# Information Processing by a Perceptron in an Unsupervised Learning Task

Jean-Pierre Nadal
*Laboratoire de Physique Statistique**
*Ecole Normale Supérieure*
*24, rue Lhomond, F-75231 Paris Cedex 05, France*

Nestor Parga
*Departamento de Física Teórica*
*Universidad Autónoma de Madrid*
*Canto Blanco, 28049 Madrid, Spain*

**Abstract**

We study the ability of a simple neural network (a perceptron architecture, no hidden units, binary outputs) to process information in the context of an unsupervised learning task. The network is asked to provide the best possible neural representation of a given input distribution, according to some criterion taken from Information Theory. We compare various optimization criteria that have been proposed : maximum information transmission, minimum redundancy and closeness to factorial code. We show that for the perceptron one can compute the maximal information that the code (the output neural representation) can convey about the input. We show that one can use Statistical Mechanics techniques, such as the replica techniques, to compute the typical mutual information between input and output distributions. More precisely, for a Gaussian input source with a given correlation matrix, we compute the typical mutual information when the couplings are chosen randomly. We determine the correlations between the synaptic couplings which maximize the gain of information. We analyse the results in the case of a one dimensional receptive field.

P.A.C.S. 05.20 Statistical mechanics
P.A.C.S. 87.30 Biophysics of neurophysiological processes
To appear in NETWORK

*Laboratoire associé au C.N.R.S. (U.R.A. 1306) et aux Universités Paris VI et Paris VII.

# 1 Introduction

In the study of formal (artificial) neural networks, two main learning schemes are generally considered, which correspond to two aspects of information processing. One is supervised learning, in which the emphasis is put on the storage of information, and on the ability of the network to infer a rule from the learned examples. The other one is unsupervised learning, where no desired output is specified for the patterns presented to the network: the network organizes itself in the absence of a teacher. This paper is concerned with this second scheme.

Consider for example a feedforward network. A set of data is sequentially encoded as patterns of activities in the first layer. For each pattern the network then produces some configuration on his output layer, and it "organizes" itself by altering its synaptic couplings according to some specific prescription that depends on the input pattern and the produced output. The prescription is chosen in order to perform some data analysis, such as clustering: one is interested in the statistical structure of the source that has generated the data. At the end of the (unsupervised) learning process, the network is encoding the data in a neural representation on which the statistical structure of the data is more easily read. One of the most well known unsupervised learning algorithm is the one proposed by Kohonen [1], which leads to a neural representation of the data where similar inputs are encoded by the activity of nearby cells. Another, and simple, example is unsupervised Hebbian learning which, with some appropriate constraints, leads to a principal component analysis [2] [3] [4]. A typical result is that an unsupervised learning scheme produces neural representations made of specialized, possibly "grand-mother" type, cells: each output cell will be active for some specific feature. For example one will have one cell whose activity gives the projection onto the first principal component, another giving the projection on the second component, and so on.

In many cases, the rule used to modify the couplings is chosen in order to minimize some cost function that characterizes the quality of the data processing. In engineering applications one is free to choose the cost function according to the type of processing and the set of constraints that are specific to the particular problem which is considered [5]. In living organisms, self-organization is known to occur in particular during post-natal development [6]. One possible approach to the modeling of such mechanisms is to assume that an optimization process is going on, and one has to guess a cost function based on what could be the aim of the neural processing and on constraints specific to biological systems. Such an approach was suggested a long time ago by Barlow [7] [8]. His ideas led to the choice of cost functions based on information theory concepts. Such choice is not, clearly, limited to brain modeling. Data analysis is a form of *information* extraction. To give an example, for a multidimensional Gaussian distribution computing the first principal component is equivalent to maximizing the amount of information that can be obtained when representing each datum by a single variable. Coming back to brain modeling, the early visual system has frequently been taken as a testground where to check the validity of organization principles[9], and recently such strategy (that is the use of information theory for deriving an appropriate cost function) has been used for the modeling of the mammalian retina and first layers of the visual cortex by Linsker [10] and Atick and Redlich [11]. The resulting predictions detailed by Atick and Redlich [12] [13] [14] for the receptive field of ganglion cells and for color processing seem to be in agreement with physiological studies. The implications of noise in the treatment of visual information has been strongly emphasized by these authors. In their work the information source is taken as a statistical

ensemble of random patterns presenting correlations similar to those observed in pictures of natural scenes as well as pictures of human faces.

If the above quoted authors consider that Information Theory is a proper tool for modeling biological organization, they disagree on the specific criterion for choosing the cost function. Barlow argues that the aim of the processing in the first stages of the sensory pathways is to eliminate the redundancy in the stimuli: two different units should encode as much as possible different (independent) information, and this leads to factorial codes (the units being statistically independent). The original proposal that redundancy reduction and factorial codes play an important role in the way the brain performs information processing is discussed by Barlow in several papers [7] [8] [15]. Stating it briefly, 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 behaviour 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. An implementation of such redundancy reduction principle has been performed by Atick and Redlich in a model of the retina[13]. Another type of redundancy is the difference between the maximal amount of information that could be encoded by the network and the amount of information actually encoded. Reduction of this redundancy has been also considered by Atick and Redlich [12]. Lastly, Linsker has considered a rather intuitive criterion, which he calls the "infomax" principle, which is to maximize the information that the output of the neural network has about the environment [10]. Despite the differences between these three points of view, the results obtained are very similar. Clearly, there is a need for a clarification. We point out that there exists other approaches making use of information theory: in particular Bialek and Zee [16] considered the performance of a neural network assuming that for the early visual system the task is to *discriminate* a signal from a noisy background. This leads to optimizing another type of cost function taken from information theory. In this paper we will not consider this alternative approach.

In this paper we discuss several aspects of elementary information processing by a neural network, in the line of [7], [10] and [11]. The success obtained by Linsker, Atick and Redlich in modeling the retina is largely due to the fact that they consider linear neurons, what allows a full analytical study of the model. However, it is of interest to consider non-linear transfer functions, and what we do in this paper is to study the extreme case of linear thresholding. In fact, it appears that the case of a simple perceptron (no hidden units) with binary output neurons, what we will call the neural encoder, presents some advantages over the case of linear units, as far as theoretical understanding is concerned. In particular, for linear units no information quantity is defined without taking into account some noise, such as quantization noise. If on the contrary the output is discrete, this drastic quantization makes everything well defined. When considering the information capacity, we will see another aspect of this type of simplification. Also, we will see that the several principles for self-organization quoted above can be easily compared. A binary perceptron was also considered (for a different task) in Bialek and Zee [16] with the additional hypotheses of translationnally invariant neurons. Such assumption, biologically motivated, is convenient for linear neurons (one can make use of Fourier transforms), but not for binary neurons. Indeed analytical results in [16] are obtained only for some specific limits which allow for approximations,

and we will see that on the contrary, without this assumption -hence allowing each neuron to chose its own transfer function-, one can perform a detailed theoretical analysis of the perceptron with binary neurons.

The simple perceptron as a neural network architecture that processes information by producing a proper code might correspond to a module anywhere in any sensory path or in the cortex. But for illustrative purpose, and to emphasize the functional role of this module one may think of a retina; from this point of view the input signal corresponds to a "visual scene". the purpose of this neural encoder is to recode the input information by translating the input message, which may be expressed in a highly correlated (redundant) code, into a more efficient one represented in its output layer. We will evaluate the quality of the code (the neural representation given by the output layer) by considering the amount of information that it gives on the input distribution. Two main questions are of interest: for a given network, what is the maximal amount of information that the net is able to extract from the source, no matter what the source is - this is the question of the information capacity, related to the Shannon capacity in communication theory; for a given input source, what is the choice of the couplings which will maximize the gain of information. We will give an exact answer to the first question, and this will allow us to compare the various optimization principles. Next we will consider a statistical formulation of the second question. In that case, we will show that Statistical Mechanics techniques, such as "replica techniques"[17], can be used to compute the typical gain of information associated to a statistical ensemble of networks. It is a remarkable fact that techniques taken from the statistical physics of disordered systems can be used for studying the ability of a network to perform an unsupervised learning task, and we expect the particular computation that we are presenting in this paper to be only a first example of what can be done.

The paper is organized as follows. Section II the neural encoder is defined and the basic information theoretical concepts and quantities that will be useful for this work are introduced. In section III we calculate the information capacity. In section IV we discuss for the perceptron the various proposals for a general principle of information processing. In section V we compute the typical mutual information of the encoder using Statistical Mechanics techniques. In particular we show how the replica technique [17] can be used to evaluate the average mutual information for an ensemble of encoders characterized by a non trivial probability distribution. The case of one-dimensional spatially correlated input patterns is considered as an example. The last section contains our discussions.

# 2 The Neural Encoder

## 2.1 Basic concepts in Information Theory

We consider the problem of processing a signal $\vec{\xi} = \{\xi_j\}_{j=1,\dots,N}$ produced by a source with a probability $P_\xi$. In a first stage along the processing line the signal $\vec{\xi}$ is received by an encoder that produces a new code using vector code symbols $\vec{V} = \{V_i\}_{i=1,\dots,p}$ representing the activities of its $p$ output neurons (figure 1). This encoder, that later on will be taken as a neural network, is completely defined by giving its architecture, the joint probabilities $P(\vec{V}, \vec{\xi})$ and a set of internal parameters such as the synaptic couplings $J_{ij}$. The encoded signal will be taken by other modules down in the line. For instance, it could be transmitted through a channel to one or several neural networks and eventually it could be decoded by another module. Although we will put the emphasis of our description in the properties of

the encoder, the different structures along the processing line can be treated in a similar way.

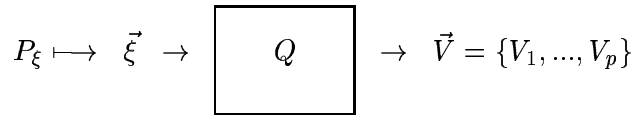


Figure 1: The processing module: the input $\vec{\xi}$ produced by the source $P_\xi$ is transformed into a new represenation $\vec{V}$ according to the transfer function $Q$.

Two relevant quantities for information processing [18] [19] [20] are the mutual information and the maximal information that the output of a module can convey about its input (keeping fixed its architecture and internal parameters), a quantity that we will call information capacity. The first gives the information that the module transmits from a given source (the information that the output $\vec{V}$ conveys on the signal), it can be expressed as

$$I(\vec{V},\vec{\xi}) = \sum_{(\vec{V},\vec{\xi})} P(\vec{V},\vec{\xi}) \log \left\{ \frac{P(\vec{V},\vec{\xi})}{P_V P_\xi} \right\} \tag{1}$$

Here and in the following logarithms are expressed in base 2 ($\log = \frac{\ln}{\ln 2}$), so that information quantities are measured in bits. In eq. (1) $P_V$ is the output state probability:

$$P_V = \int d\vec{\xi} P_\xi Q(\vec{V}|\vec{\xi}). \tag{2}$$

where $Q(\vec{V}|\vec{\xi})$ is the conditional probability to find the output state $\vec{V}$ given the input pattern $\vec{\xi}$. This "transfer function" $Q$ fully characterizes the processing module. Notice that $I(\vec{V},\vec{\xi})$ is the Kullback-Leibler distance [20] between $P(\vec{V},\vec{\xi})$ and the hypothesis that $\vec{\xi}$ and $\vec{V}$ are independent. If this were the case no information would be transmitted, i.e. $I = 0$.

The *information capacity* $C$ is defined as the maximum of the mutual information over all possible environments:

$$C = Max_{P_\xi} I(\vec{V},\vec{\xi}), \tag{3}$$

and as we have just said, it measures the maximal information that the module can transmit. From the point of view of a subsequent channel receiving this output, the object we have just discussed acts as an encoder and $C$ has the meaning of the maximal rate at which it can deliver information.

There is also a maximal possible quantity of information $I_0$ that can be extracted from the source

$$I(\vec{V},\vec{\xi}) \leq I_0 \tag{4}$$

If the source is discrete, $I_0$ is equal to the average information content per message $H(P_\xi)$, that is the source entropy [18] [19]:

$$I_0 = H(P_\xi) \equiv -\sum_{(\vec{\xi})} P_\xi \log P_\xi \tag{5}$$

For continuous inputs, the entropy of the source is infinite and the maximal information $I_0$ that can be obtained is only limited by the resolution with which the receptors can measure the signals.

Let us notice that in general for an arbitrary source the encoder will not be used efficiently. A measure of this inefficiency is given by the redundancy

$$\mathcal{R}_C = C - I. \tag{6}$$

If one achieves $I = I_0$ and $\mathcal{R}_C = 0$, then the resulting code is said to realize a compaction of the input signal. Note that it may be that not all the input information is useful, so that a minimal level $I_1 \leq I_0$ may be sufficient to achieve. In that case, the new code (i.e. after eliminating the redundancy) is said to realize a compression of the original information.

We emphasize that the capacity depends on the "transfer function" $Q(\vec{V}|\vec{\xi})$,

$$C = C(Q), \tag{7}$$

the mutual information $I$ depends on $Q$ and on the input distribution $P_\xi$, and is bounded from above by both $C(Q)$ and $I_0 = I_0(P_\xi)$.

## 2.2 Application to the noiseless perceptron

We define the neural encoder as a feedforward network that we want to analyse as an elementary module processing information. We will consider only the simplest case, that is the perceptron architecture (one input layer, no hidden units and one output layer). More specifically the network as shown on figure 2 consists of a set of $N$ input and $p$ output neurons with synaptic connections $J_{i,j}$ as shown in Fig. (1). The output neurons are taken as linear threshold elements. We will consider only a deterministic transfer function: for every output neuron $i$ $(i = 1, ..., p)$

$$V_i = sgn(\sum_{j=1}^{N} J_{i,j}\xi_j - \theta_i). \tag{8}$$

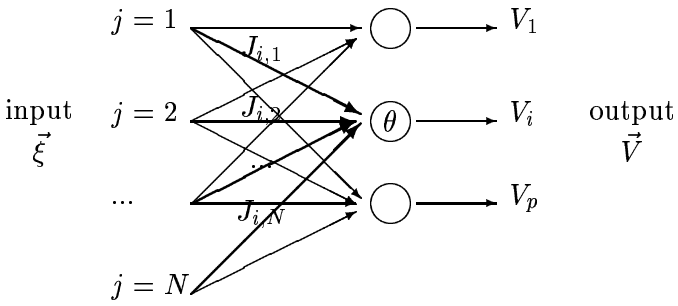


Figure 2: The neural encoder as a perceptron architecture.

The input $\vec{\xi}$ may be discrete or continuous, but in both cases, since the output is discrete the mutual information (1) is well defined. Here we will mainly consider continuous, noiseless inputs. Also in this paper we will restrict ourselves to unbiased input distributions ($< \xi_j >= 0$), so that it will always be optimal to set all thresholds $\theta_i$ to zero. This we do right now, although extension to nonzero thresholds is straightforward.

For a deterministic system the mutual information (1) is nothing but the entropy of the output distribution:

$$I(\vec{V},\vec{\xi}) = -\sum_{\vec{V}} P_V \log P_V \tag{9}$$

Here $P_V$ is the output probability distribution (2) resulting from the deterministic conditional probability $Q(\vec{V}|\vec{\xi})$ given by:

$$Q(\vec{V}|\vec{\xi}) = \prod_{i=1}^{p} \Theta(V_i \vec{J}_i.\vec{\xi}), \tag{10}$$

where $\Theta$ is the Heaviside distribution:

$$\begin{aligned} \Theta(y) &= 1 \quad y > 0 \\ &\quad 0 \quad y < 0 \end{aligned} \tag{11}$$

# 3 The Information Capacity

It is easy to see that the maximum condition in the capacity definition (3) is achieved for a source such that the probabilities for all possible output states 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 them is $2^p$, so that

$$C \leq p \tag{12}$$

(where C is measured in bits). However not every configuration may be realizable. According to (8) each one of the $p$ output neurons is associated to an oriented hyperplane which cuts the $N$-dimensional input space. For a given input, the output state $V_i$ specifies on which side of the ith hyperplane the input lies. Hence each realizable configuration $\{V_i\}_{i=1,\ldots,p}$ corresponds to one region in input space, and the number of accessible output configurations is the number of domains obtained with $p$ hyperplanes. This number, $\Delta(p,N)$, has been obtained from geometrical counting in the sixties [21] 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 $k \leq N$ vectors $\vec{J}_i$ are linearly independent). Then

$$\Delta(p,N) = \sum_{l=0}^{min(p,N)} C_p^l \tag{13}$$

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

$$C = \log \Delta(p,N). \tag{14}$$

¿From this formula one sees that the $2^p$ output sates are available up to $p = N$:

$$\begin{aligned} C &= p \quad p < N \\ &< p \quad p > N \end{aligned} \tag{15}$$

Of particular interest is the large $N$ limit. In that limit, we will measure information quantities in bits per input neuron, since the information fed into the network scales with $N$. Defining

$$\alpha = p/N \tag{16}$$

one finds that the fraction of unrealizable configurations is negligible up to $\alpha = 2$, and goes to 1 above $\alpha = 2$. More precisely, the asymptotic capacity $c(\alpha)$ in bits per input neuron for a given value of $\alpha$ is given by

$$\lim_{N\to\infty} C/N \equiv c = \begin{cases} \alpha & \text{if } \alpha \leq 2 \\ \alpha S(1/\alpha) & \text{if } \alpha > 2 \end{cases} \tag{17}$$

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

$$S(x) = -[x \log x + (1-x)\log(1-x)]. \tag{18}$$

The curve $c = c(\alpha)$ is shown on figure 3. Note that for large $\alpha$ the capacity $c$ increases as $\log \alpha$.

To conclude this section we emphasize that the capacity (14) does not depend on the particular choice of the couplings, but only on the architecture (that is on $N$ and $p$). This would not be the case for a network with non binary neurons (e.g. linear neurons).

# 4 Principles for Self-Organization

## 4.1 Maximum Information Preservation

As pointed out in section 2, the mutual information depends on the environment, that is on the input distribution $P_\xi$, and on the couplings $J_{ij}$. This implies that by modifying the couplings one may be able to increase the mutual information. This corresponds to the "infomax" principle of Linsker [10], according to which each layer in a feedforward network is asked to extract as much information as possible from the preceding layer:

$$(\mathcal{PI}) \qquad max_{J_{ij}}\ I(\vec{V}, \vec{\xi}). \tag{19}$$

For our perceptron the capacity does not depend on the couplings. Hence $C$ is also an upper bound for the best performance that can be achieved after optimization. But this means also that maximizing $I$ is equivalent to minimizing the redundancy $\mathcal{R}_C = C - I$:

$$(\mathcal{PC}) \qquad min_{J_{ij}}\ \mathcal{R}_C \tag{20}$$

One interpretation of this principle is that for optimal processing the complexity of the system should match the one in the data. This criterion $\mathcal{PC}$ has also been studied, in particular by Atick and Redlich [12] in several papers on the modeling of the visual system with multilayers of linear neurons. In that case, and more generally for non binary neurons, the capacity depends on the couplings: the minimization of redundancy must be applied under the constraint that $I$ is large enough, say $I \sim I_0$ (no information loss), and $\mathcal{PC}$ and $\mathcal{PI}$ may not be equivalent.

### 4.2 Barlow's redundancy reduction principle

Another interesting criterion comes from Barlow's observation of the need to use factorial codes [7]. A way to motivate these codes is to notice that in order to make predictions about its environment a living organism needs to know the "a priori" probabilities of all possible combinations of events occurring in that environment. To give an example, Barlow points out that in Pavlovian conditioning a good predictor of the association between the conditional stimulus $c$ and the unconditional one $u$ is $P(c,u) > P(c)P(u)$. It is clear that its implementation requires the knowledge of the a priori probabilities of at least two events. In general the number of elementary stimuli is large and what is required is the conjunction of any number of them, so that the direct knowledge of all the a priori probabilities is not possible. However the combinatorial problem is eliminated if information is represented by a factorial code; this means that the complicated statistical structure of the environment defined by $P_\xi$ is transformed, after coding, into a $P_V$ such that

$$P_V = \prod_{i=1}^{p} P_i(V_i), \tag{21}$$

that only requires to know the probabilities of single events.

This way to treat the source information solves another problem [7]. Once the system acquires knowledge about its environment, it should be able to detect the appearance of new conjunctions or "suspicious coincidences" of events. That is, it should be able to distinguish what is new from what is old. If previous experiences have been represented according to a factorial code, the existence of correlations between the $V_i's$ will signal that something not yet known requires attention.

These remarks lead 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 (the information that the $i$th neuron alone contain on the signal). For a deterministic system it is given by

$$I_i(V_i, \vec{\xi}) = -\sum_{(V_i)} P_i(V_i) \log P_i(V_i) \tag{22}$$

where $P_i(V_i)$ is the marginal distribution:

$$P_i(V_i) = \sum_{V_j, j \neq i} P_V. \tag{23}$$

Some of the information bits may be given at the same time by several neurons, so that one has a redundancy $\mathcal{R}_B$ defined by

$$\mathcal{R}_B = \sum_{i=1}^{p} I_i - I, \tag{24}$$

and Barlow's principle reads

$$(\mathcal{PB}) \qquad min_{J_{ij}}\ \mathcal{R}_B \tag{25}$$

Since the activity of each output cell is only a function of the input (to compute $V_i$ one needs to know the input and not the other $V_j$s) the above quantity is indeed positive (or null)

[19]. In fact one can show [19] that $\mathcal{R}_B$ can be rewritten as the Kullback-Leibler distance between the output probability distribution and the factorial distribution $\pi$ defined by

$$\pi_V = \prod_{i=1}^{p} P_i(V_i). \tag{26}$$

Hence achieving $\mathcal{R}_B = 0$ is indeed equivalent to realizing a factorial code. Clearly $\mathcal{R}_B = 0$ is achieved for $I = I_i = 0$, and the minimization must be performed under the constraint that $I$ is large enough, say $I \simeq I_0$. This principle $\mathcal{PB}$ has been used in one of the works by Atick and Redlich [13], and it can be shown to be directly related to the search for minimum entropy codes [22] in the case of discrete inputs and outputs [23].

Now for the perceptron, it is easy to see that for each neuron

$$I_i = 1 \tag{27}$$

Indeed, the hyperplane cuts the input space into two parts of equal weights (note that for non symmetric distributions it would be easy to choose a threshold such that $I_i = 1$). Hence one has

$$\sum_{i=1}^{p} I_i(V_i, \vec{\xi}) = p \geq C. \tag{28}$$

This implies that for the perceptron the principles $\mathcal{PC}$ and $\mathcal{PB}$ are equivalent, and thus all three principles give the same result.

We note that this is also the case for linear neurons with a Gaussian signal. In such a case, the principle $\mathcal{PI}$, *with the additional constraint that the code should be reversible*, leads to performing a principal component analysis (PCA) on the signal, a task which can be realized by unsupervised Hebbian learning schemes [2] [3]. But this PCA can be given another interpretation: it results in output neurons which are statistically uncorrelated, and this corresponds to following Barlow's requirement of finding factorial codes.

### 4.3 Optimal adaptation

Let us now summarize the discussion as illustrated on figure 3. We assume that $i_0 = I_0/N$ is finite. Then there is a particular value $\alpha_0$ of $\alpha$ for which the information capacity matches the available information, $c(\alpha_0) = i_0$. At a given value of $\alpha$, the best that can be done is to maximize $i = I/N$, hopefully up to $i = c$ for $\alpha$ below $\alpha_0$ and up to $i = i_0$ above $\alpha_0$.

By modifying the architecture (changing $N$ or $p$), one may try to achieve $i = i_0 = c$, which is the optimal point where $\mathcal{R}_C$ is minimal and there is no loss of information. Recruitment and elimination of neurons are plausible mechanisms which could account for the adaptation of the information capacity. However, it is also tempting to assume that the optimization of the architecture is rather the result of evolutionary adaptation.

To conclude, for a given architecture (given values of $N$ and $p$) the three principles $\mathcal{PI}$, $\mathcal{PC}$ and $\mathcal{PB}$ are equivalent for the perceptron. However, the possibility of adapting the architecture allows to separate two aspects of the problem: maximization of the mutual information (by modifying the couplings), and minimization of the redundancy (by modifying the architecture).

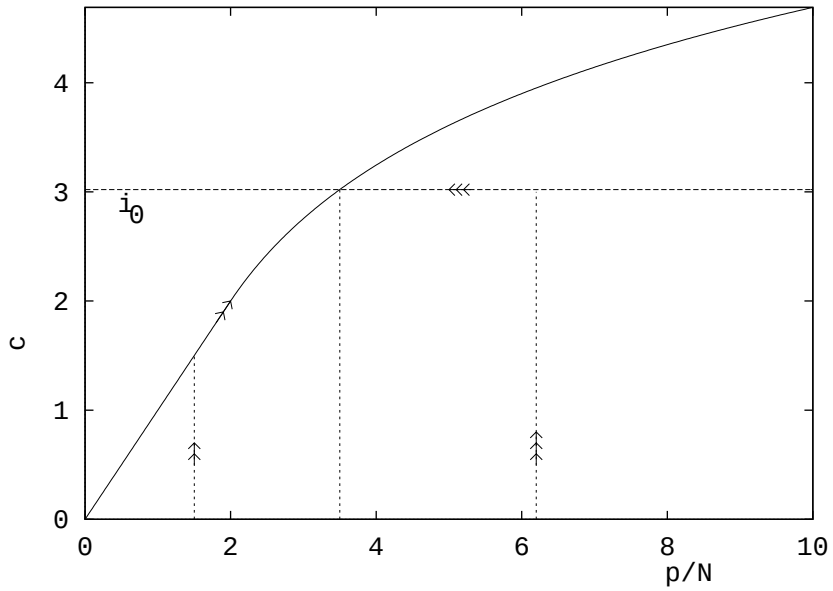


Figure 3: Solid line: The information capacity $c$ of the neural encoder (in bits per input neuron for large $N$) as a function of $\alpha = p/N$. The mutual information $i$ will be smaller than $c$ and the information content $i_0$ of the source. The arrows indicate possible trajectories when adapting the couplings below (>>) and above (>>>) $\alpha_0$.

# 5 The Typical Mutual Information

## 5.1 Statement of the Problem

What we have just seen is that for a given input distribution $P_\xi$ one should find the couplings that maximize the mutual information. Can this be done for the perceptron with binary neurons ? And what is the best value of $I$ that can actually be achieved (it is not clear whether $C$ can be reached, since $C$ is the optimum over all possible environments) ? For $\alpha$ smaller than 1, and a Gaussian input distribution, one can easily answer to these questions: exactly as for linear neurons, one should perform a Principal Component Analysis, and the mutual information is equal to the capacity (each neuron is giving one bit of information).

The problem is much more difficult for $\alpha$ larger than 1. Instead of trying to find *the* optimal couplings, we will consider a statistical ensemble of coupling vectors characterized by the correlations in their components. At the end we will look for the correlations which maximize $I$. Moreover, for any $\alpha$, we would like to know what is the information that is obtained in the absence of any optimization: this can be obtained by taking random couplings, with each component being an independent unbiased random variable. The result, compared to the capacity and to the information obtained with the best choice of correlations, will tell us how much can be gained by optimization.

The quantity we want to evaluate is thus the typical mutual information per input unit, $i$, that results when the couplings are taken from a statistical ensemble:

$$i \equiv lim_{N\to\infty} << I >> /N = lim_{N\to\infty} << H(P_V) >> /N \qquad (29)$$

where $<< . >>$ means the average over the coupling distribution $\rho$:

$$<< U >>= \int \prod_{i,j}^{p,N} dJ_{i,j} \ \prod_i \rho(\vec{J_i}) \ U(\{\vec{J_i}\}_{i=1}^p). \tag{30}$$

To evaluate the average $i$ we make use of the replica technique [17]. We will not give all the technical details, but we will point out what is specific to the present computation.

## 5.2 Application of the Replica Techniques

The mutual information $I(\vec{V}, \vec{\xi})$ is associated to a given input distribution $P_\xi$. The distributions which allow an analytical study are (unsurprisingly!) the Gaussian ones. In this paper we consider unbiased but spatially correlated Gaussian input patterns, i.e.

$$P_\xi = \frac{1}{\sqrt{(2\pi)^N detG}} \exp[-\frac{1}{2} \sum_{ij} \xi_i (G^{-1})_{ij} \xi_j] \tag{31}$$

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 a Gaussian distribution: all what will matter in the large $N$ limit is the first two moments of their distribution $\rho(\{J_{i,j}\}; j = 1, ..., N)$. We thus consider

$$<< J_{i,j} >>= 0 \tag{32}$$

$$<< \ J_{i,j} \ J_{i',k} >>= \delta_{i,i'} \ \Gamma_{jk}. \tag{33}$$

At the end we will look for the correlation matrix $\Gamma$ which maximizes $i$. We want to compute

$$i = \frac{1}{N} \sum_V \ll - P_V \log P_V \gg \tag{34}$$

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. (2). According to the replica method eq. (34) is written as

$$\ll I(V, \vec{\xi}) \gg = -\sum_V \ll P_V \lim_{n \to 0} ((P_V)^n - 1)/n \gg, \tag{35}$$

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

$$\sum_{\vec{V}} P_V = 1, \tag{36}$$

one can write also

$$\sum_V \ll [P_V]^{n+1} \gg \overset{n \to 0}{\sim} \exp(-nNi). \tag{37}$$

The computation of the left hand side of (37) follows closely the one of standard Gardner's calculations [24]. Under the "replica symmetry" ansatz, the result is as follows. Defining $s$ by

$$s = \tau[G \ \Gamma], \tag{38}$$

with

$$\tau[.] \equiv lim_{N\to\infty}\frac{1}{N}Tr\ (.), \tag{39}$$

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

$$i = extr_{q,\hat{q}}\{\hat{q}\frac{1-q}{2\ln 2} + \frac{1}{2}\,\tau[\log(1-\hat{q}\,\mathcal{G})] + \alpha\int_{-\infty}^{\infty} DzS(h)\} \tag{40}$$

where $\mathcal{G}$ is the normalized matrix:

$$\mathcal{G} = \frac{G\,\Gamma}{s} \tag{41}$$

(note that by definition of $s$ $\tau[\mathcal{G}] = 1$). In (40) $Dz$ is the Gaussian measure

$$Dz = \frac{dz}{\sqrt{2\pi}}\exp(-\frac{1}{2}z^2), \tag{42}$$

and $S$ is the entropy function defined in eq.(18), with its argument $h$, function of $z$ and $q$, being given by

$$h = H(\sqrt{\frac{q}{1-q}}z) \tag{43}$$

with $H$ the error function:

$$H(y) = \int_y^{\infty} Dz. \tag{44}$$

The order parameter $q$ is defined from the scalar product, with the metric induced by $\Gamma$, of two inputs associated to two different replicas $a$ and $b$, but giving the same output configuration:

$$q = q^{ab} \equiv \frac{1}{sN}\sum_{i,j} << \xi_i^a\Gamma_{ij}\xi_j^b >> \tag{45}$$

Note the normalization by $s = \tau[G\ \Gamma]$. The saddle point equations for the order parameters $q$ and $\hat{q}$ are

$$q - 1 + \tau[(1-\hat{q}\,\mathcal{G})^{-1}\,\mathcal{G}] = 0 \tag{46}$$

$$\hat{q} - 2\alpha\ln 2\frac{d}{dq}\int_{-\infty}^{\infty} DzS(h) = 0, \tag{47}$$

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

### 5.3 Technical remarks: link with Gardner's approach to supervised learning

The above calculation is related to computations done for the study of the storage capacity of a perceptron in a supervised learning task. To see this, consider the perceptron having a *single* output unit, $N$ inputs and for which the *couplings* are $\xi_j, j = 1, ..., N$. We ask this perceptron to learn the $p$ input-output pairs $\{\vec{J_i}, V_i\}, i = 1, ..., p$, $\vec{J_i}$ being here the $i$th input pattern. Then the fractional volume of the *couplings* (the $\xi$s) which realize these associations is equal to $P_V$ when the prior measure is $\rho$. Gardner [24] has shown that one can compute, with the replica techniques and for various choices of the distributions $\rho$ and $P_\xi$, the typical value of $P_V$. More precisely, for a configuration $\vec{V}$ chosen at random, the logarithm of the fractional volume is a self averaging quantity, that is $lim_{N\to\infty}\log P_V/N$ exists and is equal to

$$\bar{l} = lim_{N\to\infty} << \log P_V) >> /N \tag{48}$$

It is this quantity, $\bar{l}$, that is computed in a standard Gardner type calculation [24]. For the perceptron we are considering in the context of unsupervised learning, $-\log P_V$ is the gain of information resulting from the occurrence of the particular output configuration $\vec{V}$, and $-\bar{l}$ is thus the typical gain of information for an output configuration chosen at random (among the $2^p$ configurations). But what we have seen is that the relevant quantity is rather the entropy (the mutual information): the computation of $\bar{l}$ in the context of supervised learning should be contrasted to the one of the entropy in the case of unsupervised learning. We detail this relationship between supervised and unsupervised learning in [25]. From the technical point of view, the formulae for $i$ and the order parameters should be compared with those obtained for $\bar{l}$ that give the storage capacity of the perceptron with continuous couplings [24], and in particular with those for the case of correlated patterns recently studied by Monasson [26].

Apart from the fact that we have here $n+1$ replicas, instead of the $n$ that appear in storage capacity calculations, the evaluation of $i$ is done in the same way. From the expression (2) one has a product of $n+1$ integrals (replicas), with a product of $p$ $\Theta$ distributions in each of them. These $\Theta$ distributions are written in an integral representation. Then one performs the average over the couplings according to (30). After integrating 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 make 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 peculiarity of our calculation is the normalization condition (36), which means that when setting $n = 0$ exactly in (37) one should find 1. This condition fixes half of the order parameters. The parameter $s$ (which is one of the two parameters fixed by the $n \equiv 0$ condition), is equivalent to the $s$ parameter of [26]. Finally we note that the term proportional to $\alpha$ can easily be understood using the cavity method [17]: it is the gain of information due to an infinitesimal increase of $\alpha$.

### 5.4 Analysis of the result

First we note the interpretation of the parameters $s$ and $q$: the typical norm of an input vector, with the metric induced by $\Gamma$, is equal to $\sqrt{s}$; $q$ is the typical value of the scalar product (with the metric induced by $\Gamma$, and divided by $s$) 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 $\alpha$ limit $q$ goes to 0. When the number of output neurons becomes very large, the typical domain size decreases to zero, hence when $\alpha$ goes to infinity one finds that $q$ goes to 1. This particular limit can be easily analyzed for any given matrix $\mathcal{G}$. As $q$ tends to one $\hat{q}$ goes to infinity and one has the asymptotic expression

$$i \overset{\alpha \to \infty}{\sim} \log \alpha + \frac{1}{\ln 2} + \frac{1}{2}\tau[\log \mathcal{G}] + \log A \tag{49}$$

where $\mathcal{G}$ is the normalized matrix given in (41) and $A$ is a constant given by

$$A = \int_{-\infty}^{\infty} \frac{dy}{\sqrt{2\pi}} \left\{ -H(y) \ln H(y) - (1 - H(y)) \ln(1 - H(y)) \right\} \tag{50}$$

where $H(.)$ is the error function defined in (44) (note that in the above expression for $A$ we have Neperian logarithms). Numerical evaluation of $A$ gives

$$A \sim 0.72 \tag{51}$$

The asymptotic expression (49) for $i$ has to be compared with the asymptotic expression of the information capacity, namely:

$$c \overset{\alpha \to \infty}{\sim} \log \alpha + \frac{1}{\ln 2}. \tag{52}$$

This shows that, whatever the correlations in the couplings and in the inputs, the asymptotic mutual information (in bits per input) scales exactly as the information capacity, and is below it by a constant which depends on $\mathcal{G} = \frac{G\,\Gamma}{s}$.

From the asymptotic expression (49) of $i$, it is clear that the best choice of correlations for the couplings is $\mathcal{G} = 1$, that is

$$\Gamma = G^{-1} \tag{53}$$

This result is in fact valid at any value of $\alpha$. In expression (40) for the mutual information $i$ the matrix $\mathcal{G}$ enters only in one term. Once $q$ and $\hat{q}$ are taken as the values which maximize $i$, optimization with respect to $\Gamma$ is obtained by maximizing $\tau[\log(1 - \hat{q}\,\mathcal{G})]$. Since $\tau[\mathcal{G}] = 1$ the optimum is realized for the unit matrix, $\mathcal{G} = 1$. 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. The optimal value is still below the capacity. It is clear that with random couplings the number of domains in input space will be typically equal to the number $\Delta(p, N)$, and this is certainly why the typical mutual information scales as the capacity for large $\alpha$. But the difference by a constant term means that the domains cannot be given the same weight even when optimizing the statistical characteristics of the network (even below $\alpha = 1$). It would be interesting to see whether one can find for $\alpha$ larger than 1 an efficient algorithm allowing to come closer to the capacity.

Lastly we comment on a point evoked in the introduction. In most related works, one assumes that all the neurons have the same transfer function :

$$J_{i,j} = J(\|\vec{r}_i - \vec{r}_j\|) \tag{54}$$

where $\vec{r}_i$ and $\vec{r}_j$ are the locations of the output neuron $i$ and of the input unit $j$ respectively. In the present paper we did not restrict the couplings. In fact, imposing such a translationnal invariance would render the analytical study much harder (see [16] for a related study where (54) is assumed). However, in our statistical approach we impose a statistical invariance by translation: the correlation matrix $\Gamma$ does not depend upon the site $i$. In fact, our result (53) for the optimal *statistical* properties of the couplings is similar to the *exact* result obtained for linear neurons in the low noise limit and under the hypotheses (54) [12]. In particular one finds that the receptive fields have a mexican-hat shape, as in the particular example detailed below.

### 5.5 1-d example

In this subsection we discuss with more detail one particular example. We have studied a one dimensional example with the correlation matrix

$$G_{ij} = x^{|i-j|} \tag{55}$$

where $x$ is in the interval $[0,1]$. This correlation has been considered previously in [26] to study storage capacity properties of neural networks with spatially correlated patterns. We consider the average mutual information obtained with uncorrelated couplings ($\Gamma = 1$, $\mathcal{G} = G$). For this particular case one has just to replace, in the formulae giving $i$ and the order parameters, the trace $\tau$ by the sum over the eigenvalues $\lambda$ of $G$:

$$\tau[f(G)] = \int_0^{2\pi} \frac{d\phi}{2\pi} f(\lambda(\phi)) \tag{56}$$

where $\lambda(\phi)$ is given by

$$\lambda(\phi) = \frac{1-x^2}{1-2x\cos\phi + x^2} \tag{57}$$

In particular

$$s = \tau[G] = \int_0^{2\pi} \frac{d\phi}{2\pi} \lambda(\phi) \tag{58}$$

Evaluation of the right hand side of the above formula gives 1, as it should (from the definition (55) of $G$, $\frac{1}{N}TrG$ is 1 for any $N$).

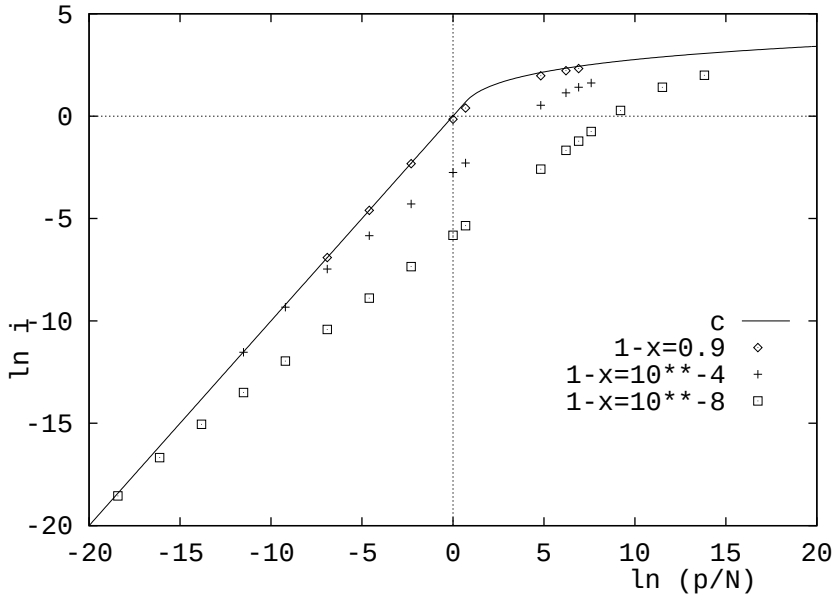


Figure 4: The capacity of the neural encoder and the mutual information as a function of $\ln\alpha$ for the one dimensional example discussed in the text. The mutual information is shown for several values of the correlation parameter $x$. Full line: $\ln c$. The symbols correspond to $\ln i$ for $1-x = 0.9, 10^{-4}$ and $10^{-8}$ (from top to bottom). All information quantities are measured in bits.

The resulting average mutual information is shown in figure 4 as a function of $\ln\alpha$ for several values of $x$, together with the information capacity. The limits $q \to 1$ and $q \to 0$ can be obtained analytically. The first corresponds either to the large $\alpha$ limit keeping $x$ fixed or to $x \to 1$ with $\alpha$ fixed, that is to the two limits where a given output codes a unique

(macroscopically speaking) input. As we have already said the small $\alpha$ behavior is obtained for $q \to 0$. More explicitly

$$i = \alpha[1 - \alpha F(x)] \tag{59}$$

for small $\alpha$. For large $\alpha$ and $x$ fixed it behaves as $\log \alpha$, i.e. it increases in the same way as $c$ but remains asymptotically below it, while in the limit $x \to 1$ and $\alpha$ fixed it is

$$i = K(1-x)^{\frac{1}{3}}\alpha^{\frac{2}{3}} \tag{60}$$

with $K$ a numerical constant. As can be seen in Fig.(3) this behaviour appears for intermediate values of $\alpha$ in an interval that increases as $x$ approaches 1.

Since all our equations depend only on $G\,\Gamma$ it is clear that the same result holds when the correlation $G$ is the unit matrix and the synapses converging to a given output neuron are correlated according to eq. (55).

# 6 Conclusion

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. 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 conveyed by individual output neurons as well as the information capacity do not depend on the synaptic couplings.

The evaluation of the typical mutual information of an ensemble of networks can be done with the replica technique. The order parameter $q$ that appears is the typical distance between two inputs belonging to the same domain. Since every domain in input space, associated to a given output configuration, is convex, the replica symmetric ansatz should be exact. We have considered a family of neural encoders characterized by a two-point correlation $\Gamma$ between synapsis. Maximization of the typical mutual information then results in $\Gamma = G^{-1}$. As in [12] the synapsis develop in such a way that they cancel the source correlations. For the one-dimensional example these correlations give to the receptive field of the output units a mexican hat shape.

A particular feature of the perceptron is that the deterministic rule defining the output neuron state $\vec{V}$, eq.(8) with zero thresholds, 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 $i$th 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 a neural encoder from what is known for the perceptron as an information storage device. We present this duality in details in [25]. In particular the computation of the typical mutual information for continuous inputs is closely related to the Gardner's calculation of the storage capacity for continuous couplings - and the origin of the validity of the replica symmetric ansatz is the same in these calculations. This relationship also explains the similarity of our results with the Gardner type computation of the critical storage capacity for the case of spatially correlated patterns [26].

Work on extensions of the present paper is in progress, some important questions being the effect of noisy inputs and stochastic outputs. The replica technique can again be used

for these cases [27]. A slightly 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 [25]. This case most probably will require the use of one-step replica symmetry breaking.

## Acknowledgements

We would like to thank M. Mézard, R. Monasson, N. Sourlas, G. Toulouse and S. Verdu for fruitful discussions. This work was partly supported by the program Cognisciences of C.N.R.S..

## References

[1] Kohonen T.O. *Self-organization and associative memory.* Springer, Berlin, 1984.

[2] Oja E. *Int. Journ. of Neur. Syst.*, 1:61, 1989.

[3] Sanger T. D. *Neural Network*, 2:459, 1989.

[4] Hertz J., Krogh A., and Palmer R. G. *Introduction to the Theory of Neural Computation.* Addison-Wesley, Cambridge MA, 1990.

[5] Buhmann J. and Kuhnel H. Complexity optimized data clustering by competitive neural networks. *Neural Comp.*, 5:75–88, 1993.

[6] Blakemore C. and Cooper G. F. Development of the brain depends on the visual environment. *Nature*, 228:419–478, 1970.

[7] Barlow H. B. Possible principles underlying the transformation of sensory messages. In Rosenblith W., editor, *Sensory Communication*, page 217. M.I.T. Press, Cambridge MA, 1961.

[8] Barlow H. B. Cerebral cortex as model builder. In Rose D. and Dobson V. G., editors, *Models of the Visual Cortex.* John Wiley, New-York, 1985.

[9] Bienenstock E., Cooper L. N., and Munro P. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosc.*, 2:32–48, 1982.

[10] Linsker R. Self-organization in a perceptual network. *Computer*, 21:105–17, 1988.

[11] Atick J. J. Could information theory provide an ecological theory of sensory processing. *NETWORK*, 3:213–251, 1992.

[12] Atick J. J. and Redlich A. Towards a theory of early visual processing. *Neural Comp.*, 2:308, 1990.

[13] Atick J. J. and Redlich A. What does the retina know about natural scenes. *Neural Comp.*, 4:196–210, 1992.

[14] 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.

[15] Barlow H. B. Unsupervised learning. *Neural Comp.*, 1:295–311, 1989.

[16] Bialek W. and Zee A. Understanding the efficiency of human perception. *Phys. Rev. Lett.*, 61:1512–1515, 1988.

[17] Mézard M., Parisi G., and Virasoro M. *Spin Glass Theory and Beyond.* World Scientific Pub., Singapore, 1987.

[18] Shannon C. E. and Weaver W. *The Mathematical Theory of Communication.* The University of Illinois Press, Urbana, 1949.

[19] Blahut R. E. *Principles and Practice of Information Theory.* Addison-Wesley, Cambridge MA, 1988.

[20] Kullback S. *Information Theory and Statistics.* John Wiley, New-York, 1959.

[21] Cover T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.*, 14:326, 1965.

[22] Barlow H. B., Kaushal T. P., and Mitchison G. J. Finding minimum entropy codes. *Neural Comp.*, 1:412–423, 1989.

[23] Nadal J.-P. and Parga N. On redundancy reduction with applications to feedforward and attractor networks. *in preparation.*

[24] Gardner E. The space of interactions in neural networks models. *J. Phys. A: Math. and Gen.*, 21:257, 1988.

[25] Nadal J.-P. and Parga N. Duality between learning machines: a bridge between supervised and unsupervised learning. *LPSENS preprint, submited to Neural Computation*, 1992.

[26] Monasson R. Properties of neural networks storing spatially correlated patterns. *J. Phys. A: Math. and Gen.*, 25:3701–3720, 1992.

[27] Nadal J.-P. and Parga N. *in preparation.*