SENSORY CODING: INFORMATION MAXIMIZATION AND REDUNDANCY REDUCTION

J.-P. NADAL
Laboratoire de Physique Statistique de l'ENS
(C.N.R.S. URA 1306, associated to ENS and Universities Paris VI and Paris VII)
Ecole Normale Supérieure
24, rue Lhomond, F-75231 Paris Cedex 05, France
E-mail: nadal@lps.ens.fr

N. PARGA
Departamento de Física Teórica
Universidad Autónoma de Madrid
Cantoblanco, 28049 Madrid, Spain
E-mail: parga@delta.ft.uam.es

This text on some information theoretic approaches to neural coding is based on lectures given at the summer school "Neural Information Processing", Cargèse, July 1997, and on a talk given at the meeting "Towards a theoretical brain", Les Treilles, April 1995.

1 Criteria for sensory coding

Taking over the original ideas of H. Barlow (1960) and F. Attneave (1954), a systematic approach to the modeling of sensory systems is being developed since the last ten years. The general scheme is as follows:

1. bet on what is the task fulfilled by the particular sensory system under consideration;

2. define a criterion (an objective function) that characterizes the performance of a system performing (or trying to perform) this task;

3. compute the optimal performance that could be obtained, from a mathematical point of view, given the signal to noise ratio (in particular taking into account the noise at the level of receptors) and other constraints specific to the studied system;

4. and eventually, compare with experimental data from neurophysiology - e.g. on receptive fields and/or the "preferred" stimulus of neural cells -, and also with some psychophysical empirical data such as contrast sensitivity curves.

From step (3), one may also hope to derive plausible learning mechanisms that may lead to this organization, providing models for the epigenetic development.

Step (1) may be relatively easy when dealing with simple animals. For instance one finds motion detectors in the fly visual system, which provide velocity estimations readily used by the motor system. In such case (see e.g. W. Bialek et al, 1991) the Bayesian inference framework (step 2) allows to define the optimal estimator, given the statistics of the signal and the noise level in the receptors (step 3). A different situation is when dealing with, say, the human visual system. Even though independent channels exists, it is clear that many different tasks have to be solved from the same incoming optical flow. One may thus assume that the first layers in the sensory pathway are building a non specific neural representation, or "code", a priori efficient for further processing. Yet, this hypothesis is not enough for specifying an objective cost function (step 2). Indeed, various criteria have been proposed in the literature, among which several based on information theoretic criteria (for an introduction to Information Theory, see e.g. Blahut 1988). The simplest one, studied by various authors, is what has been called the "infomax principle" by R. Linsker (1988): one ask for a neural network which will maximize the *mutual information* between the output (the neural representation) and the input (say the visual stimuli). The receptors and neural noises, and the finite amount of available resources (number of neurons, synaptic resources) limit the amount of information that can be conveyed by the network on the input, and this limitation renders the maximization a conceptually interesting problem and a generally difficult practical task.

Barlow's proposal was qualitatively different. According to him, it is not only the preservation of information that matters, but more importantly the information *presentation*: the neural code should be easily readable by the system behind. This imply a compression of information (one should take advantage of the regularities in the stimuli, coding only what makes each stimulus unique), and the search for a code where each neuron is coding for features statistically independent from those coded by the other neurons. These aspects are subsumed in the notion of "redundancy reduction", and the optimal code that achieve redundancy reduction is a *factorial code*.

It is worth emphazing the differences between these two criteria. Infomax is a *quantitative* criterion, based on the measure in bits of the statistical relationship between the input and the output of the network. Asking for a factorial code is a *qualitative* requirement, on the statistical relationship between the output neurons. Departure from factorization is however quantified by a *redundancy* cost function (to be minimized) expressed in term of information quantities (Barlow 1989, Atick 1992).

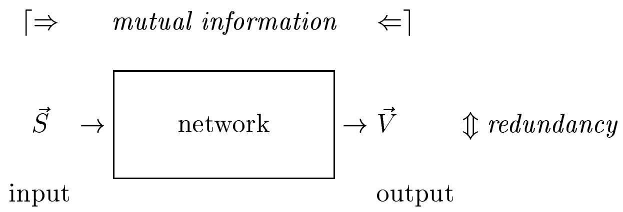

Figure 1: *At each time step the environment provides an input* $\vec{S} = \{S_j, j = 1, ..., N\}$ *to the network, which itself produces an output* $\vec{V} = \{V_i, i = 1, ..., p\}$. *The* mutual information *between the two random variables* $\vec{S}$ *and* $\vec{V}$ *quantifies how much information the output conveys about the input. The* redundancy *is a measure of the statistical dependency between the output activities* $\{V_i, i = 1, ..., p\}$.

2 The linear-Gaussian case

The most detailed analytical studies have been performed for simple feedforward linear networks, with Gaussian input distributions, for both the *infomax* principle (Linsker 1988, van Hateren 1992) and various implementations of the redundancy reduction principle (see e.g., Barlow et al 1989, Atick 1992, Redlich 1993, Li and Atick 1994). Taking account some features specific to particular visual systems, the results have been applied to the visual system of the fly (van Hateren 1992) and to mammalian retina (Atick 1992). Some of these works include a model for the neural code in V1, which takes into account altogether contrast, color and motion sensitivities, as well as stereo-vision, in a multiscale representation (Li and Atick 1994). The predictions from these calculations are in qualitative agreement with known facts on RF of ganglion and V1 cells, and in some cases in quantitative agreement with contrast sensitivity curves obtained in psychophysical experiments (Atick 1992). One can note also that these predicted contrast sensitivity curves are qualitatively very similar to those obtained from a completly different approach, where the retina is modeled as a linear electric filter, the architecture of it being based on the detailed knowledge we have on the retina organization (see J. Herault 1997).

From the theoretical point of view, what is striking is the extreme similarity between the predictions derived from these different criteria: clearly one cannot claim to have pointed out to a basic organization principle if other criteria lead to almost identical results! One may think that the similarity in the results is due to the use of a linear processing onto a Gaussian distribution. In fact, one can see that, in that case of a linear network with Gaussian inputs, any "reasonable" criterion will lead to a principal component analysis, with details depending on the particular constraints under which optimization is performed. Still, the linear-Gaussian system remains quite interesting. One can show (Del Giudice et al, 1995) that the maximization of mutual information, with given input and output additive noises, leads to the following features:

1. there exists a large family of equivalent solutions (one solution being characterized by a particular choice of synaptic couplings - or RF -); this freedom allows for taking into account various constraints if needed;

2. the optimal processing, performed by the network onto the input signal with any of these solutions, amounts to performing two steps: (1) a redundancy reduction, finding the m largest principal component (m depending on the noises levels and on the constraints); (2) a redundancy increase in order to increase the signal-to-noise ratio in each one of these m channels, allocating a specific amount of resource to each component (e.g. several neurons), again according to the noises levels and to the constraints.

It is interesting to see that one may choose a compact solution with exactly m output neurons, or distributed solutions with any number (at least equal to m) of output neurons. This may be relevant for the understanding of the huge increase in the number of cells from LGN to V1.

3 Infomax for a single nonlinear cell

However, as soon as one takes into account any nonlinear aspect in the processing (in particular the saturation of the transfer functions), one readily finds (Linsker 1988) that the large freedom we had in the choice of the solution disappears. It is thus quite important to understand the role of the nonlinearities. Furthermore, one may ask for the optimization in the choice of nonlinearities. On this aspect, a remarckable work, both theoretical and experimental, has been performed by Laughlin (1981) on contrast coding in the fly visual system. Considering a single cell responding to a local contrast (so the input to the cell is a single scalar), Laughlin derived the optimal transfer function from an information theoretic point of view, and showed that this prediction compares very well with its experimental measurement of the cell respons curve. The theoretical result is that, in order to maximize the amount of information conveyed about the input signal, the cell should perform what is known in image processing as "sampling equalization": for a bounded output activity (say between 0 and 1), the derivative of the transfer function should be equal to the probability distribution function (p.d.f.) of the input (here the contrast), so that every possible output occurs with equal probability (Laughlin 1981).

4 Infomax leads to redundancy reduction

The next step is to consider an array of output neurons, coding for a multidimensional stimulus. It turns out that a very general statement can be derived for feedforward networks with non linear transfer functions and arbitrary input (signal) distributions. We have shown (Nadal and Parga 1994) that, in the low noise limit, the maximization of mutual information between the input and the output, if performed over *both* the synaptic efficacies and the choice of the transfer functions, leads to a factorial code - hence to redundancy reduction à la Barlow! To be more explicit, consider the simplest feedforward network, where each output neuron has its activity equal to a (specific) nonlinear transfer function applied to a post-synaptic potential (PSP). This PSP is a linear superposition of the inputs, the coefficients being the synaptic efficacies. Our result states that the mutual information will be maximum if: (1) the linear part of the processing, that is the choice of the synaptic efficacies, is such that the PSP's (hence the activities of the output cells) are statistically independent; (2) and for each cell the transfer function is chosen according to the sampling equalization rule.

Interestingly, this result is related to studies in signal processing on *blind source separation* (BSS) (which is decorrelation in the time domain, see e.g. Comon, 1994). In particular, it implies that the mutual information can be used as a cost function for performing blind source separation (Nadal and Parga 1994). This has been turned into algorithms showing promising performance on particular BSS applications (Bell and Sejnowski 1995). It has been then realized that this infomax approach, for that BSS case, is equivalent to a maximum likelihood approach (Gaeta and Lacoume 1990, Pham *et al* 1992), since they both lead to the very same cost function (Cardoso 1997).

Recently, we have also considered feedforward networks with arbitrary *stochastic* output activities (Nadal, Brunel and Parga 1997). We have shown that the result on factorization remains valid in the limit of vanishing input noise, whenever it is the *probability distribution* of each output (and not the activity itself) which depends on a (cell dependent) deterministic function of the input: the maximization of the mutual information between the input and the output, over the choice of these deterministic functions, leads to a factorial code. This is an important extension since it shows that the result applies in particular in the more realistic case of spiking neurons.

One should emphasis that this factorization, as part of the optimal solution when information maximization is performed, occurs also with *any non linear processing* before the output layer. This means that, for any input distribution being a non linear mixture of independent spatio-temporal signals, there exists a nonlinear network, with one or more hidden layers, with which the output will convey as much information as possible by providing a factorial representation. This representation, although not unique, reflects directly the statistical structure of the input data. In addition, the infomax cost function, already tested for the simplest networks (that is for feedforward networks with no hidden layer, see Bell and Sejnowski 1995), leads to simple unsupervised backprobagation algorithms for multilayer networks (Nadal and Parga 1997).

5 Conclusion

To conclude, we have shown that, at least in the low input noise limit, maximization of information implies factorization. However, understanding the possible consequences of the infomax principle requires a much better understanding of the statistical structure of the signal (in the case of the visual system, one may ask for the structure at the "pixel" level, or may be at the "object" level.

It remains also to understand what subsists when one consider noisy inputs. A reasonnable guess is that one will get a result similar to the one discussed for the linear-Gaussian network: infomax may lead to redundancy reduction, that is to the factorization of the input signal into its independent components, together with a redundancy increase in each one of these independent channels.

Finally, we stress that, although the infomax principle offers a simple unifying framework for studying sensory coding, it is certainly not the only possible approach. An interesting alternative approach has been proposed by Olshausen and Field (1996), who argue that sparseness might be the relevant quality criterion for the neural code. We note, however, that the requirement for sparse coding might be considered as a constraint in an infomax approach, which should be easily taken into account at least in the linear case (as we have seen, optimal solutions exists with a large number of ouput cells).

Acknowledgements

The most recent part of our work presented here is the result of a collaboration with Nicolas Brunel (LPSENS).

References

1. F. Attneave. Informational aspects of visual perception. *Psychological Review* **61** 183–193 (1954).
2. J. J. Atick, Could information theory provide an ecological theory of sensory processing. *Network: Computation in Neural Systems* **3** 213–251 (1992).
3. H. B. Barlow 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).
4. H. B. Barlow, T. P. Kaushal, and G. J. Mitchison. Finding minimum entropy codes. *Neural Computation* **1** 412–423 (1989).
5. A. Bell and T. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation* **7** 1129–1159 (1995).
6. W. Bialek, F. Rieke, R. de Ruyter van Steveninck, & D. Warl, Reading a neural code, *Science* **252** 1854–57 (1991).
7. R. E. Blahut. *Principles and Practice of Information Theory.* Addison-Wesley, Cambridge MA (1988).
8. J.-F. Cardoso. Infomax and maximum likelihood for blind separation. *IEEE Signal Processing Letters* **4** 112-114 (1997).
9. P. Comon. Independent component analysis, a new concept ? *Signal Processing* **36** 287–314 (1994).
10. P. Del Giudice, A. Campa, N. Parga, and J.-P. Nadal. Maximization of mutual information in a linear network: a detailed study, *Network: Computation in Neural Systems* **6** 449–468 (1995).
11. M. Gaeta & J.L. Lacoume. Source separation without apriori knowledge: the maximum likelihood approach *Signal Processing V*, proceedings of EUSIPCO 90, L. Tores, E. MasGrau & M.A. Lagunas eds, pp 621-624 (1990).
12. J. H. van Hateren, Theoretical predictions of spatio-temporal receptive fields of fly LMCs, and experimental validation. *J. Comp. Physiology A* **171** 157-170 (1992).
13. J. Herault, This volume (Cargèse summer school “Neural Information Processing”, June 30-July 12, 1997).
14. S. B. Laughlin, A simple coding procedure enhances a neuron's information capacity, *Z. Naturf.* **C36** 910-2 (1981).
15. Z. Li and J. J. Atick, Efficient stereo coding in the multiscale representation. *Network: Computation in Neural Systems* **5** 1–18 (1994).
16. R. Linsker, Self-organization in a perceptual network. *Computer* **21**

105–17 (1988).

17. J.-P. Nadal & N. Parga, Nonlinear neurons in the low noise limit: a factorial code maximizes information transfer *Network: Computation in Neural Systems* **5** 565-581 (1994).
18. J.-P. Nadal & N. Parga, Redundancy reduction and independent component analysis: Algebraic and adaptive approaches. *Neural Computation* **9** 1421-1456 (1997).
19. J.-P. Nadal, N. Brunel & N. Parga, Nonlinear feedforward networks with stochastic ouputs: infomax implies redundancy reduction. LPSENS preprint (1997), to appear in *Network: Computation in Neural Systems.*
20. B.A. Olshausen & D. J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381** 607-609 (1996).
21. D.-T. Pham, Ph. Garrat & Ch. Jutten Separation of a mixture of independent sources through a maximum likelihood approach. in *Proc. EUSIPCO*, pp 771–774 (1992).
22. A. N. Redlich. Redundancy reduction as a strategy for unsupervised learning. *Neural Computation* **5** 289–304 (1993).