered advice while this work was in progress, amongst whom Drs Marc Mézard, Albrecht Rau and Michael Wong deserve particular gratitude. RWP would like to thank Jesus College, Oxford, for the generous award of a scholarship. We gratefully acknowledge financial support from the SERC (under grant number 9130068X).

References

Derrida B, Gardner E J and Zippelius A 1987 *Europhys. Lett.* **4** 167

Gardner E J 1988 *J. Phys. A: Math. Gen.* **21** 257

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

Krauth K and Mézard M 1989 *J. Physique* **50** 3057

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

Rau A, Wong K Y M and Sherrington D 1992 *Europhys. Lett.* **17** 649

—— 1993 Pattern selectivity in neural networks as a means of understanding basin structures *J. Phys. A: Math. Gen.* **26** 2901