International Journal of Neural Systems, Vol. ?, (Supp. 1995) 000–000
Proceedings of the third Workshop: *Neural Networks: from Biology to High Energy Physics*

INFORMATION TRANSMISSION BY NETWORKS OF NON LINEAR NEURONS

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

and

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

Received (to be inserted
Revised by Publisher)

We investigate the consequences of maximizing information transfer in a simple neural network, with bounded and invertible transfer functions. In the case of a vanishing additive output noise, and an even smaller input noise, the main result is that maximization of information (over receptive fields and transfer functions) leads to a factorial code - hence to the same solution as required by the redundancy reduction principle of Barlow.

1. Introduction

In the present paper we consider the *infomax* principle of Linsker[1] (that is, the maximization of information transfer) and its relationship with the *redundancy reduction* principle of Barlow[2]. Our main concern will be the study of the adaptation of a network with non linear transfer functions according to those criteria. Most of the published works deal with linear neurons. In the literature there exists only a few and not systematic studies of non linear processing. Still, there are works on the optimization of the transfer function [3], on the study of input distributions to which a given transfer function is optimally adapted [4]; on the use of redundancy reduction for binary, more generally discrete, coding [5,6]; on networks of binary neurons studied with the tools of statistical mechanics [7,8]; on the effect of a weak non linearity [9], and on neurons with non linear transfer functions in the limit of large output noise[10].

We will consider non linear processing in the limit of a small output noise. Although we will not deal with a specific realistic case, we note that this small noise situation has been discussed in the theoretical modelling of the early visual system [11]. In a previous work[8], 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

[*]Laboratoire associé au CNRS (URA 1306) et aux Universités Paris VI et Paris VII.

are equivalent. Here 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.

In the present paper we just sketch the main results. Details and extensions can be found in [12].

2. Joined adaptation of synaptic efficacies and transfer functions

We consider a feedforward neural network with N inputs and p outputs. We assume the output activities $\{V_i\}, i = 1, ..., p$ to be given by

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

The potentials h_i are deterministic functions of the input signal ξ, and of a set of parameters. A particular case is the one where they are linear functions of the input signal, the parameters being then synaptic efficacies. For simplicity, we will always call the parameters "synaptic efficacies". The input signal has a distribution (not necessarily Gaussian), which induces a distribution $\Psi(\vec{h})$ for the potential.

The z_i's are noise terms with an arbitrary distribution $\nu(\vec{z})$ (the z_i's need not to be independent random variables). Its strength is given by the total variance, $\sum_i (< z_i^2 > - < z_i >^2) = pT$. In this section there is no input noise.

For an additive output noise, the mutual information I between the input and the output code $\vec{V}$, can be expressed as

$$I = H(q) - H(\nu) \tag{2}$$

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

$$dV_i = f_i'(h_i) dh_i, \quad i = 1, ..., p. \tag{3}$$

This gives

$$H(q) = -D(\Psi \mid \prod_{i=1}^{p} f_i') \tag{4}$$

with

$$D(\Psi \mid \prod_{i=1}^{p} f_i') = \int d\vec{h} \Psi(\vec{h}) \ln \frac{\Psi(\vec{h})}{\prod_{i=1}^{p} f_i'(h_i)} \tag{5}$$

Hence, one finds 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 one is that the mutual information will be maximized with synaptic efficacies realizing the factorization

$$\Psi(\vec{h}) = \prod_{i=1}^{p} \Psi_i(h_i), \tag{6}$$

together with the individual adaptations of the transfer functions according to

$$f_i'(h_i) = \Psi_i(h_i), \quad i = 1, ..., p. \tag{7}$$

Hence, we obtain in particular the remarkable fact that the *infomax* principle of Linsker [1] and the *redundancy reduction* principle of Barlow [2,11], 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. We will see in the next section how a small input noise lifts this degeneracy.

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 (7). 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 discussed in [11], *in the low noise limit*, can be understood as just a practical way of finding a code which will maximize information transfer. Note however that, if it is not possible to find a factorial code for the potentials, it is not obvious whether such strategy will be the most efficient.

3. Taking into account input noise

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 $T \rightarrow 0$ and $\Delta \rightarrow 0$ 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* $\rightarrow \vec{h}$. 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 previous calculation at first order in Δ - and we should still have that I goes to infinity as $T \rightarrow 0$. This is obtained by computing first the Δ expansion at a finite value of T, and then taking the limit $T \rightarrow 0$. We will see that the relevant small parameter is in fact $\frac{\Delta}{T}$.

We will assume Gaussian input and output noise: the output of the ith neuron is given by

$$V_i = f(h_i + y_i) + z_i \tag{8}$$

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

$$< y_i y_{i'} > - < y_i >< y_{i'} >= \Delta\, C_{ii'}, \tag{9}$$

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

$$< z_i z_{i'} > - < z_i >< z_{i'} >= T\, \delta_{i,i'}. \tag{10}$$

After some algebra [12] one gets in the $T \rightarrow 0$ limit:

$$I = I_0 - \frac{\Delta}{2T} \sum_{i=1}^{p} C_{ii} \int dh_i \Psi_i(h_i) f_i'^2 \; + \; O(\Delta) \tag{11}$$

where I_0 is the value at $\Delta = 0$ computed in the previous section. Hence, at leading order in Δ/T, the optimal solution is still a factorial code. In the particular case of a Gaussian input distribution, one then finds that the above mutual information is maximized when only the p largest principal components are selected.

4. Conclusions

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 and arbitrary input distributions. 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 implies the *redundancy reduction* principle of Barlow. Moreover we have shown that this result still holds for linear processing whenever infomax is performed under some constraint which can be written as a sum of terms, each one depending on one output unit only [12]. This explains why the results obtained by Atick and coworkers [11] and by Linsker[9] are so similar.

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.

One may say that an optimized network is extracting *qualitative* information, (looking at statistically independent features), and *quantitative* information, looking at the most relevant features only (the input noise providing a scale for measuring the input signal).

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* [13,14]. 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 neuronlike units[15]. Source separation algorithms have also been proposed as odor *coding* algorithms in the olfactory system of insects [16], hence an approach very similar to the one of Barlow for the visual system.

As we suggested [12], it should be interesting to see whether decorrelating algorithms, with a two stage strategy where both transfer functions and synaptic efficacies are adapted, could be defined from gradient ascent on the mutual information with non linear output units. Very recently [17] this use of infomax for decorrelation (but with a fixed transfer function) has been explored in the case of zero noise.

Acknowledgments

This work was partly supported by a Human Capital and Mobility project of the EU and by the French-Spanish program 'Picasso'.

References

1. R. Linsker, " Self-organization in a perceptual network," *Computer*, **21**, 105–17 (1988).
2. H. B. Barlow, " Possible principles underlying the transformation of sensory messages," In W. Rosenblith, editor, *Sensory Communication*, page 217. M.I.T. Press, Cambridge MA (1961).

3. S. B. Laughlin, " A simple coding procedure enhances a neuron's information capacity," *Z. Naturf.*, **C 36**, 910–2 (1981).
4. F. Chapeau-Blondeau, "Information entropy maximization in the transmission by a neuron nonlinearity," *C.R.A.S.*, to appear (1994).
5. H. B. Barlow, T. P. Kaushal, and G. J. Mitchison, "Finding minimum entropy codes," *Neural Comp.*, **1**, 412–423 (1989).
6. A. Redlich, " Redundancy reduction as a strategy for unsupervised learning," *Neural Comp.*, **5**, 289–304 (1993).
7. W. Bialek and A. Zee, " Understanding the efficiency of human perception," *Phys. Rev. Lett.*, **61**, 1512–1515 (1988).
8. J.-P. Nadal and N. Parga, "Information processing by a perceptron in an unsupervised learning task," *NETWORK*, **4**, 295–312 (1993).
9. R. Linsker, " Deriving receptive fields using an optimal encoding criterion," in S. J. Hanson, J. D. Cowan, and C. Lee Giles, editors, *Neural Information Processing Systems 5*, pages 953–60. Morgan Kaufmann - San Mateo (1993).
10. H. G. Schuster, " Learning by maximizing the information transfer through nonlinear noisy neurons and noise breakdown," *Phys. Rev.*, **A 46**,2131–2138 (1992).
11. J. J. Atick, " Could information theory provide an ecological theory of sensory processing," *NETWORK*, **3**, 213–251 (1992).
12. J.-P. Nadal and N. Parga, " Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer," *NETWORK*, **5**, 565–581 (1994).
13. C. Jutten and J. Herault "Blind separation of sources. Part I: an adaptive algorithm based on a neuromimetic architecture" *Signal Proc.*, **24**,1–10 (1991).
14. J. J. Hopfield "Olfactory computation and object perception" *Proc. Natl. Acad. Sci. USA*, **88**, 6462–6466 (1991).
15. G. Burel "Blind separation of sources: a nonlinear neural algorithm." *Neural Networks*, **5**, 937–947 (1992).
16. J.-P. Rospars and J.-C. Fort "Coding of odor quality: roles of convergence and inhibition." *NETWORK*, **5**, 121–145 (1994)
17. A. J. Bell and T. J. Sejnowski "An information-maximisation approach to blind separation and blind deconvolution" to appear in *Neural Computation* (1995).