Non linear neurons in the low noise limit: a factorial code maximizes information transfer

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 investigate the consequences of maximizing information transfer in a simple neural network (one input layer, one output layer), focussing on the case of *non linear* transfer functions. We assume that both receptive fields (synaptic efficacies) and transfer functions can be adapted to the environment.

The main result is that, for bounded and invertible transfer functions, in the case of a vanishing *additive* output noise, and no input noise, maximization of information (Linsker's*infomax* principle) leads to a factorial code - hence to the same solution as required by the redundancy reduction principle of Barlow.

We show also that this result is valid for *linear*, more generally unbounded, transfer functions, provided optimization is performed under an additive constraint, that is which can be written as a sum of terms, each one being specific to one output neuron.

Finally we study the effect of a non zero input noise. We find that, at first order in the input noise, assumed to be small as compared to the - small - output noise, the above results are still valid, provided the *output* noise is uncorrelated from one neuron to the other.

P.A.C.S. 87.30 Biophysics of neurophysiological processes
Short title: Information maximization with non linear neurons
To appear in NETWORK
INDEX: nadalparga.infomaxredred.ps.Z nadal@physique.ens.fr 19 pages Infomax applied to non linear neurons, in the low noise limit, leads to redundancy reduction.

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

1 Introduction

In the theoretical approaches to the analysis of sensory systems, one would like both to understand the organization of the architecture and of the receptive fields, and to provide models of self-organization which could account for the epigenetic developement. It is natural to consider that the *function* of the processing in the first stages of sensory processing (e.g. in the case of the visual system, the retina and the very first layers behind), is to realize some particular neural representation (or code) of the environment. One possible approach is then to assume that the observed nervous systems, and the self-organization process, result from the optimization of some cost function which characterizes the quality of the code. This defines a whole program: choose a cost function - and possibly a set of relevant constraints -; derive from it the (set of) optimal network(s) - and self-organization algorithms in order to reach an optimal solution-; compare with biological data. The first question is thus to define what could be a relevant (set of) cost function(s).

Already a long time ago[1, 2] it has been suggested that *information theory*[3, 4] could provide appropriate tools. In particular the general ideas developped by Barlow have been at the origin of many theoretical and experimental studies (see e.g. [9, 10]). Barlow[2, 8] insists on the need of building a neural representation that could be easily used in subsequent processing. This leads to the idea of *factorial code*: each output unit should be statistically independent from each other unit. Hence the network decorrelates independent features that are mixed in the input signal. This means that one should minimize the *redundancy* in the neural code, a fact that can be quantified in terms of an information theoretic criterion. For modeling the first layers of the visual system, this redundancy reduction principle has been first studied by Barlow in the case of discrete and noiseless coding, and then by Atick et al[11, 12] for continuous and noisy neural states. This approach has then been systematically developped in order to account for color, scale invariant and stereo perception [13, 14, 15].

A less demanding requirement is that the system should simply maximize the amount of information that the output conveys about the input signal. This suggests in particular a way for modelling how the transfer function of a given sensory neuron is adapted to the particular environment in which the animal lives. Some remarkable experimental tests have been performed, in particular by Laughlin[5] and van Hateren[6], indicating the validity of such hypothesis. The appropriate cost function is taken as the *mutual information*[4] between the output and the input of the network. This idea of "information preservation" has been also developped by Linsker[7] under the name of "infomax principle" in a model of the first layers of the visual system. In Atick *et al*, Linsker and van Hateren studies, a network of linear neurons is studied, and the maximization of the chosen cost function is performed over the synaptic efficacies (receptive fields); the study is done with a Gaussian input distribution, so that analytical results can be obtained. As it has been pointed out, in particular in [20], the prediction of these two criteria can be very similar especially for large signal to noise ratios (depending on the particular constraint which is chosen), and can differ for large noise.

One should note that there are alternative approaches, not necessarily based on information theoretic criteria. In particular, one may ask for the possibility of *reconstructing* the signal from the neural representation. It has been experimentally shown that the detailed statistics of observed spikes can indeed be used to reconstruct with little error the input signal[16]. We note that reconstruction, which is a *decoding* task, can be viewed as a *supervised* learning task, dual of the coding task[19]. One may ask what would be the optimal network if the criterion is to minimize a quadratic error between the reconstructed signal and the true signal. It has been shown [17, 18] that optimization leads to different predictions than those derived from the information theoretic criteria of Linsker and Barlow, in particular for small signal to noise ratios[18].

In the present paper, we will only consider the maximization of information transfer (the *infomax* principle of Linsker), and its relationship with the redundancy reduction of Barlow. Our main concern will be the study of a network with non linear transfer functions. Indeed, most of the papers that we have quoted dealt with linear neurons. May be fewer, or at least less systematic, studies have been devoted to non linear processing. Still, there are works on the optimization of the transfer function[5], on the study of input distributions to which a given transfer function is optimally adapted[21]; on the use of redundancy reduction for binary, more generally discrete, coding[22, 23]; on networks of binary neurons studied with the tools of statistical mechanics[24, 25]; on the effect of a weak non linearity[20], and on neurons with non linear transfer functions in the limit of large output noise[26]. Although this is not said in [26], it is easy to see that in this large noise limit the optimal transfer function is the step function (so that the neuron is a McCulloch and Pitts neuron), with a threshold chosen in such a way that statistically the neuron is equally often "ON" and "OFF".

In the following we will consider the opposite limit, that of a small output noise. Although we will not discuss a specific realistic case, we note that this small noise situation has been considered in theoretical studies applied to the modelling of the first layers of the visual pathway [11]. In a previous work[25], we studied the case of a noiseless perceptron with binary (McCulloch and Pitts) neurons. We showed in particular that for such a network the *infomax* and redundancy reduction principles are equivalent. In the present paper we consider the case of neurons with arbitrary *invertible* transfer functions, in the presence of a small output noise. We will ask what is the consequence of maximizing information transfer, the optimization being both over the synaptic efficacies and over the transfer functions. One outcome of our work is precisely to partly elucidate the origin of the similarity of results obtained with the infomax and the redundancy reduction principles.

The paper is organized as follows. In the following section 2 we formalize the information processing problem in the case of a single neuron with a non linear, bounded, transfer function. We rederive a standard result giving, in the low noise limit, the optimal transfer function for a given signal distribution. Then in section 3 we generalize the derivation to the case of a network with several outputs, each neuron having its own non linear bounded transfer function. We obtain the main result of this paper, showing that, in the limit of a vanishing additive output noise, optimal information transfer is obtained with a factorial code. In the next section, 4, we extend the approach to unbounded (in particular linear) transfer functions, showing that the result for bounded transfer functions remains only under certain conditions. Finally, we show in section 5 how these results subsist in the presence of a small input noise. We discuss the results on the particular case of Gaussian signal distributions, and we also compute the *information capacity* of the network in this particular small noise limit. Perspectives are given in the last section, 6.

2 Transfer of information by a single neuron

In this section we review basic properties of the information transfer by a single neuron with a non linear, e.g. sigmoidal, transfer function f. At each instant of time some signal activates the sensory units, leading to a total postsynaptic potential h at the neuron connected to these input units. The output V of the neuron is given by

$$V = f(h) + z \tag{1}$$

Here f is any nonlinear transfer function which, for simplicity, we will assume to be bounded between 0 and 1, and invertible, as on figure 1a (we will comment shortly later on non invertible cases). We have assumed no input noise, but the presence of some additive output noise z, with a (not necessarily Gaussian) probability distribution ν. The noise strength is measured by, say, the noise variance T:

$$< z^2 > - < z >^2 = T \tag{2}$$

where the brackets means averaging with the ν distribution.

Of interest here is the amount of information conveyed by V about the signal, and we would like to choose the transfer function f in order to maximize information transfer. A basic result is that the information transfer will be maximum, in the vanishing noise limit, if the output distribution is uniform (maximum entropy distribution). This is achieved if what is known in image processing as *sampling* (or *histogram*) *equalization*[27] is performed, that is if (fig. 1):

$$\frac{d}{dh} f(h) = \Psi(h) \tag{3}$$

where $\Psi(h)$ is the probability distribution of h induced by the input distribution.

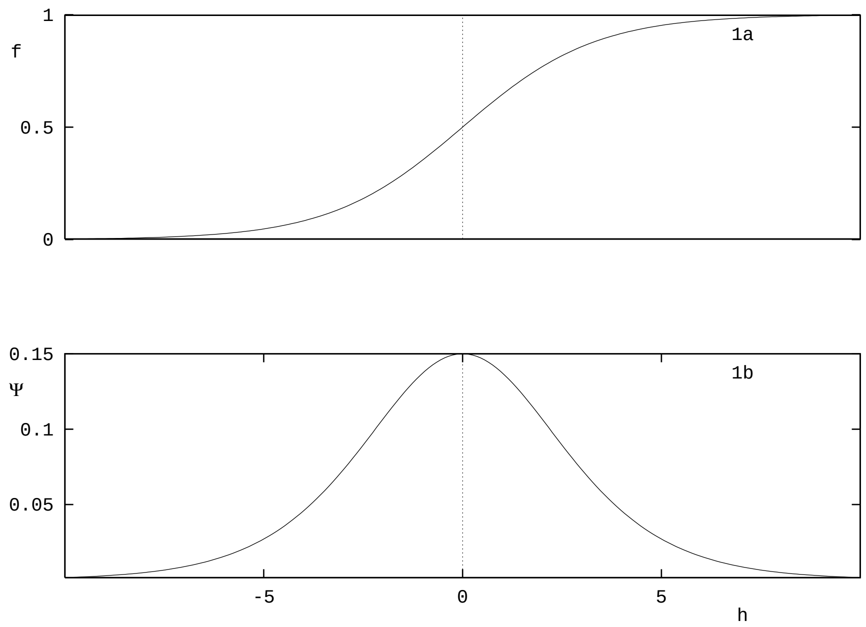

Figure 1: Optimal transfer function $f(h)$ (1a) adapted to the particular input probability distribution shown on 1b, $\Psi = .5\beta / \cosh^2(\beta h)$, with $1/\beta = .3$.

An elegant derivation of what we have just said is given in [5] (see also [11]). The physical meaning is easily understood: a large amount of information is obtained if one can discriminate finely the input signal. If the potential distribution is known through a sample of p values $h^1, h^2, ..., h^p$, more values are observed nearby the h values for which $\Psi(h)$ is large; to discriminate between these, the slope of the transfer function has to be large, in such a way that the outputs are as far apart as possible. This argument is correct because the noise, infinitely small but non zero, provides a resolution scale on the output. There are experimental evidence for adaptations of sensory systems leading to (3) [5].

In order to deal in the next section with several output neurons - and also because this is instructive -, we now formalize a little more this information transfer problem. We thus consider a neuron of output activity $V = f(h)$. In the absence of input noise, the postsynaptic potential h is assumed to be a deterministic function of the input signal (figure 2a), which, for illustrative purpose, one can think of being of the type shown on figure 2b: the neuron is receiving N inputs and, ξ_j denoting the activation of the jth input unit and

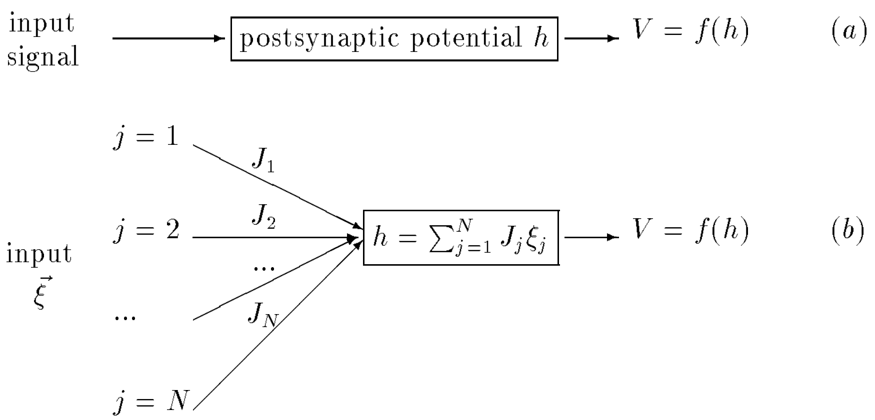

Figure 2: (a) Generic model: the output activity is obtained by applying a non linear transfer function f on the postsynaptic potential h. (b) The simplest neuron model: the potential h is a linear combination of the input components.

J_j the associated synaptic efficacy, the total postsynaptic potential h is

$$h = \sum_{j=1}^{N} J_j \xi_j = \vec{J}.\vec{\xi} \tag{4}$$

However, we emphazise that what follows (as well as in all this paper except partly in section 5), is independent of the particular model considered: the transformation from signal to potential could include non linearities, delays in synaptic transmission, and so on.

The general goal is to choose the synaptic efficacies $\vec{J}$ (and/or any other parameters on which the potential h depend), and the transfer function f in order to maximize information transfer. In this section we only consider the optimization of the transfer function. Given a set of couplings, the input (signal) distribution induces a probability distribution $\Psi(h)$ for the postsynaptic potential h. The quantity of interest is the mutual information between the random variables V and $\vec{\xi}$. In the absence of input noise, this is equal to the mutual information $I(V, h)$ between V and h:

$$I(V, h) = \int dh dV \ \Psi(h) \ Q(V \mid h) \ln \frac{Q(V \mid h)}{Q(V)} \tag{5}$$

where $Q(V \mid h)$ is the conditional probability of observing V knowing the input h, and $q(V)$ the resulting output distribution:

$$\begin{aligned} Q(V \mid h) &= \nu(V - f(h)) \\ q(V) &= \int dh \Psi(h) Q(V \mid h) \end{aligned} \tag{6}$$

Because the noise is additive, the mutual information I can be rewritten as

$$I = H(q) - H(\nu) \tag{7}$$

The first term in (7) is the differential entropy of the probability distribution q:

$$H(q) = -\int dV q(V) \ln q(V) \tag{8}$$

The second term depends only on the noise distribution:

$$H(\nu) = -\int dz \nu(z) \ln(\nu(z)) \tag{9}$$

At this point, one should notice that, whatever the *additive* noise level (and in the absence of input noise), the maximization of the mutual information is equivalent to the maximization of the entropy of the output distribution. In the case of a Gaussian noise, $H(\nu)$ is equal to $\frac{1}{2}\ln(2\pi e T)$; since the Gaussian distribution has the largest entropy among all distributions of a given variance,

$$H(\nu) \leq \frac{1}{2}\ln(2\pi e T) \tag{10}$$

Hence, as T goes to zero, the second term in (7) goes to infinity. What matters is the correction to this infinite constant, that is $H(q)$ which has a finite limit in the zero noise limit, obtained by taking q as

$$q(V) = \int dh \Psi(h) \delta(V - f(h)) \tag{11}$$

In $H(q)$ one can make the change of variable $V \to h$, using:

$$dV q(V) = dh \Psi(h) \tag{12}$$

together with

$$dV = f'(h) dh \tag{13}$$

The result can be written as

$$H(q) = -D(\Psi \mid f') \tag{14}$$

where

$$D(\Psi \mid f') \equiv \int dh \Psi(h) \ln \frac{\Psi(h)}{f'(h)} \tag{15}$$

Since we have $0 < f < 1$, one can consider the derivative of f, f', as a probability distribution. Then one recognizes $D(\Psi \mid f')$ as the Kullback distance [4] (or relative entropy) of the probability distribution $\Psi(h)$ to the probability distribution f'. This quantity is positive or null, and is zero if and only if the two probability distributions are identical (except possibly on a zero measure set). Hence, *maximizing the mutual information is equivalent,* in this low noise limit, *to minimizing the Kullback distance between the input distribution and the one defined from the derivative of the transfer function* - in particular one gets the result announced in (3).

One remark: if we had not chosen f to be with a positive derivative, one would have to replace f by its absolute value in (14), and the result (3) would read

$$\mid \frac{d}{dh} f(h) \mid = \Psi(h) \tag{16}$$

As an alternative solution to (3), one has then the possibility to chose $f' = -\Psi$. For a signal distribution with a single bump (as it is the case for a Gaussian), one obtains an "OFF" cell: its activity is shut down for large inputs.

3 Several output neurons

Now we consider the generalization of the above result, (14), to a network with a number p of outputs. The ith output neuron has the postsynaptic potential h_i, on which acts a transfer function f_i (possibly different from the others). As noted by Atick [11], *if* one realizes a *factorial code*, that is if it is possible to find couplings such that

$$\Psi(\vec{h}) = \prod_{i=1}^{p} \Psi_i(h_i), \tag{17}$$

then one can apply the preceeding reasoning to each output neuron. As a result, the optimal set of transfer functions is given simply by

$$f_i'(h_i) = \Psi_i(h_i), \quad i = 1, ..., p. \tag{18}$$

What we show now is that, with the same conditions as above (no input noise and small additive output noise), a factorial code (17) together with the individual adaptations (18), gives precisely the maximum information transfer.

Our working hypothesis is thus that the output activities are given by

$$V_i = f_i(h_i) + z_i, \quad i = 1, ..., p \tag{19}$$

with an arbitrary noise distribution $\nu(\vec{z})$ (the z_i's need not to be independent random variables). We we will define the noise strength from the total variance:

$$\sum_i (< z_i^2 > - < z_i >^2) = pT \tag{20}$$

Again because the noise is additive, the mutual information can be written as

$$I = H(q) - H(\nu) \tag{21}$$

where now $q = q(\vec{V})$ and $\nu = \nu(\vec{z})$. In the limit $T \to 0$, one can make the change of variable $\vec{V} \to \vec{h}$, with

$$\prod_{i=1}^{p} dV_i q(\vec{V}) = \prod_{i=1}^{p} dh_i \Psi(\vec{h}) \tag{22}$$

and

$$dV_i = f_i'(h_i) dh_i, \quad i = 1, ..., p. \tag{23}$$

This gives

$$H(q) = -D(\Psi \mid \prod_{i=1}^{p} f_i') \tag{24}$$

with

$$D(\Psi \mid \prod_{i=1}^{p} f_i') = \int d\vec{h}\Psi(\vec{h}) \ln \frac{\Psi(\vec{h})}{\prod_{i=1}^{p} f_i'(h_i)} \tag{25}$$

Hence, one finds the direct generalization of (15), giving that the mutual information is, up to a constant, equal to minus the Kullback distance of the potential distribution to the probability defined by the product of the f_i'.

This fact has several important consequences. The main consequence, as announced above, is that the mutual information will be maximized with synaptic efficacies realizing a factorial code (17), together with the individual adaptations of the transfer functions according to (18). Hence, we obtain in particular the remarkable fact that the *infomax* principle of Linsker [7] and the redundancy reduction principle of Barlow [28, 11], which precisely requires to build a factorial code, lead to identical predictions for the receptive fields (within our working hypotheses of zero input noise and low output noise). Note however that it is only the maximization of mutual information which predicts *both* the receptive fields and the transfer functions.

One should notice that *any* factorial code will optimize the information transfer. For example, if one has a Gaussian input distribution and a given number of $p < N$ output units, *any* choice of p different principal components will give the *same* optimal information transfer. This degeneracy comes from the absence of input noise: there is no scale with which to compare the different directions. We thus expect that, when taking into account a small amount of input noise, only the p largest principal components will be selected. What is less obvious is whether the factorial code will remain the optimal choice. We have thus considered the effect of a small input noise, and we present our analysis in section 5.

Another consequence, from the algorithmic point of view, is that the optimization with respect to the couplings, and the adaptation of the transfer functions, may be considered separately: one can first deal with the linear part of the processing (that is the transformation *input* $\rightarrow \vec{h}$, asking for a factorial code for the potential distribution), and then compute the transfer functions from (18). It is remarkable that receptive fields can be predicted from the analysis of a purely linear system, even when non linear processing is taken into account. The application to linear processing of the principle of redundancy reduction *à la Barlow*, as done by Atick et al [12, 11], *precisely in the low noise limit*, can be understood as just a practical way of finding a code which will maximize information transfer. One should point out, however, that dealing separately with the linear and the non linear parts of the processing leads to the optimal solution only if it is indeed possible to find a factorial code for the potential distribution. If this is not the case, it is not obvious whether such strategy will be the most efficient. It would thus be very interesting to study non Gaussian distributions.

The main result of the present section seems to be specific to the case of non linear, bounded, transfer functions. Hence we would like to know what happens for purely linear, and more generally unbounded, transfer functions. This is what we consider in the next section.

4 Unbounded transfer functions

We now assume that the transfer functions are restricted to a class of unbounded functions. Then the optimization of the mutual information I given by (24) has no solution: one must introduce some constraint in order to have a well defined problem. This is well known for linear processing: in such case, the mutual information is equal to the logarithm of the signal to noise ratio, and no upper bound exists unless one restrict the optimization to, say, a given value of the signal variance. In our case, the constraint can be on the couplings, on the potentials distribution, or on the outputs distribution. As one would expect, the optimum will strongly depend on the choice of the constraint. We will study below the effect of different kind of constraints.

For simplicity we first consider a unique output neuron, and we define a cost function by adding to the Kullback distance in (14) a term which enforces a constraint:

$$D_\rho \equiv D(\Psi \mid f') - \rho\, \Gamma \tag{26}$$

The constraint may be on the potential, with for example

$$\Gamma = \Gamma(\Psi) = \int dh \Psi(h) G(h) \tag{27}$$

for some given function G (e.g. $G(h) = h^2$), or on the output,

$$\Gamma = \Gamma(\Psi, f) = \int dV q(V) G(V) = \int dh \Psi(h) G(f(h)) \tag{28}$$

For such constraints, one can rewrite the cost function (26) as

$$D_\rho = \int dh \Psi(h) \ln \frac{\Psi(h)}{\Phi(h)} - \ln Z \tag{29}$$

with the distribution Φ defined by

$$\Phi(h) = f'(h) e^{-\rho G}/Z, \tag{30}$$

Z being the normalization factor which ensures that

$$\int dh \Phi(h) = 1. \tag{31}$$

In the particular case of a constraint on the output distribution as (28), Z does not depend on f:

$$Z = \int dV e^{-\rho G(V)} \tag{32}$$

In any case, one sees that the optimization task is equivalent to the one with an effective bounded transfer function whose derivative is Φ. One can easily check that, for $G = h^2$ and $f(h) = h$, one recovers the standard formula for the mutual information of a linear channel. The generalization to several outputs is straightforward. For constraints such as

$$\Gamma(\Psi) = \int d\vec{h} \Psi(\vec{h}) G(\vec{h}) \;\text{ or: }\; \Gamma(\Psi, f) = \int d\vec{V} q(\vec{V}) G(\vec{V}), \tag{33}$$

one gets an expression analogous to (29) with now

$$\Phi(\vec{h}) \equiv \prod_{i=1}^{p} f_i'(h_i) e^{-\rho G}/Z \tag{34}$$

This shows that, in the general case, Φ is not a factorial distribution, hence the optimal code is not factorial. However, *if* the constraint can be written as a sum of individual constraints, e.g. with

$$G = \sum_i G_i(h_i) \quad \text{or: } G = \sum_i G_i(V_i) \tag{35}$$

then

$$\Phi(\vec{h}) = \prod_{i=1}^{p} \Phi_i(h_i) \tag{36}$$

(with $\Phi_i(h_i) = f_i'(h_i)e^{-\rho G_i}/Z_i$), and the optimal solution is a factorial code. This result allows to understand the similarities and differences between several works based on different criteria. In a forthcoming paper[29], we will compare the model discussed by van Hateren [6] using an information maximization criterion, with the one proposed by Atick *et al* [12, 11] with an approach based on the reduction of redundancy. To conclude this section, maximizing mutual information with linear, and more generally unbounded transfer functions, leads to a factorial code only if the constraint is additive. We note, however, that additive constraints are what is usually considered.

5 Taking into account input noise

5.1 First order correction

We want to see the effect of a non zero input noise of strength Δ. To do so, one has to pay attention to the fact that the limits $T \to 0$ and $\Delta \to 0$ do not commute. Indeed, I is finite whenever any noise is present, whether it is on the inputs or on the outputs. Consider the case of zero output noise and finite input noise: then going from the (noisy) postsynaptic potential to the output is nothing but a (reversible) change of variable, so that the mutual information is equal to the one given by the linear system $\vec{\xi} +$ *noise* $\to \vec{h}$. In that case, considerations of the preceeding section apply. In the present section we are interested in the opposite limit: what we want is the perturbation of the calculation of section 3 at first order in Δ - and we should still have that I goes to infinity as $T \to 0$. This is obtained by computing first the Δ expansion at a finite value of T, and then taking the limit $T \to 0$. We will see that the relevant small parameter is in fact $\frac{\Delta}{T}$.

5.1.1 One output

Let us first consider the case of one output neuron:

$$V = f(h + y) + z \tag{37}$$

where z is the output noise as before, and y the input noise of variance Δ:

$$< y^2 > - < y >^2 = \Delta \tag{38}$$

We assume Gaussian input and output noise. Using the small Δ expansion of the Gaussian:

$$\frac{e^{-h^2/2\Delta}}{\sqrt{2\pi\Delta}} = \delta(h) + \frac{\Delta}{2}\delta\text{"}(h) + O(\Delta^2) \tag{39}$$

one can write the conditional probability (6) as

$$Q(V|h) = Q_0(V|h) + \frac{\Delta}{2}\frac{\partial^2}{\partial h^2} Q_0(V|h) \tag{40}$$

with

$$Q_0(V|h) = \frac{1}{\sqrt{2\pi T}} \exp -\frac{[V - f(h)]^2}{2T} \tag{41}$$

Then one gets, after some algebra, that the mutual information I can be written as

$$I = I_0[T] + \Delta\, I_1[T] \tag{42}$$

where $I_0[T]$ is the mutual information in the absence of input noise, at a finite value of T, and the correction I_1 is given by

$$I_1[T] = -\frac{1}{2}\int dV \ln q_0(V) \int dh \frac{d^2\Psi(h)}{dh^2} Q_0(V|h) \ - \ \frac{1}{2}\int dh \Psi(h) \frac{f'^2(h)}{T}, \tag{43}$$

q_0 being the output distribution in the absence of input noise as given in (6). When T goes to zero, the first term I_0 takes its asymptotic expression obtained in section 2:

$$I_0 = -\frac{1}{2}\ln(2\pi e T) - D(\Psi \mid f'). \tag{44}$$

In I_1, the first term in (43) has a finite limit, so that, for $\Delta << T << 1$, the only relevant term is the second one in (43). Hence, one can write, at first order in Δ/T:

$$I = I_0 - \frac{\Delta}{2T}\int dh \Psi(h) f'(h)^2 \tag{45}$$

Optimization with respect fo the transfer function f gives

$$f'(h) = \Psi(h) + \frac{\Delta}{T}\left[\overline{\Psi^3}\Psi(h) - \Psi(h)^3\right] \tag{46}$$

where

$$\overline{\Psi^3} \equiv \int dh \Psi(h)^3. \tag{47}$$

For this optimal transfer function, one has

$$I = -\frac{1}{2}\ln(2\pi e T) - \frac{\Delta}{2T}\overline{\Psi^3} \tag{48}$$

Hence, as expected, not all Ψ distributions are equivalent. In the particular case of a Gaussian input,

$$\Psi(h) = \frac{e^{-h^2/2\omega}}{\sqrt{2\pi\omega}} \tag{49}$$

one gets

$$\overline{\Psi^3} = \frac{1}{2\sqrt{3}\pi\omega} \tag{50}$$

Hence, the optimal solution is the one which maximizes the potential variance, a result identical to what would be obtained from the optimization of a linear neuron in the limit of small noise.

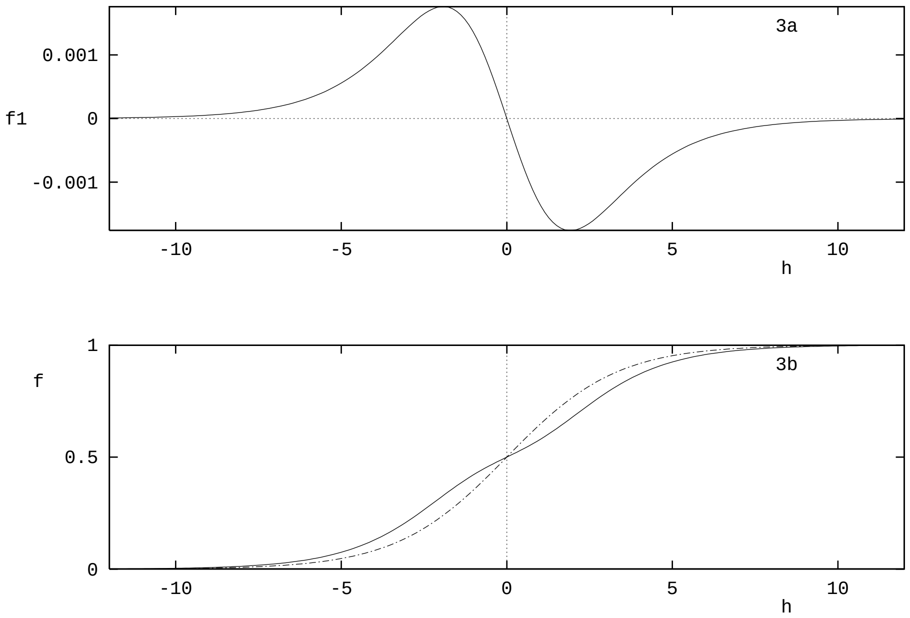

Figure 3: For a small input noise, the optimal transfer function is $f(h) = \Psi(h) + \frac{\Delta}{T} f_1(h)$ (see text). *a)* f_1 for the input distribution $\Psi(h)$ shown on figure 1b. *b)* Full line: the resulting transfer function f; because f_1 is very small, the correction has been artificially enhanced by taking $\frac{\Delta}{T} = 50$. The dashed line is the optimal transfer function in the absence of noise.

5.1.2 Several outputs

Let us consider now the case of several output neurons. Again we will assume Gaussian input and output noise: the output of the ith neuron is given by

$$V_i = f(h_i + y_i) + z_i \tag{51}$$

where z_i is the output noise as before, and y_i the input noise of correlation matrix $\Delta\, C$:

$$< y_i y_{i'} > - < y_i >< y_{i'} >= \Delta\, C_{ii'}, \tag{52}$$

C being $O(\Delta^0)$. We assume *uncorrelated* output noise:

$$< z_i z_{i'} > - < z_i >< z_{i'} >= T\ \delta_{i,i'}. \tag{53}$$

The conditional probability, at any given value of T, is now given by

$$Q(\vec{V}|\vec{h}) = Q_0(\vec{V}|\vec{h}) + \frac{\Delta}{2} \sum_{j,k} C_{jk} \frac{\partial}{\partial h_j} \frac{\partial}{\partial h_k} Q_0(\vec{V}|\vec{h}). \tag{54}$$

One gets the expression for $I_1[T]$:

$$I_1 = -\frac{1}{2} \int d\vec{V} \ln q_0(\vec{V}) \int d\vec{h} \Psi(\vec{h}) \sum_{i,i'} C_{ii'}\ \frac{\partial}{\partial h_i} \frac{\partial}{\partial h_{i'}} Q_0(\vec{V}|\vec{h})\ -\ \frac{1}{2} \sum_i C_{ii} \int dh_i \Psi_i(h_i) \frac{f_i'^2}{T} \tag{55}$$

Taking the limit $T \to 0$, one gets the generalization of (45):

$$I = I_0 - \frac{\Delta}{2T} \sum_{i=1}^{p} C_{ii} \int dh_i \Psi_i(h_i) f_i'^2 \; + \; O(\Delta) \tag{56}$$

where I_0 is the value at $\Delta = 0$. One sees that non diagonal terms, introducing correlations between the i's, appear only at order Δ. Hence, at leading order in Δ/T, the optimal solution is still a factorial code. For a given choice of input distribution $\Psi(\vec{h})$, for each output neuron i the optimal transfer function is given by equation (46) replacing $\Psi(h)$ by the marginal distribution $\Psi_i(h_i)$, and $\frac{\Delta}{T}$ by $C_{ii}\frac{\Delta}{T}$. For this set of optimal transfer functions, the mutual information is then

$$I = I_0 - \frac{\Delta}{2T} \sum_{i=1}^{p} C_{ii} \overline{[\Psi_i]^3} \tag{57}$$

with

$$\overline{[\Psi_i]^3} \equiv \int dh_i [\Psi_i(h_i)]^3 \tag{58}$$

The consequence of this expression (57) is that, in order to maximize the mutual information, one has to find couplings realizing a factorial code such that the $\overline{[\Psi_i]^3}$ are minimized.

One should note, however, that the above result is valid only for uncorrelated output noise: generalization of the calculation to correlated noise is straightforward, and one finds that the term of order $\frac{\Delta}{T}$ introduces correlations between output units. This is to contrast with the zero input noise case, for which the correlations in the output noise does not matter.

5.2 Discussion and comparison with the linear case

5.2.1 Gaussian input distribution

To illustrate the result (56), we consider the particular case of neurons of the type shown on figure 2b, with synaptic efficacies $\{\vec{J}_i, i = 1, ..., p\}$. We assume a Gaussian input distribution with correlation matrix R:

$$< \xi_j \xi_k >_c = R_{jk} \tag{59}$$

with a Gaussian noise $\vec{\nu}$ on the inputs, so that:

$$\begin{aligned} y_i &= \sum_{j=1}^{N} J_{ij} \nu_j \\ < \nu_j \nu_k >_c &= \Delta \; \delta_{j,k} \end{aligned} \tag{60}$$

Then the correlation matrix C is given by

$$C_{ii'} = \delta_{i,i'} \sum_j J_{ij}^2 = \delta_{i,i'} [JJ^T]_{ii}, \tag{61}$$

and one gets after optimization over the transfer functions (equ. (57)):

$$I = -\frac{p}{2} \ln(2\pi e T) + \frac{1}{2} \ln \frac{det JRJ^T}{\prod_i [JRJ^T]_{ii}} - \frac{\Delta}{2T} \frac{1}{2\pi\sqrt{3}} \sum_i \frac{[JJ^T]_{ii}}{[JRJ^T]_{ii}} \tag{62}$$

For $p < N$, maximization of I over the couplings J leads to taking, for the directions of the p vectors $\vec{J}_i$, the p largest eigenvectors of the correlation matrix R. Note that the solution is independent of the scale $[JJ^T]_{ii}$ of the couplings: this is because for any given scale there exists an adaptated transfer function. Fixing the norm of the couplings is here equivalent to fixing the range of the potentials on which the transfer function goes from 0 to 1.

This expression (62) of the mutual information (obtained after optimizing the transfer function) should be compared with the redundancy reduction cost function for a set of linear neurons, under the same conditions of weak noise. More precisely, consider the linear network $\vec{\xi}$ + *noise* $\rightarrow$ *output* = $\vec{h}$ + *output noise*. The particular limit we have considered corresponds to taking a large output noise *as compared to the coupling strength.* If we set to J the global scale of the couplings:

$$\vec{J_i} = J\vec{u_i} \tag{63}$$

where $\vec{u_i}$ has a norm of order 1 (one may choose, e.g., either $\vec{u_i}^2 = 1$ or $\sum_i \vec{u_i}^2 = p$), then T should be defined as

$$T = [\text{output noise variance}]/J^2. \tag{64}$$

The cost function defined in [11] as a measure of redundancy is

$$\Re \equiv \sum_{i=1}^{p} I_i - I \tag{65}$$

where I_i is the amount of information conveyed by the ith ouptut neuron alone. The redundancy $\Re$, which is zero only if one has a factorial code, has to be *minimized*[11] under some appropriate constraint[1]. With the above notations, $\Re$ can be written as

$$\Re = \frac{1}{2} \ln \frac{det[1 + \frac{1}{T} uRu^T + \frac{\Delta}{T} uu^T]}{\prod_i [1 + \frac{1}{T}[uRu^T]_{ii} + \frac{\Delta}{T}[uu^T]_{ii}]} \tag{66}$$

At first order in Δ/T, and in the limit of small output noise, one finds that the term of order Δ/T disappears, being equal to

$$-\frac{\Delta}{2T}\{Tr(uu^T\ ) - \sum_i [uu^T]_{ii}\} = 0. \tag{67}$$

Hence, one has simply, at this order:

$$\Re = -\frac{1}{2} \ln \frac{det uRu^T}{\prod_i [uRu^T]_{ii}} \tag{68}$$

which is minus the second term in (62). However, one has to take into account some constraint. The one which will leads to an expression closely related to the one of (62) is the choice of a constraint on the output variances. Defining ρ as the Lagrange multiplyer needed to enforce a global constraint, one has then to maximize

$$\frac{1}{2} \ln \frac{det uRu^T}{\prod_i [uRu^T]_{ii}} - \rho \sum_i [uRu^T]_{ii}. \tag{69}$$

[1]We note that this criterion should be used with care, since it is not always true that $\Re$ is positive. To give a simple example, consider an input signal which can be in two states, A and B. This signal is redundantly encoded with two binary units, (V_1, V_2), in a "XOR" representation: A is coded with equal probability as (0,1) or (1,0), and B with equal probabilities as (0,0) or (1,1). Then no information is obtained by looking at a single unit, so that $I_1 = I_2 = 0$, whereas there is no loss of information at all, $I = 1$ bit.

Comparing this expression with (62), one sees that ρ plays a role similar to the one of the parameter Δ/T. Hence, it appears that, in this low noise limit, maximization of the mutual information, in a network with non linear transfer functions, leads to essentially the same predictions for the receptive fields as the redundancy reduction criterion of Barlow-Atick applied to a network of linear neurons under some constraint on the output variances.

5.2.2 Information capacity

Finally we address the question of the *information capacity* of the network. We have discussed the adaptation of the network to a given environment, that is a given input distribution. The problem of the information capacity[25] is then to determine the environment for which the network will be the most efficient - the most able to extract information from the signal after optimization of the parameters (synaptic efficacies and transfer functions). This problem is analogous to the one of the channel capacity in information theory[4], except that here we consider a *family* of channels: because the system we consider is allowed to adapt to the environment, the relevant information capacity is the largest channel capacity within the accessible family of networks, each one being characterized by a choice of synaptic efficacies and transfer functions. Knowing the information capacity C, we will know that, whatever the environment, a network even after optimization will not be able to extract more information than C.

We consider the family of networks having p output neurons, with bounded, invertible, transfer functions, as in section 3. Because the optimization requires a factorial code, the information capacity C_p will be equal to p times the capacity C_1 of a single neuron. We can then consider the case of a single output neuron. If the transfer function was fixed, then in the vanishing noise limit the information capacity C_1 would be equal to the mutual information obtained when the relation (3) is fulfilled: the optimal input distribution equals the derivative of the transfer function, and $C_1 = -\frac{1}{2}\ln(2\pi e T)$. A discussion of this relation (3) for some standard transfer functions is given in [21]. Now we allow for the optimization over the tranfer function as well, and we take into account the input noise. Then, according to (48), the optimal environment is the one for which $\overline{\Psi^3}$ is minimal. This optimization problem is ill defined without adding a constraint on the input distribution. We will search the optimal input distribution among those with a given differential entropy:

$$S = -\int dh \Psi(h) \ln \Psi(h) \tag{70}$$

Minimization of $\overline{\Psi^3}$ imposes a flat distribution, hence the optimum is $\Psi = 1/2a$ on the interval $[-a, a]$, where a is obtained from (70):

$$2a = e^S. \tag{71}$$

Then one has the capacity

$$C_1 = -\frac{1}{2}\ln(2\pi e T) - \frac{\Delta}{2T}\, e^{-2S}. \tag{72}$$

For this particular input distribution, the term of order $\frac{\Delta}{2T}$, in the equation (46) for the transfer function, is zero. The associated optimal transfer function is then a ramp (that is f goes linearly from 0 to 1 on the interval $[-a, a]$). It is interesting to compare C_1 with the mutual information for the Gaussian input distribution (49) with a variance ω giving the same entropy S:

$$S = \frac{1}{2}\ln(2\pi e\omega) \tag{73}$$

We have seen that $\overline{\Psi^3} = 1/2\pi\sqrt{3}\omega$, so that we can write the mutual information I_{Gauss} after optimization of the network as:

$$I_{Gauss} = -\frac{1}{2}\ln(2\pi eT) - \frac{\Delta}{2T}\frac{e}{\sqrt{3}}e^{-2S} \tag{74}$$

The dependence of C_1 and I_{Gauss} on S is the same, with a prefactor larger in I_{Gauss} ($\frac{e}{\sqrt{3}} \sim 1.57$).

Remark: if we had chosen the optimal distribution among those having a given variance ω, we would have the same solution, except that in this case a would be given by

$$\frac{a^2}{3} = \omega, \tag{75}$$

leading to

$$C_1 = -\frac{1}{2}\ln(2\pi eT) - \frac{\Delta}{2T}\frac{1}{12\omega}, \tag{76}$$

to be compared with the mutal information for the Gaussian distribution with the same variance:

$$I_{Gauss} = -\frac{1}{2}\ln(2\pi eT) - \frac{\Delta}{2T}\frac{1}{2\sqrt{3}\pi\omega}. \tag{77}$$

In that case the relative prefactor of $1/\omega$ is $\frac{12}{2\sqrt{3}\pi} \sim 1.1$.

It is not surprising to find a flat distribution as the optimal one: in the vanishing noise limit, the optimal output distribution is flat on the interval $[0,1]$; hence the optimal input distribution is nothing but a rescaled version of that distribution. Still, this is to contrast with the case of a linear channel, in which case the optimal input is the Gaussian distribution[4].

6 Conclusion

In this paper we considered the problem of maximizing information transfer with a network of neurons made of N inputs and p outputs, focussing on the case of non linear transfer functions. We assumed that both the transfer functions and the synaptic efficacies could be adapted to the environment.

The main consequence of our analysis is that, in the limit of small *additive* output noise (and an even smaller input noise), the *infomax* principle of Linsker, and the redundancy reduction criterion of Barlow, are equivalent when non linear processing is taken into account. Moreover, this result subsists for linear processing whenever optimization is performed under some constraint which can be written as a sum of terms, each one depending on one output unit only. This explains why the results obtained by Atick et al[30] and by Linsker[20] are so similar. We will detail this comparison also for finite noise in a forthcoming paper [29].

A practical consequence is that optimization of receptive fields, that is of the synaptic efficacies, and of transfer functions can be done separately: one may first look for a linear transformation which realizes a factorial code, and then adapt the transfer functions independently for each output neuron. Of course, this is true only if a factorial code does exists. However, this two step strategy is still valid if one considers that it is sufficient to act *as if* the input distribution was a Gaussian, paying attention only to the mean and covariance of the input distribution (see also [20, 11] for more detailed discussions on the motivation for such an approximation).

In the absence of input noise, *any* factorial code will maximize information transfer. Provided such codes exist (and then, if the number of output units is smaller than the number of inputs, many such codes will exist), one may say that an optimized network is extracting *qualitative* information, looking at statistically independent features. It is only in the presence of a small input noise, which provide a scale for measuring the input signal, that the network can extract *quantitative* information, looking at the most relevant features only.

Although our main results are valid for *any* input distribution, we mainly discussed their consequences assuming that a factorial code could be found, and in particular we have considered the case of a Gaussian distribution. Clearly, it would be very interesting to study the case of non Gaussian distributions - in particular the empirical distributions derived from the analysis of natural scenes [31, 32].

Finally, we note that the same analysis can be done in the time domain. In such case, maximizing information will lead to, again, decorrelation, which in this case has the meaning of *source separation* [33, 34]. Recently an algorithm has been proposed for source separation, based on an *ad hoc* cost function related to the statistical correlations of a set of neuron like units[35]. It would be interesting to see whether similar algorithms could be defined from gradient descent on an information theoretic criterion (mutual information with non linear output units and/or redundancy reduction cost function as in [11] with linear units). Conversely, it would be interesting to see whether the efficient empirical algorithms developped for source separation[33, 35] could be adapted to decorrelation in the spatial domain (known algorithms for spatial inputs assume a Gaussian distribution[36, 37]). Source separation algorithms have also been proposed as odor *coding* algorithms in the olfactory system of insects [38], hence an approach very similar to the one of Barlow and Atick *et al* for the visual system. It seems thus quite plausible that a unique framework - say, maximization of mutual information - could be used to study encoding of spatio-temporal signals, leading to a joined decorrelation in space and time.

Acknowledgements

This work was partly supported by the French-Spanish program ”Picasso”.

References

[1] Attneave F. Informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954.

[2] Barlow H. B. The coding of sensory messages. In W. H. Thorpe and O. L. Zangwill, editors, *Current Problems in Animal Behaviour*, pages 331–360. Cambridge University Press, 1960.

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

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

[5] Laughlin S. B. A simple coding procedure enhances a neuron's information capacity. *Z. Naturf.*, C 36:910–2, 1981.

[6] van Hateren J.H. Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation. *J. Comp. Physiology A*, 171:157-170, 1992.

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

[8] Barlow H. B. and Foldiak P. Adaptation and decorrelation in the cortex. In R. Durbin, C. Miall, and G. Mitchison, editors, *The Computing Neuron*, pages 54–72. Addison-Wesley, Cambridge MA, 1989.

[9] Bialek W., editor. *Princeton Lectures on Biophysics.* World Scientific Pub., Singapore, 1992.

[10] Bialek W., editor. *Princeton Lectures on Biophysics.* NEC, Princeton, 1993.

[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. What does the retina know about natural scenes. *Neural Comp.*, 4:196–210, 1992.

[13] Atick J. J., Li Z., and Redlich A. Understanding retina color coding from first principles. *Neural Comp.*, 4:559–572, 1992.

[14] Li Z. and Atick J. J. Towards a theory of the striate cortex. *Neural Comp.*, 6:127–146, 1994.

[15] Li Z. and Atick J. J. Efficient stereo coding in the multiscale representation. *Neural Comp.*, 1993, to appear.

[16] Bialek W., Rieke F., de Ruyter van Steveninck R., and Warland D. Reading a neural code. *Science*, 252:1854–57, 1991.

[17] Linsker R. Sensory processing and information theory. In Grassberger P. and Nadal J.-P., editors, *From Statistical Physics To Statistical Inference and Back*, pages 237–247. Kluwer Acad. Pub., Dordrecht, 1994.

[18] Ruderman D. Designing receptive fields for highest fidelity. *NETWORK*, 5:147–155, 1994.

[19] Nadal J.-P. and Parga N. Duality between learning machines: a bridge between supervised and unsupervised learning. *Neural Comp.*, 6:489–506, 1994.

[20] Linsker R. Deriving receptive fields using an optimal encoding criterion. In Hanson S. J., Cowan J. D., and Lee Giles C., editors, *Neural Information Processing Systems 5*, pages 953–60. Morgan Kaufmann - San Mateo, 1993.

[21] Chapeau-Blondeau F. Information entropy maximization in the transmission by a neuron nonlinearity. *C.R.A.S.*, 1994, to appear.

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

[23] Redlich A. N. Redundancy reduction as a strategy for unsupervised learning. *Neural Comp.*, 5:289–304, 1993.

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

[25] Nadal J.-P. and Parga N. Information processing by a perceptron in an unsupervised learning task. *NETWORK*, 4:295–312, 1993.

[26] Schuster H. G. Learning by maximizing the information transfer through nonlinear noisy neurons and noise breakdown. *Phys. Rev.*, A46:2131–2138, 1992.

[27] Russ J. C. *The image processing handbook.* CRC Press, 1992.

[28] 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.

[29] Nadal J.-P. and Parga N. Infomax with constrained variances: from noise reduction to decorrelation. *In preparation.*

[30] Atick J. J. and Redlich A. Towards a theory of early visual processing. *Neural Comp.*, 2:308–20, 1990.

[31] Field D. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am.*, 4:2379, 1987.

[32] Ruderman D. and Bialek W. Statistics of natural images: scaling in the woods. In Cowan J. D., Tesauro G., and Alspector J., editors, *Neural Information Processing Systems 6*, pages –. Morgan Kaufmann - San Mateo, 1994.

[33] Jutten C. and Herault J. *Signal Proc.*, 24:1–10, 1991.

[34] Hopfield J.J. -. *Proc. Natl. Acad. Sci. USA*, 88:6462–6466, 1991.

[35] Burel G. Blind separation of sources: a nonlinear neural algorithm. *Neural Networks*, 5:937–947, 1992.

[36] Linsker R. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Comp.*, 4:691–702, 1992.

[37] Atick J. J. and Redlich A. N. Convergent algorithm for sensory receptive field development. *Neural Comp.*, 5:45–60, 1993.

[38] Rospars J.-P. and Fort J.-C. Coding of odor quality: roles of convergence and inhibition. *NETWORK*, 1993, to appear.