Analysis of the effects of lesions on a perceptron

1989 J. Phys. A: Math. Gen. 22 2227

(http://iopscience.iop.org/0305-4470/22/12/021)

J. Phys. A: Math. Gen. **22** (1989) 2227-2232. Printed in the UK

Analysis of the effects of lesions on a perceptron

M A Virasoro

Dipartimento di Fisica, Università di Roma 'La Sapienza', Piazzale Aldo Moro 2, I-00185, Roma, Italy, and INFN, Sezione di Roma

Received 21 February 1989

Abstract. Gardner's analysis is used to study the evolution of a perceptron trained on a rule-controlled mapping with exceptions after it has gone through an irreversible random process of deterioration or lesion. It is shown that entropy considerations lead to useful statistical inferences based on partial information about the consequences of the lesions. In particular it is shown that patterns that follow the rule are more robust than patterns that have an exceptional response.

1. Introduction

Gardner's analytical calculation of the available volume in the phase space of interactions (Gardner 1988) has had a deep influence on our way of thinking. Her analysis demonstrated one possible way of handling new difficult concepts like the idea of the *entropy* of a neural network (Carnevali and Patarnello 1987, Denker *et al* 1987).

As my modest contribution to honour the memory of such an extraordinary woman I wish to show a simple example of how her calculation can be used in this way. In statistical mechanics growth of entropy controls the direction of evolution of a system. I will show how similar ideas can be used to analyse the fate of a network which after having learned a certain task is lesioned in a random way.

In a previous letter (Virasoro 1988, hereafter referred to as V88) a network that has stored categorised patterns reacting after partial, random destruction of synapses was studied. Here I will show that the same result can be obtained solely from entropy considerations. The advantage of this new way of looking at the problem is that it allows for immediate generalisation to more general cases. The limitations will be discussed at the end.

Just for the sake of novelty I will consider an apparently different kind of task for the system, but one that will turn out to be essentially isomorphic to the memorisation of categorised patterns. Let me assume that a feedforward two-layer system (a perceptron) is trained to map certain inputs to certain outputs but that, in the mapping presented, most of the corresponding pairs obey a simple rule except in a certain number of cases where the rule is not obeyed and the response is exceptional.

The main conclusion will be that after the lesion if one looks at the ultimate fate of an equal number of regular and exceptional cases one will find that the regular patterns are more robust: the probability for one particular regular pattern to fail is smaller than the corresponding probability for an exceptional one. This is reminiscent of the general phenomenon of *regularisation* widely observed in patients.

0305-4470/89/122227+06$02.50

2. The model

Let me consider a perceptron with N_0 input units and for simplicity a single output unit. For the sake of concreteness let me imagine that the machine is supposed to learn to read in a language where pronunciation is rather simple: to each letter there corresponds just one phoneme except when that letter is in a particular context defined by a string of characters which belong to a predefined set of possible exceptions.

These exceptional strings of characters are chosen at random. Therefore in this simple language the difference between rule-governed behaviour and exceptional behaviour is maximal. In a real language the situation is more complicated because the exceptions in general show subregularities, i.e. they obey rules of a higher level of complexity. I believe that the results I am going to derive apply to this more realistic case.

The central character is coded by a finite number F of neurons. The context will be assumed to include N neurons where N will eventually tend to infinity. It will follow from this analysis that the number of exceptions that the system can learn is strictly smaller than N and therefore can be much smaller than the number of regular responses.

The response function is

$$s^\alpha = \operatorname{sgn}\left(\sum_{k=1}^{F} U_k \xi_k^\alpha + \sum_{i=1}^{N} J_i \xi_i^\alpha \right) \tag{1}$$

where ξ_k^α is the kth bit ($k = 1, N+F$) in the αth ($\alpha = 1, P$) input pattern, U_k ($k = 1, F$) are the synapses converging from the F central input neurons, while J_i ($i = 1, N$) are the synapses converging from all the other ones.

If there were no exceptions then

$$J_i \simeq 0 \qquad \sum_{k=1}^{F} U_k \xi_k^\alpha - M s_R^\alpha \qquad M > 0 \tag{2}$$

where the subscript R reminds us that by definition it is the regular response. But there are exceptions; i.e. among the P patterns there is a set R of P_R regular patterns but there is also another set I with P_I patterns for which the regular response is not correct and must be superseded by an irregular one. In order to learn such a situation the system must decrease M and make J_i different from zero so that

$$s^\alpha = \operatorname{sgn}\left(M s_R^\alpha + \sum_{i=1}^{N} J_i \xi_i^\alpha \right) \tag{3}$$

or equivalently:

$$\begin{aligned} -s_R^\alpha \sum_{i=1}^{N} J_i \xi_i^\alpha &< M \qquad \text{for regular patterns} \\ -s_R^\alpha \sum_{i=1}^{N} J_i \xi_i^\alpha &> M \qquad \text{for exceptional patterns.} \end{aligned} \tag{4}$$

The calculation 'a la Gardner' of the volume V in J_i, M space such that equations (4) are satisfied is straightforward. Notice that the natural normalisation condition is

$$\delta\left(\sum_{i=1}^{N} (J_i)^2 + \frac{M^2}{N} - 1 \right) \tag{5}$$

so that the central synapses are typically larger than the J_i. This is a direct consequence of the task imposed on the network: the signal from the central character and the one

from the contextual string must be roughly of the same order. Using the same notations as in Gardner (1987):

$$\langle \ln V\rangle = \left\langle \ln \int \left(\prod_i \mathrm{d}J_i\right) \mathrm{d}M \prod_{\alpha\in R} \theta\left(s_R^\alpha \sum_{i=1}^N J_i\xi_i^\alpha + M\right) \right.$$
$$\left. \prod_{\alpha\in I} \theta\left(-s_R^\alpha \sum_{i=1}^N J_i\xi_i^\alpha - M\right)\theta(U)\delta\left(\sum_{i=1}^N (J_i)^2 + \frac{M^2}{N-1}\right)\right\rangle \tag{6}$$

while the final result becomes

$$s = \langle \ln V\rangle = (N/2)[\ln(1-q) + q/(1-q)]$$
$$+ P_R \int Dt \ln H((-M + t\sqrt{q})/\sqrt{1-q})$$
$$+ P_I \int Dt \ln H((M + t\sqrt{q})/\sqrt{1-q}) \tag{7}$$

where P_R (P_I) is the number of regular (exceptional) cases learned,

$$H(x) = \int_x^\infty \frac{\mathrm{e}^{-t^2/2}}{\sqrt{2\pi}} \tag{8}$$

and Dt is the gaussian measure $\mathrm{e}^{-t^2/2}\,\mathrm{d}t/\sqrt{2\pi}$. The variables M, q are determined from the saddle point equations

$$\partial S/\partial U = \partial S/\partial q = 0. \tag{9}$$

This result is identical to the one derived for correlated patterns once the parameters are conveniently translated. When $P_R \gg P_I$, M will tend to infinity. For $q \neq 1$

$$S = (N/2)[\ln(1-q) + q/(1-q)] - P_R(1/M)\sqrt{(1-q)2\pi}\,\mathrm{e}^{-M^2/2} - P_I M^2/2(1-q) \tag{10}$$

while equations (9) imply:

$$-\tfrac{1}{2}M^2 = \ln(P_I/P_R) + \tfrac{1}{2}\ln[2\ln(P_R/P_I)] + \mathrm{O}(1) \tag{11}$$

$$q = (2P_I/N)\ln(P_R/P_I) + \mathrm{O}(1). \tag{12}$$

It is noteworthy that the number of exceptions P_I can at most grow linearly with N while P_R can grow exponentially.

3. Entropy per pattern

Perhaps the farthest reaching consequence of Gardner's analysis is that it de-emphasises the choice of a learning procedure. S is an entropy decrease due to learning but rather than applying to one particular type of learning it can be better understood as a lower bound on the entropy decrease necessary to store the given patterns. How relevant this is from a biological point of view can be argued forever, particularly given the present day situation when there is no candidate for a biologically relevant learning mechanism. Eventually, when this mechanism is discovered, one will have to examine whether it is ergodic with respect to Gardner's measure. Among those proposed in the literature there are some that are not ergodic. However in V88 we have already argued that they have to pay a cost for that: for instance, their storage capacity is not

optimal (Amit *et al* 1989). It was also argued that any learning rule able to reach the maximal storage capacity and near the saturation of this capacity will lead the system towards a configuration whose properties can be obtained from Gardner's approach.

If this is the case, S is the physical entropy of the network after learning and we can use it as in statistical mechanics to characterise the evolution of the system when lesioned or deteriorated in an irreversible way. Entropy considerations provide a concise language in which to take into account the different availability of phase space.

The entropy in equations (7) and (10) can be analysed as a function of the parameters. The derivatives $\partial S/\partial P_R$ and $\partial S/\partial P_I$ are the average entropy decrease per regular and exceptional patterns. More precisely, if a system with P_R, P_I patterns has to learn a new pattern, the additional decrease of entropy will be different according to which pattern is added. As the latter are generated with a certain probability distribution then we can talk of a probability distribution of the additional entropy. The average values are the above-mentioned derivatives but the fluctuations can also be calculated.

Let me define as ΔV the decrement in available volume when the new pattern has been stored. The moments of the distribution:

$$\langle (1-\Delta V/V)^m \rangle \tag{13}$$

can be calculated by the standard replica method or more directly using the cavity method (Mézard 1989). For future discussions it is convenient to consider a slightly more general quantity $\Delta_\kappa V$ which corresponds to a storage with a stability parameter κ. It is obtained by changing the constraints on the J introduced by the new pattern so that instead of equations (4) we consider

$$\begin{aligned} -s_R^\alpha \sum_{i=1}^{N} J_i \xi_i^\alpha &< M-\kappa \qquad \text{for regular patterns} \\ -s_R^\alpha \sum_{i=1}^{N} J_i \xi_i^\alpha &> M+\kappa \qquad \text{for exceptional patterns.} \end{aligned} \tag{14}$$

The result for (13) is

$$\langle (1-\Delta_\kappa V/V)^m \rangle = \begin{cases} \displaystyle\int Dt[H((-M+t\sqrt{q}+\kappa)/\sqrt{1-q})]^m & \text{for a regular pattern} \\ \displaystyle\int Dt[H((M+t\sqrt{q}+\kappa)/\sqrt{1-q})]^m & \text{for an exceptional pattern.} \end{cases} \tag{15}$$

It is now possible to reconstruct the probability distribution of $-\Delta_\kappa V/V$ from the moments

$$P_{(I)}^{(R)}(\Delta_\kappa V/V) = \int Dt\, \delta[\Delta_\kappa V/V - 1 + H(-(\pm M - t\sqrt{q}-\kappa)/\sqrt{1-q})]. \tag{16}$$

4. Effects of lesions

Suppose now that the lesion has occurred and the only thing known about it is that it has provoked the loss of a total number ΔP of patterns. For technical reasons ΔP is supposed to be small so that the state of the system has only been slightly perturbed and the values of U, q can be considered unchanged. I propose to use the maximum entropy principle to estimate other details about the way the system has evolved.

The first patterns to be lost will be the ones for which ΔV is larger. Let me work on the region $P_R \gg P_I$ where the analysis can be done explicitly. From (16) for $\kappa = 0$ it is easy to see that for regular patterns the distribution P_R is concentrated around $\Delta V/V \simeq 0$ while for the exceptional ones it is concentrated around $\Delta V/V \simeq 1$. If one checks the failures among an equal number of regular and exceptional patterns it follows that the number of regular patterns in that group that are lost will be exponentially small: i.e. the probability of a regular pattern failure will be exponentially negligible. This is the meaning of *regularisation.*

However, this should not be read as implying that no regular patterns are lost. On the contrary the probability of a regular pattern failure is exponentially small but the total number of regular patterns is exponentially large so that the leading divergences compensate. A detailed analysis yields that the total number of regular pattern failures and the total number of irregular pattern failures are roughly of the same order.

A detailed comparison between these results and the analysis in V88 shows both the limitations and the power of this approach. The conclusions (in particular all those stressed in both papers) are qualitatively the same but details differ. The reason lies in the fact that in V88 a detailed model for the process of lesion was assumed while here it is treated as a black-box process subject to some kind of statistical inference based on maximum entropy constrained by our information. In other words we are using entropy, as in Shannon, to complement our lack of knowledge.

In fact in V88 if a pattern was lost during the lesion it did not mean that all traces of it disappear from the couplings J. Instead it meant that the corresponding constraint became:

$$\theta\left(s^\alpha\left(\sum_j J_j\xi_j^\alpha + Ms_R^\alpha\right)\right) \to \theta\left(s^\alpha\left(\sum_j J_j\xi_j^\alpha + Us_R^\alpha\right) + \kappa\right) \tag{17}$$

with κ small and positive. With this modification in mind one can reapply the entropy argument. The volume decrement per pattern is now:

$$\delta S = (\Delta V - \Delta_\kappa V)/(V - \Delta V) \tag{18}$$

and the probability distribution becomes:

$$P_{(\uparrow)}^{(R)}(\delta S) = \int Dt\delta\left[\delta S - \left(\frac{H(-(\pm M - t\sqrt{q} - k)/\sqrt{1-q}) - H(-(\pm M - t\sqrt{q})/\sqrt{1-q})}{H(-(\pm M - t\sqrt{q})/\sqrt{1-q})}\right)\right]. \tag{19}$$

This distribution exactly coincides with the distribution of the stability field that appears in V88. From there on, both analyses will necessarily coincide.

It is clear that in all statistical inferences based on maximum likelihood the results depend on our information about the process. This is a clear limitation and should always be kept in mind particularly when comparing the predictions with experimental data (for instance those coming from patients) or simulations.

However, the enormous advantage of entropy considerations is that they can be easily generalised. Abstracting from this type of argument to that which necessarily will be valid in more general cases is easier. For instance, the fact that the entropy decrement per regular pattern is much smaller than the entropy decrement per exceptional one can be seen as a direct consequence of the definition of regularity. Therefore it is reasonable to expect that phenomena like *regularisation* will in general be present.

Acknowledgments

I thank Tim Shallice for useful conversations on the regularisation phenomena in neuropsychology. A fellowship from the Guggenheim foundation is gratefully acknowledged.

References

Amit D, Franz S and Virasoro M A 1989 to be published
Carnevalli P and Patarnello S 1987 *Europhys. Lett.* **4** 1199
Denker J, Schwartz D, Wittner B, Sollas S, Hopfield J J, Howard R and Jackel L 1987 *Complex Systems* **1** 877
Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
Mézard M 1989 *J. Phys. A: Math. Gen.* **22** 2181
Virasoro M A 1988 *Europhys. Lett.* **7** 293