International Journal of Modern Physics C

NEURAL NETWORKS AS OPTIMAL INFORMATION PROCESSORS

ALESSANDRO CAMPA, PAOLO DEL GIUDICE
Physics Laboratory, Istituto Superiore di Sanità
and INFN Sezione Sanità
Viale Regina Elena 299, 00161 Roma, Italy
e-mail: campa@vaxsan.iss.infn.it giudice@vaxsan.iss.infn.it

JEAN-PIERRE NADAL
Laboratoire de Physique Statistique
(Laboratoire associé au C.N.R.S. (U.R.A. 1306) et aux Universités Paris VI et Paris VII)
Ecole Normale Supérieure
24, rue Lhomond, 75231 Paris Cedex 05, France
e-mail: nadal@physique.ens.fr

NESTOR PARGA
Departamento de Fisica Teorica, Universidad Autonoma de Madrid
Ciudad Universitaria de Cantoblanco, 28049 Madrid, Spain
e-mail: parga@vaxrom.roma1.infn.it

ABSTRACT

We explore the properties of a feed-forward neural network whose couplings are chosen in such a way as to maximize the input-output mutual information, in the case in which the input-output channel is affected by noise.

Keywords: Neural Networks, Information Theory, Signal Analysis.

1. Introduction

Even if not directly related to WWW and its applications, the work we briefly report on in this paper lies in the general frame underlying the theme of the workshop, in that it deals with the problem of information processing. In particular, the subject of this short account is the analysis of a neural network as an information processor; while no applications will be considered in what follows, the results obtained in this kind of approach can be directly relevant to problems such as information compression, signal analysis and many others.

Neural networks have become a widely used computing tool in the context of data analyzing and processing, as well as a promising theoretical framework for the modelling of neural processing in the cerebral cortex.

It is customary to classify these models as feed forward or attractor neural networks, according to their architectural and dynamical characteristics. We will not review here the subject,[a] assuming that the reader is familiar with the basics of neural network theory; we only point out, just to define the context for our work, that we will consider here feed forward networks with unsupervised learning.

It has become progressively clear that meaningful quantities characterizing the performance of a network can be borrowed from information theory. While information has a precise mathematical definition, the essence of its meaning can be understood intuitively in this context quite easily. In a feed forward network a given output pattern can be in general the result of different input patterns, and therefore the question arises of the knowledge one gets from the output about the input; the smaller the uncertainty about the input, given the output, the larger the knowledge one gets on the input, and the mathematical information gives a measure of the quality of this knowledge. Also in attractor networks it has long been recognized that, more than the number of different stored patterns, the relevant quantity is the number of bits needed to specify them, which is therefore the number of bits actually stored in the network; for example, this number can be much higher for few uncorrelated patterns than for many highly correlated patterns. Again, the theory of information provides a precise definition for this.

The use of information theory in feed forward neural networks modelling the processing of data flow coming from the external world, possibly in a multistage data processing, has already been considered by several authors.[3–8]

2. Information

We give here some definitions in information theory, without any attempt of completeness, mainly to establish our notations; we refer the interested reader, for example, to the book by Cover and Thomas.[9] In the formulas below we restrict ourselves to random variables that take on only discrete values, with a given probability distribution; one has to be careful to extend the definitions to continuous variables, but our relevant quantity, the mutual information, is well defined also in this case.

Given a random variable x that can take on some discrete values x_1, ..., x_n with probabilities $P(x_1)$, ..., $P(x_n)$, we denote by X the set of the possible values x_i. Then the following quantity defines the entropy H:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{2.1}$$

where the base of the logarithm defines the unit of H; usually one assumes base 2, so the entropy is measured in bits. The entropy is the average value of the random variable $-\log P(x)$. The nonnegative quantity $-\log P(x_i)$ is interpreted as

[a]For a general introduction to the subject, we refer the reader to Refs. 1,2.

the amount of information required to specify that the variables x has taken on the value x_i, and it is called the self-information of x_i. This is intuitively satisfying, since the smaller $P(x_i)$ the larger the self-information; on the other hand, in the limit $P(x_i) = 1$ the self-information vanishes, since we need not any information to specify the occurrence of an event that is certain. In conclusion, the entropy is the average value of the self-information. It is relevant to consider the case in which we have events specified by the values of two random variables, x and y, and we are interested in what we can learn, from the knowledge of the value of one variable, about the other. Thus, with X the set of possible values of x (x_1, ..., x_n) and with Y the set of possible values of y (y_1, ..., y_m) an event is specified by the couple (x_i, y_j), which occurs with the joint probability distribution $P(x_i, y_j)$. The probability of occurrence of a value of x, regardless of the value of y, is given by $P(x_i) = \sum_{j=1}^{m} P(x_i, y_j)$, $i = 1, \ldots, n$, and analogously for the occurrence of a value of y regardless of that of x, i.e., $P(y_j) = \sum_{i=1}^{n} P(x_i, y_j)$, $j = 1, \ldots, m$. To avoid burdening the notation, we have used here, as we will do troughout the paper, the same symbol P for different probability distributions. Then the mutual information provided about the occurrence of $x = x_i$ by the occurrence of $y = y_j$, or, symmetrically, provided about the occurrence of $y = y_j$ by the occurrence of $x = x_i$ is defined by:

$$I(x_i, y_j) = \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}. \tag{2.2}$$

Its average value is called the average mutual information (or mutual information for short):

$$I(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}, \tag{2.3}$$

which can be shown to be a nonnegative quantity. We just point out that, as one expects, for x and y independent one has $I(X, Y) = 0$, since in that case $P(x_i, y_j) = P(x_i)P(y_j)$.

For continuous random variables, which are characterized by probability density functions $P(x, y)$, $P(x)$ and $P(y)$, the mutual information is given by:

$$I(X, Y) = \int dx dy P(x, y) \log \frac{P(x, y)}{P(x)P(y)}. \tag{2.4}$$

The mutual information is the relevant quantity in our study, that we present below.

3. Model and Results

We study a feed forward neural network that is supposed to simulate the way in which the stimuli coming from the external world are processed. Our network has no hidden layers; the input neurons represent the response of the "receptors" to external signals. Through the synapses connecting the input layer to the output layer we have, from the output neurons, a new representation of the external stimulus.

We consider neurons with continuous values of their activities. It is supposed that the external stimuli to the input neurons can be extracted from a given distribution function. Therefore, if we have N input neurons and we indicate their activity with $\xi_j, (j = 1, \ldots, N)$, we will have some probability density $P(\vec{\xi})$ for $\vec{\xi} \equiv (\xi_1, \ldots, \xi_N)$. The total weighted input to the $i - th$ output neuron $(i = 1, \ldots, p)$ is given by $\sum_{j=1}^{N} J_{ij}\xi_j$, where the weight J_{ij} connects the $i - th$ output neuron to the $j - th$ input neuron; we will denote in the following by J the p by N matrix with elements J_{ij}; we consider only the case $p \leq N$. The activity $\vec{V} \equiv (V_1, \ldots, V_p)$ of the output neurons, for a given $\vec{\xi}$, will in turn be distributed according to some probability density $P(\vec{V}/\vec{\xi})$, rather that being exactly determined, due to the noise present in the network. For given functions $P(\vec{\xi})$ and $P(\vec{V}/\vec{\xi})$ we can in principle compute the information as given in (2.4) (using that $P(\vec{\xi}, \vec{V}) = P(\vec{\xi})P(\vec{V}/\vec{\xi})$). We will then suppose that the optimal weights J_{ij} for the network will be those for which I is maximum.

Here we consider a case where it is possible to compute analytically the maxima of I, and we will not focus on the problem of how the network finds them, being only concerned with the analysis of their properties.

The probability distribution which characterizes the environment will be chosen as gaussian:

$$P(\vec{\xi}) = \frac{1}{\sqrt{\pi^N \det(C)}} \exp\Big\{-\sum_{i=1}^{N}\sum_{j=1}^{N} \xi_i \big(C^{-1}\big)_{ij} \xi_j\Big\}, \tag{3.1}$$

where C is the correlation matrix defined by $\frac{1}{2}C_{ij} =< \xi_i\xi_j >$. This choice can be justified on the basis of a principle of "minimum knowledge" about the environment[10] since this defines the maximum entropy probability distribution, given the correlations $< \xi_i\xi_j >$. The noisy nature of the channel is expressed through the conditional probability

$$P(\vec{V}/\vec{\xi}) = \left(\frac{1}{\pi b^2}\right)^{\frac{p}{2}} \exp\Big\{-\frac{1}{b^2}\sum_{i=1}^{p}\big(V_i - \sum_{j=1}^{N} J_{ij}\xi_j\big)^2\Big\}, \tag{3.2}$$

where the parameter b^2 characterizes tha amount of noise present in the network. Given these choices, it is possible to calculate the probability distribution of the output $P(\vec{V})$ and the mutual information I. We first define the p by p matrix $Q = b^2\ \mathbf{1} + JCJ^T$, where $\mathbf{1}$ is the unit matrix and J^T is the transpose matrix of J. Then we have:

$$P(\vec{V}) = \frac{1}{\sqrt{\pi^p \det(Q)}} \exp\Big\{-\sum_{i=1}^{p}\sum_{j=1}^{p} V_i \big(Q^{-1}\big)_{ij} V_j\Big\} \tag{3.3}$$

and

$$I = \frac{1}{2}\log\frac{\det(Q)}{b^{2p}}. \tag{3.4}$$

Notice that, as intuition suggests, $I \to 0$ as $b^2 \to \infty$. From eq. (3.4) it also follows that for $b^2 \to 0$ (noiseless channel, for which we have the deterministic relation

$V_i = \sum_{j=1}^{N} J_{ij}\xi_j$) the mutual information tends to infinity. This is related to the general fact that for continuous input variables the input entropy itself is infinite.[b] If $p = N$ and $b^2 = 0$ there is no loss of information through the channel, since once the output $\vec{V}$ is known the input $\vec{\xi}$ is uniquely detemined (the matrix J being invertible), so I is infinite. For the case $p < N$ and $b^2 = 0$, let us consider for simplicity the case $p = 1$; then the relation for the only output neuron, $V = \sum_{j=1}^{N} J_j\xi_j \equiv \vec{J}\cdot\vec{\xi}$, implies that, once V is known, the input $\vec{\xi}$ is determined apart from an arbitrary vector orthogonal to $\vec{J}$, so that, although there is an infinite loss of information through the channel, the power of this infinite is lower than that of the input information. This results again in an infinite mutual information. However, since from Eq. (3.4) we see that the part to be maximized, $\log\det(Q)$, is always finite, one can choose to give a meaning also to the case $b^2 = 0$: in this case, the representation of the input that the network builds through the J's resulting from the maximization procedure is related to the principal components of the input distribution (see below and the comments in the last Section). If, on the other hand, a noise in input is introduced[c] (which can be interpreted as the fact that the receptors coding for the input signals have a finite analog depth), this makes the mutual information finite even for a noiseless channel; this case will be considered in a paper to be published shortly.

The expression for I also shows that the mutual information is an unbounded function of the J's: the couplings can grow without limits, increasing the signal to noise ratio. Therefore I has to be maximized subjected to some limitations on the J's, which can be implemented as a constraint or, as we have done in the study presented here, introducing a penalty term, i.e., adding to I a simple function that penalizes the growth of the J's, and finding the maxima of this new function. We have chosen to add the term $(-1/2)\sum_{i=1}^{p}\sum_{j=1}^{N} J_{ij}^2$ (that can be considered as representing in a rough way a tendency to "forget"). In a gradient ascent dynamics for the J's, that can be chosen to find the maxima by computer simulation when it is not possible to make an analytical calculation, this results in the appearance of an exponential decay term in the equation for $\delta(J_{ij})$:

$$\delta(J_{ij}) = \eta\left(\frac{\partial I}{\partial J_{ij}} - J_{ij}\right) \tag{3.5}$$

where the parameter η fixes the time scale of the dynamics. However, for the case studied in this paper, we can compute the maxima, i.e., the stable zeroes of the right hand side of Eq. (3.5). We defer to a fuller account to be published an analysis of possible alternatives for the J dynamics, together with a study of the effects of input noise.

We list below a brief summary of the results obtained for the stable fixed points, without detailing the calculations. We just show the expression that gives the fixed

[b] Here we refer to the actual continuum limit of the discrete entropy and not to the so called "differential entropy" $\int d\vec{\xi} P(\vec{\xi}) \log P(\vec{\xi})$, which is finite (see, e.g., Ref. 9).

[c] This can be obtained by, e.g., adding to each ξ_i a random variable with gaussian distribution.

points:

$$Q^{-1}JC - J = 0. \tag{3.6}$$

We solve this equation and perform a stability analysis. Then we have the following situation.

$b \to 0$:
Let us define the rows of the coupling matrix J_{ij} as the set of N components vectors $\vec{J}_i, i = 1, \ldots, p$; it turns out that the $\vec{J}_i$ at the fixed point form an arbitrary orthonormal basis for a p dimensional subspace spanned by p eigenvectors of C. However, the stability analysis shows that the stable fixed points are only the $\vec{J}_i$ lying in the subspace spanned by the first p eigenvectors of C (having sorted the eigenvectors according to the value of the corresponding eigenvalues $\lambda_i, i = 1, \ldots, N$, ordered in decreasing way). This means that the $\vec{J}_i$ at the stable fixed point are related to the principal components of the input distribution[11–14]; in particular, if $p = 1$ the fixed point equations are the same as those deriving from the Oja's algorithm,[11] resulting in the projection of the input distribution on the axis of maximal variance.

b *finite*:
In this case, we denote by q the number of eigenvalues of C that are greater than b^2. Then, if $q \geq p$, for $i = 1, \ldots, p$ the vectors $\vec{J}_i$ lie in the same subspace spanned by the first p eigenvectors of C. The vectors $\vec{J}_i$ now, in general, are neither orthogonal to each other nor of unit length; the p by N matrix J at the stable fixed point can be obtained, from the p by N matrix formed by the first p eigenvectors of C, each one of square modulus $(\lambda_i - b^2)/\lambda_i, i = 1, \ldots, p$, applying from the left any p by p orthogonal matrix.[d]
If $q < p$, $(p - q)$ vectors $\vec{J}_i$ become linearly dependent on q vectors $\vec{J}_i$, which lie in the subspace spanned by the first q eigenvectors of C. The $\vec{J}_i$ can be obtained as in the case $q \geq p$, with the difference that, in the p by N matrix of the first p eigenvectors of C, the modulus of the last $(p - q)$ eigenvectors is 0.
In particular, if $q = 1$ all the $\vec{J}_i$ are parallel to the first eigenvector of C.
When $q = 0$, i.e., when b^2 is greater than the largest eigenvalue of C, the only stable solution corresponds to the vanishing of all the vectors $\vec{J}_i$.

Therefore, the effect of the noise present in the channel is that of destabilizing the lower principal components; in other words, a redundancy is introduced via the linear dependence of the $\vec{J}_i$ in the channel in order to compensate for the effects of noise.

4. Discussion

Several authors have explored the possibility of deriving from information theory design principles for the definition of "optimal" neural networks.[3–8] These developments have been mainly motivated by an attempt to clarify the possible com-

---

[d] It is to be noted that all the eigenvalues of C are positive.

putational strategies involved in the visual processing taking place in the nervous system. The idea of a "factorial code"[15,16] as a "novelty detector" with respect to correlations present in the environment, has proved to be particularly relevant in this respect, as it can be formulated as a maximum entropy principle; in the case of linear neurons, under suitable conditions on the input distribution, it is also equivalent to the projection of the input distribution on the principal components axes.

As we have seen, this in turn is related to what is obtained when maximizing the input-output mutual information; we have shown in particular what is the effect of the channel noise in such a situation, proving that it gradually destabilizes the solutions for the J's which correspond to higher and higher eigenvalues of C.

If this framework is to be interpreted as relevant to the understanding of neurobiological visual processing , the possibility is clearly relevant to formulate a learning algorithm for the development of the J's which is local, with the dynamics of J_{ij} determined only by the activities V_i and ξ_j. In this respect, while we have examined in this paper the properties of the maxima of the mutual information function, it has been suggested[6,17] that, as far as the dynamics of the J's is concerned, it is possible to reformulate the problem in such a way as to obtain a local learning algorithm, provided a suitable redefinition of the network architecture is introduced.

Several details remain to be clarified; however, a general frame seems to emerge, in which optimization principles derived from information theory appear as good candidates to serve as theoretical guidelines for neural network modelling of early visual processing.

As part of the work in progress, we mention that if "real life" conditions have to be considered, one has to resort to numerical simulations, in order to sample the input distribution characterizing the environment.

Acknowledgements

One of us (NP) thanks the kind hospitality received at the Laboratorio di Fisica, Istituto Superiore di Sanità (Rome). This work has been partially supported by INFN through the RM5 collaboration, and by the French-Spanish program "PICASSO".

References

1. J. A. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*, (Addison-Wesley, 1991).
2. D. J. Amit, *Modelling Brain Function*, (Cambridge Univ. Press, 1989).
3. J. J. Atick and A. N. Redlich, *Neur. Comp.* **2** (1990) 308.
4. J. J. Atick and A. N. Redlich, *Neur. Comp.* **4** (1992) 196.
5. R. Linsker, *Computer* **21** (1988) 105.
6. R. Linsker, *Neur. Comp.* **4** (1992) 691.
7. J. P. Nadal and N. Parga, in *Neural Networks: from Biology to High Energy Physics* (special issue of *Int. J. Neur. Syst.*), eds. O. Benhar, C. Bosio, P. Del Giudice, and M. Grandolfo, (World Scientific, 1993), pp. 41-50.

8. J. P. Nadal and N. Parga, *Network* **4** (1993) 295.
9. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, (Wiley, 1991).
10. J. J. Atick, *Network* **3** (1992) 213.
11. E. Oja, *J. Math. Biol.* **15** (1982) 267.
12. E. Oja, *Int. J. Neur. Syst.* **1** (1989) 61.
13. T. D. Sanger, *Neur. Networks* **2** (1989) 459.
14. A. Krogh and J. A. Hertz, in *Parallel Processing in Neural Systems and Computers*, eds. R. Eckmiller, G. Hartmann, and G. Hauske, (Elsevier, 1990), pp.183-186.
15. H. B. Barlow, *Neur. Comp.* **1** (1989) 295.
16. H. B. Barlow, T. P. Kaushal, and G. J. Mitchison, *Neur. Comp.* **1** (1989) 412.
17. J. J. Atick and A. N. Redlich, *Neur. Comp.* **5** (1993) 45.