International Journal of Neural Systems, Vol. 3, (Supp. 1992) 000–000
Proceedings of the Neural Networks from Biology to High Energy Physics

INFORMATION PROCESSING BY A PERCEPTRON

J.-P. Nadal
Laboratoire de Physique Statistique, Ecole Normale Supérieure, 24, rue Lhomond
F-75231 Paris Cedex 05, France

and

N. Parga *
Istituto Superiore di Sanita, Physics Laboratory, INFN Sezione Sanita, Viale Regina Elena 299
00161 Roma, Italy

Received (to be inserted
Revised by Publisher)

The information coming into a module, which is part of a global system, will in general require pre-processing that consists in building a representation - expressing it in a new code - convenient to the task to be performed by this particular module. This information, coming either from the environment or from another module in the system, will undergo this operation in what we can call an encoder. In this work we describe our results for a perceptron architecture viewed as an encoder, using encoding principles based on Information Theory. In particular we show how to evaluate the information capacity and the typical mutual information, quantities which are relevant to analize the different criteria used to code the information. Techniques taken from statistical mechanics of disordered systems will be shown to be useful for these calculations.

1. Introduction

The notion that Information Theory may provide optimization principles determining the synaptic weights $J_{i,j}$ of the neural network has been used by many authors [1 2 3 4 5]. Several different principles of this kind have been proposed. In particular the "infomax" principle of Linsker [3], maximizes the transmission of information taking place in the network. Various implementations of the search for minimal redundancy and decorrelating codes have also been carried out [1 2 4 5]. These principles specify cost functions, and define methods of unsupervised learning and self-organization: in the resulting algorithms, the couplings are modified according to the patterns presented to the input layer, but no desired output is specified.

Most of the works done so far refer only to linear output units, so little is known about the properties of information processing by non-linear neurons. Apart from this, some clarification of the relation between the different proposals seems to be necessary. Note that choosing one particular cost function is a statement on what is (might be) useful for the subsequent modules.

*Permanent address: Departamento de Física Teórica, Universidad Autónoma de Madrid, Canto Blanco, 28049 Madrid, Spain.

In this paper we discuss several aspects of elementary information processing by a neural network, in the line of [1] [3] and [5]. This processing, performed by an encoder that we will take as a neural network with a perceptron architecture, consists in finding a convenient encoding of the input information that implements one of the criteria that we mentioned before. Notice that since our output neurons are binary units this example represents an extreme case of non-linearity. We will not be concerned with the algorithmic aspects, but with the theoretical analysis of the performance of a network, comparing several optimization principles. For further details about this work the reader is referred to [6] and [13].

This work is organized as follows. In section 2 we describe the encoder as a perceptron architecture and in section 3 we show how to evaluate the maximal rate of information that it can process. In these two sections we also define some basic quantities in information theory. We begin section 4 with a discussion of different criteria for efficient coding, at the end of this section these are applied to the perceptron. Section 5 is devoted to describe how to evaluate information theoretical quantities such as the mutual information with techniques taken from the statistical mechanics of disordered systems. The conclusions are contained in section 6.

2. The Encoder

We consider the problem of processing a signal $\vec{\xi} = \{\xi_j\}_{j=1,\ldots,N}$ produced by a source with statistical structure P_ξ. In a first stage along the communication line the signal $\vec{\xi}$ is received by an encoder that produces a new code as the set $\vec{V} = \{V_i\}_{i=1,\ldots,p}$ of the activities of its p output neurons.

We take for this encoder a perceptron architecture (that is, a feedforward network with one input layer, no hidden units and one output layer) that we want to analyze as an elementary module processing information. It consists of a set of N input and p output neurons with synaptic connections $J_{i,j}$. The output neurons are taken as linear threshold elements. We will consider only a deterministic transfer function: for every output neuron i $(i = 1, ..., p)$

$$V_i = sgn(\sum_{j=1}^{N} J_{i,j}\xi_j). \tag{1}$$

The input $\vec{\xi}$ may be discrete or continuous, but in both cases, since the output is discrete the mutual information [8] [9]

$$I(\vec{V},\vec{\xi}) = \sum_{(\vec{V},\vec{\xi})} P(\vec{V},\vec{\xi}) \ln \left\{ \frac{P(\vec{V},\vec{\xi})}{P_V P_\xi} \right\} \tag{2}$$

is well defined. Here P_V and $P(\vec{V},\vec{\xi})$ are the output state probability and the joint probability of the input and output vectors. Logarithms will be always expressed in base 2, so that information quantities are measured in bits. We will consider continuous, noiseless inputs and unbiased input distributions ($<\xi_j>= 0$). Let us notice that for a deterministic system the mutual information can be written as

$$I(\vec{V},\vec{\xi}) = -\sum_{\vec{V}} P_V \ln P_V, \tag{3}$$

and that the output state probability can be expressed as:

$$P_V = \int d\vec{\xi}\, P_\xi Q(\vec{V}|\vec{\xi}) \tag{4}$$

where $Q(\vec{V}|\vec{\xi})$ is the conditional probability to find the output state $\vec{V}$ given the input pattern $\vec{\xi}$.

3. The Information Capacity

The information capacity is defined as

$$C = Max_{P_\xi} I(\vec{V}, \vec{\xi}), \tag{5}$$

and measures the maximal information that the system can transmit. From the point of view of a subsequent module receiving this output, the system acts as an encoder and C has the meaning of the maximal rate at which information can be delivered to the next module.

It is clear that the maximum condition in the capacity definition is achieved for a source such that the probabilities for all possible output states (i.e. codewords) are equal. The evaluation of the capacity is then equivalent to the calculation of the number of different possible output states. Since there are p binary output units, the maximal possible number of codewords is 2^p, so that

$$C \le p. \tag{6}$$

However not every codeword may be realizable. Each one of the p output neurons cut the N-dimensional input space with an hyperplane, so that the number of accessible output states is the number of domains obtained with p hyperplanes: the state V_i specify on which side of the ith hyperplane the input pattern lies, and $\{V_i\}_{i=1,\dots,p}$ code for one particular domain of the input space. This number of domains, $A(p, N)$, has been obtained from geometrical counting by Cover [10] in the context of supervised learning. A remarkable result is that it depends only on p and N, whenever the hyperplanes are "in general position" (that is any $k \le N$ vectors $\vec{J_i}$ are linearly independent). Then

$$A(p, N) = \sum_{l=0}^{min(p,N)} C_p^l \tag{7}$$

where C_p^l is the combinatorial number $\frac{p!}{l!(p-l)!}$. The information capacity is thus

$$C = \ln A(p, N). \tag{8}$$

From formula (7) one sees that the 2^p output sates are available up to $p = N$. Above this value there is a change in the behavior of the capacity. In the large N limit (since the information fed into the network scales with N we will measure information quantities in bits per input neuron) we have

$$\lim_{N\to\infty} C/N \equiv c = \begin{cases} \alpha & \text{if } \alpha \le 2 \\ \alpha S(1/\alpha) & \text{if } \alpha > 2 \end{cases} \tag{9}$$

Here

$$\alpha = p/N \tag{10}$$

and $S(x)$ is the entropy function (measured in bits):

$$S(x) = -[x \ln x + (1-x)\ln(1-x)]. \tag{11}$$

One sees that the fraction of unrealizable codewords for α in the interval between one and two is negligeable and goes to one above $\alpha = 2$. Note that for large α the capacity c increases as $\ln \alpha$.

4. Criteria for Efficient Coding

4.1. *Information preservation and redundancy*

Now for a given source P_ξ the mutual information depends on the couplings. It is then desirable to optimize the transmission of information by the system, searching for the best possible choice of couplings. Maximization of the mutual information [3] provides us with a reasonable criterion to determine the synaptic couplings:

$$(\mathcal{PI}) \qquad \max_{J_{ij}} I(\vec{V}, \vec{\xi}). \tag{12}$$

However this is not the only possibility to use information theory in unsupervised learning. The concept of what an efficient code is will depend on the operating conditions of the system and on the way the module receiving the coded information will use it. According to these requirements one can propose alternative coding principles. One possibility, which has been considered in several papers by Atick et al. [4], aims at reducing the redundancy of the code. Indeed, for a given input source and a given system the information I cannot be larger than the information capacity C of the system, and the redundancy

$$\mathcal{R}_C = C - I \tag{13}$$

is a measure of how inefficiently the module is being used. Then, if the information capacity were higher than the source information I_0, it might be convenient to optimize the network parameters in such a way that the redundancy $\mathcal{R}_C$ is as small as possible:

$$(\mathcal{PC}) \qquad min_{J_{ij}} \mathcal{R}_C \tag{14}$$

Now we have seen that for the perceptron the information capacity depends only on p and N (whereas for linear neurons[4], and more generally for continuous transfer functions, the information capacity depends on the couplings). Hence in our case for a given architecture (i.e. for given values of p and N), minimizing the redundancy $\mathcal{R}_{\mathcal{C}}$ is equivalent to maximizing the mutual information I, that is to applying the principle $\mathcal{PI}$.

4.2. *Barlow's proposal of redundancy reduction*

Another point of view was introduced long ago by Barlow [1 2]. Its original proposal is that factorial codes play an important role in the way the brain performs information processing He argues that the current knowledge an organism has about its environment comes from the previous observation of correlations. The realization that the occurrence of two successive events is not casual should be taken into account by proper changes in its brain that will alter the future behavior of the animal. Its ability to recognize what is new from what is old in the environment would allow it to decide to which features of a given natural scene or event to pay attention and consequently to take fast decisions. Barlow's proposal of factorial coding is a way to solve this problem: the brain would code the input visual scene in such a way that the occurrence of correlations in the coded message is a signal that something unusual is happening. This means that the complicated statistical structure of the environment, given by P_ξ, after coding is transformed into a P_V such that

$$P_V = \prod_{i=1}^{p} P_i(V_i) \tag{15}$$

i.e., a factorial or decorrelating code. Here $P_i(V_i)$ is obtained from P_V by summing over all possible states of all the other output units.

These remarks lead one to consider the mutual information conveyed by a single output neuron, independently of all the others. Let us indicate by $I_i(V_i, \vec{\xi})$ the mutual information associated with the i-th output neuron; for a deterministic system it is given by

$$I_i(V_i, \vec{\xi}) = -\sum_{(V_i)} P_i(V_i) \ln P_i(V_i). \tag{16}$$

Since [9]

$$\sum_{i=1}^{p} I_i(V_i, \vec{\xi}) \geq I(\vec{V}, \vec{\xi}) \tag{17}$$

one can define a redundancy

$$\mathcal{R}_B = \sum_{i=1}^{p} I_i - I, \tag{18}$$

that measures how far from being factorial is the code obtained at the output layer. Looking for codes with neurons as uncorrelated as possible means minimizing the redundancy $\mathcal{R}_B$

$$(\mathcal{PB}) \qquad \min_{J_{ij}} \mathcal{R}_B \tag{19}$$

where the minimization is under the constraint that I is large enough, say $I \simeq I_0$.

4.3. *Application to the perceptron*

Let us now come back to the case of the perceptron. We have thus first to consider one given output neuron i. Under the conditions of unbiased inputs and no thresholds, the hyperplane defined by the couplings $\vec{J_i}$ cuts the input N-dimensional space into two parts with statistically half of the input patterns in each of them. Hence each neuron gives one bit of information, so that

$$\sum_{i=1}^{p} I_i(V_i, \vec{\xi}) = p \tag{20}$$

Since $p \geq C \geq I$, this implies that all three principles, $\mathcal{PI}$, $\mathcal{PC}$ and $\mathcal{PB}$ are equivalent for the perceptron.

The situation can be summarized as follows: if $i_0 = I_0/N$ is finite, there is an optimal value $\alpha = \alpha_0$ for which the information capacity matches the available information, $c = i_0$. At a given value of α, the best that can be done is to maximize $i = I/N$, hopefully up to $i = c$ for α below α_0 and up to $i = i_0$ above α_0. And optimizing the architecture may allow to reach the optimal point where $i = i_0 = c$.

5. The Mutual Information and the Replica Technique

The result of the preceeding section is that one should find the couplings which maximizes the mutual information I. For p smaller than N ($\alpha \leq 1$) and for a Gaussian input distribution, it is not difficult to find the optimal solution: one has to decorrelate the inputs exactly as for a linear system [3]. The problem is much harder for p larger than N, and for a non-Gaussian distribution. In the case of a Gaussian distribution however, it is possible to replace the search for the optimal solution by a simpler problem, namely the search for the best *statistical ensemble* of couplings:

one can compute the typical mutual information per input unit, $\bar{\imath}$, when the coupling vectors are taken from an ensemble, with given correlations between their components. At the end we will look for the correlations which maximize $\bar{\imath}$. Another motivation for computing $\bar{\imath}$ is that it will give us the amount of information given by a "naive" network - that is in the absence of any optimization. This will tell us how much information can be gain by learning.

Denoting by $<< . >>$ the average over the coupling distribution ρ, the information $\bar{\imath}$ is

$$\bar{\imath} \equiv lim_{N\to\infty} << I >> /N = lim_{N\to\infty} << H(P_V) >> /N. \tag{21}$$

To evaluate this average we make use of the replica technique [7]. We will not give all the technical details, but we will point out what is specific to the present computation.

The mutual information $I(\vec{V}, \vec{\xi})$ is associated to a given input distribution P_ξ. In this paper we consider unbiased but spatially correlated Gaussian input patterns, i.e.

$$P_\xi = \frac{1}{\sqrt{(2\pi)^N detG}} \exp[-\frac{1}{2}\sum_{ij} \xi_i (G^{-1})_{ij} \xi_j] \tag{22}$$

with G the correlation matrix.

We take the coupling vectors as p independent random vectors, each one having unbiased but correlated components. However we do not need to assume any specific distribution: all what will matter in the large N limit is the first two moments of $\rho(\{J_{i,j}\}; j = 1, ..., N)$. We thus consider

$$<< J_{i,j} >>= 0 \tag{23}$$

$$<< \ J_{i,j} \ J_{i',k} >>= \delta_{i,i'} \ \Gamma_{jk}. \tag{24}$$

At the end we will look for the correlation matrix Γ which maximizes $\bar{\imath}$.

We want to compute

$$\ll i(V, \vec{\xi}) \gg = -\frac{1}{N}\sum_V \ll P_V \ln P_V \gg \tag{25}$$

in the large N limit and for a fixed ratio $\alpha = p/N$. In the language of statistical physics the $\{J_{i,j}\}$ and the $\{V_i\}$ are "quenched" variables, whereas the input patterns $\{\xi_j\}$ are "annealed" variables as can be seen from the definition of P_V in eq. (4). According to the replica method eq. (25) is written as

$$\ll I(V, \vec{\xi}) \gg = -\sum_V \ll P_V \lim_{n\to 0}((P_V)^n - 1)/n \gg, \tag{26}$$

where the small n limit is taken at the end of the calculations, after the large N limit. Due to the normalization

$$\sum_{\vec{V}} P_V = 1, \tag{27}$$

one can write also

$$\sum_V \ll [P_V]^{n+1} \gg \overset{n\to 0}{\sim} \exp(-nNi). \tag{28}$$

The computation of the left hand side of (28) follows closely the one of standard Gardner's calculations [11]. Under the "replica symmetry" ansatz, the result is as follows. Defining s by

$$s = \tau[G\,\Gamma], \tag{29}$$

with

$$\tau[.] \equiv lim_{N\to\infty}\frac{1}{N}Tr\ ., \tag{30}$$

one obtains that the asymptotic mutual information can be written as:

$$\bar{\imath} = \ll i(V,\vec{\xi}) \gg = \frac{1}{2}extr_{q,\hat{q}}\left[\hat{q}(s-q) + \tau[\ln(1-\hat{q}G\,\Gamma)] + 2\alpha\int_{-\infty}^{\infty} DzS(h)\right] \tag{31}$$

where Dz is the Gaussian measure

$$Dz = \frac{dz}{\sqrt{2\pi}}\exp(-\frac{1}{2}z^2), \tag{32}$$

and S is the entropy function defined in eq. (11), with its argument h a function of z and q given by

$$h = H(\sqrt{\frac{q}{1-q}}z) \tag{33}$$

with $H(y) = \int_y^\infty Dz$.

The saddle point equations for the order parameters q and $\hat{q}$ are

$$q - s + \tau[(1-\hat{q}G\,\Gamma)^{-1}G\,\Gamma] = 0 \tag{34}$$

$$\hat{q} - 2\alpha\frac{d}{dq}\int_{-\infty}^{\infty} DzS(h) = 0, \tag{35}$$

that give q and $\hat{q}$ as functions of α.

A few technical remarks follow. Let us outline the main steps leading to the above formulae. From the expressions (4) and (28) one has a product of $n+1$ integrals (replicas), with a product of p Θ-distributions in each of them. These Θ-distributions are written in an integral representation. Then one performs the average over the couplings according to (24). After integration over all the "microscopic" degrees of freedom, one ends up with an integral over a small number of macroscopic parameters, the "order parameters". The integrand being the exponential of N times a function of these parameters, one can apply the saddle point method. This can be done under some hypothesis on what the saddle point is. We made here the replica symmetric ansatz, which is reasonable (in analogy with Gardner's calculation for continuous couplings), and one ends up with four order parameters. A particularity of our calculation is the normalization condition (27), which means that when setting $n=0$ exactly in (28) one should find 1. This condition fixes half of the order parameters. The parameter s is one of the two parameters fixed by the $n \equiv 0$ condition. We note also that the cavity method[7] can be used, and that the term proportional to α is easily understood within this framework: it is the gain of information due to an infinitesimal increase of α.

Finally we note that, apart from the fact that we have here $n+1$ replicas, instead of the n that appear in storage capacity (Gardner type) calculations, the evaluation of $\bar{\imath}$ is done in the very same way. In fact, as explained in [6] [13], it is interesting to compare our computation with those done for the storage capacity of a perceptron with one output unit, continuous couplings, for which the task is to memorize $p = \alpha N$ input-output pairs [11]. Of particular interest is the case of correlated patterns recently studied by Monasson [12].

Let us now analyze the result. The parameter q has an interpretation: it is the value of the typical scalar product between two input patterns having a same output $\vec{V}$. When the number of output neurons is small the volume of input space associated to a given output is large, two patterns taken at random in it are statistically orthogonal: in the small α limit q goes to 0. When the number of output neurons becomes very large, the typical domain size decreases to zero, hence when α goes to infinity one finds that q goes to 1. This particular limit can be easily analyzed for any given matrices G and Γ. As q tends to one $\hat{q}$ goes to infinity and one finds the the asymptotic expression

$$\bar{\imath} \overset{\alpha\to\infty}{\sim} \ln\alpha. \tag{36}$$

In section 3 the same asymptotic behavior was found for the information capacity c. This means that, whatever the correlations in the couplings and in the inputs, the asymptotic mutual information scales exactly as the information capacity, and one finds that $c-\bar{\imath}$ goes to a constant which depends on $G\,\Gamma$. The best choice of correlations for the couplings is easily found to be

$$\Gamma = G^{-1}, \tag{37}$$

a result quite similar to the case of linear neurons in the low noise limit [4]. For this optimal choice, the mutual information is equal to its value for uncorrelated patterns and uncorrelated couplings (i.e. $G = 1$ and $\Gamma = 1$). This is not surprising: the largest possible gain of information corresponds to signals with the largest entropy.

6. Conclusions

We have shown in this work how general principles based on Information Theory can be implemented when a perceptron architecture is taken as an encoder of signals coming either from the environment or from other modules in the brain. Different principles such as maximal information transmission, minimal redundancy and decorrelating codes become, for this system, equivalent. This is a result of the peculiarity of this network architecture that the information capacity and the information conveyed by individual output neurons do not depend on the sypnaptic couplings.

When these criteria are applied to a typical (noiseless) encoder characterized by two-point correlations Γ between synapsis, the result $\Gamma = G^{-1}$ immediately follows. As in [4] the synapsis develop in such a way that they cancel the source correlations.

The evaluation of the typical mutual information of a statistical ensemble of encoders was done with the replica technique. For the case of real valued input units the replica symmetric ansatz is correct. This is expected because the domain of input states ξ producing a given codeword is connected. Note also that it respects the upper bound given by the information capacity, approaching it from below in the large α limit.

An interesting feature of the perceptron is that the deterministic rule eq.(1) defining the output neuron state $\vec{V}$ can be interpreted in two ways. One is, as in this work, that we have p output neurons where the $J_{i,j}$ are the couplings and the $\vec{\xi}$ are the input patterns. But one can as well say that we have a perceptron with only one output neuron, $\vec{\xi}$ being the coupling vector. In that case we have p input patterns $\vec{J_i}$, the ith one having V_i as output. This observation establishes an interesting connection between unsupervised and supervised learning that allows to make some predictions about the perceptron as an encoder from what is known for the perceptron as an information storage device. We present these results in [13]. We have seen here several consequences of this relation. One of them is the expression for the information capacity found in section 3, it is to be compared with the information storage capacity of the single output neuron perceptron found in [14]. In particular the critical storage capacity has to be the same as the value of α where the information capacity changes behavior. The duality between the two systems is also reflected in the fact that the results found in section 5 for the mutual information depend only on the product $G\,\Gamma$ of input pattern and synaptic correlations.

Work on extensions of the present communication is in progress, some important questions are the effect of noisy inputs and stochastic outputs. The replica technique can again be used for these cases [15]. A more difficult problem from the technical point of view is the case of discrete input neurons, although some predictions can be made from what is known about supervised learning [13]. This case most probably will require the use of one-step replica symmetry breaking.

Acknowledgments

We thank G. Toulouse for stimulating discussions. One of us (NP) wants to thank the kind hospitality received at both the Laboratoire de Physique Statistique of the Ecole Normale Supérieure (Paris) and the Laboratorio di Fisica and the INFN of the Istituto Superiore di Sanita (Rome) . This work was partly supported by the programme Cognisciences of C.N.R.S.

References

1. Barlow H. B., in *"Sensory Communication"*. Ed. Rosenblith W. MIT Press (Cambridge, MA, 1961) pp. 217; "Cerebral Cortex as Model Builder" in *Models of the Visual Cortex.* Eds. Rose D. and Dobson V. G. (John Wiley, 1985); *Neural Comp.* **1**, 295 (1989).
2. Barlow H. B., Kaushal T. P. and Mitchison G. J., *Neural Comp.* **1**, 412 (1989).
3. Linsker R., *Computer* **21**, 105 (1988). *Proc. Natl. Acad. Sci. USA* **83**, 7508,8390,8779 (1986).
4. Atick J. J. and Redlich A., *Neural Comp.* **2**, 308 (1990); Atick J. J., Li Z. and Redlich A. "Understanding retina color coding from first principles" *preprint* IASSNS-HEP-91/1, to appear in Neural Comp. (1992).
5. Atick J. J. and Redlich A., *Neural Comp.* **4**, 196 (1992).
6. Nadal J.-P. and Parga N., "Information Transmission by a Perceptron", *preprint* LPSENS (1992). Submitted to NETWORK.
7. Mezard M., Parisi G. and Virasoro M., *"Spin Glass Theory and Beyond"*. World Scientific (Singapore,1987).
8. Shannon C. E. and Weaver W., *"The Mathematical Theory of Communication"*. The University of Illinois Press (Urbana,1949).
9. Blahut R. E., *"Principles and Practice of Information Theory"*. Addison-Wesley (1988).
10. Cover T., *IEEE Trans. Electron. Comp.* **14**, 326 (1965).
11. Gardner E., "The space of interactions in neural networks models" *J. Phys. A Math. Gen.* **21**, 257 (1988).
12. Monasson R., "Properties of Neural Networks Storing Correlated Patterns" *J. Phys. A Math. Gen.* **25**, 3701 (1992).
13. Nadal J.-P. and Parga N., "Dual Learning Machines: a bridge between Supervised and Unsupervised Learning", *preprint* LPSENS (1992).
14. Brunel N., Nadal J.-P. and Toulouse G., "Information Capacity of a Perceptron" to appear in *J. Phys. A Math. Gen.* (1992)
15. Nadal J.-P. and Parga N., in preparation.