Information theoretic approach to neural coding and parameter estimation: a perspective

Jean-Pierre Nadal

Laboratoire de Physique Statistique de l'E.N.S.*
Ecole Normale Supérieure
24, rue Lhomond - 75231 Paris cedex 05, France
nadal@lps.ens.fr
http://www.lps.ens.fr/~nadal/

Abstract

In this contribution I put in perspective results on neural coding and parameter estimation derived from general principles based on information theoretic criteria or, equivalently, on the Bayesian framework.

I describe generic behaviours in the amount of information conveyed by a network, as a function of the number p of coding cells: the mutual information between input stimuli and coding cells activities grows at best linearly with p for p small, whereas for large p the growth is typically logarithmic. I show how the Bayesian approach to parameter estimation leads to the very same formalism, p being in this context the number of observations, and the relevant quantity the mutual information between parameters and observations. I explain the linear and logarithmic behaviours of the mutual information in terms of code redundancy in the neural coding context and learning performance in the parameter estimation context.

I also propose some lines of research in order to extend such approaches, in particular the *infomax* approach, to the analysis of efficient sensory-motor coding.

This document is to appear as a chapter of the book
Statistical Theories of the Brain, edited by R. Rao, B. Olshausen, and M. Lewicki, MIT press 2000 (proceedings of the NIPS98'Workshop on Statistical Theories of Cortical Function).

*Laboratory associated with C.N.R.S. (U.M.R. 8550), ENS, and Universities Paris VI and Paris VII

1 Introduction

The idea that Shannon's Information Theory (Shannon and Weaver, 1949; Blahut 1988) is relevant for studying neural coding goes back to Attneave, 1954, and Barlow, 1960. Despite some very important early works (e.g. Stein, 1967; Laughlin, 1981), it has received considerable attention only since the late 80's (after the works of, e.g., Linsker, 1988; Barlow et al 1989; Bialek et al, 1991; Atick, 1992; van Hateren, 1992, and with many other works since then). The relevance of Information Theory to the field of parameter estimation has been acknowledged much more recently. On the one hand it has been shown that, in the Bayesian framework, the Bayes cumulative risk is precisely equal to the mutual information between parameters and data (see e.g. Clarke and Barron 1990; Haussler and Opper, 1995; and references therein). On the other hand, a simple relationship between the *Fisher Information*, a basic quantity in Estimation Theory, and the (Shannon) mutual information between parameters and data, has been shown to exist only very recently (Clarke and Barron, 1990; Rissanen, 1996; Brunel and Nadal 1998). At the same time, the fact that neural coding and parameter estimation are intimately related subjects has been realized. In particular a duality has been shown between neural architectures which allows to translate a learning (or parameter estimation) task into a sensory coding problem (Nadal and Parga, 1994a).

In the next section I introduce the general information theoretic framework which allows to discuss both parameter estimation (from the Bayesian point of view) and neural (especially sensory) coding. Then I present in section 4 the generic behaviour of the mutual information between parameters and data (equivalently between stimuli and neural code), putting into perspective results published in the literature in both fields. In the last section I evoke current and possible lines of research for going beyond the feedforward sensory coding problem.

2 General framework

2.1 Neural coding

One possible approach to the modelling of the building of neural representations, or "neural codes", is illustrated on Figure 1. The environment is assumed to produce at each instant of time a new stimulus Θ with some probability distribution $\rho(.)$. Depending on the particular problem considered, Θ might be, say, the orientation of a bar presented in the visual field, or the multidimensional pattern elicited at the input of the neural system (e.g., the activities of the photoreceptors). The "neural code" is given by the activities $D = \{d^1, ..., d^p\}$ of p neural cells. The activation rule is given by some probability distribution $P(D|\Theta)$, which depends on some set of parameters W (e.g. the set of synaptic efficacies).

The neural representation D is further processed by the neural system, and it may be the case - but this is not necessarily the case - that there is *reconstruction* of the stimulus, that is an *estimation* $\hat{\Theta}$ is produced as a result of the processing of D.

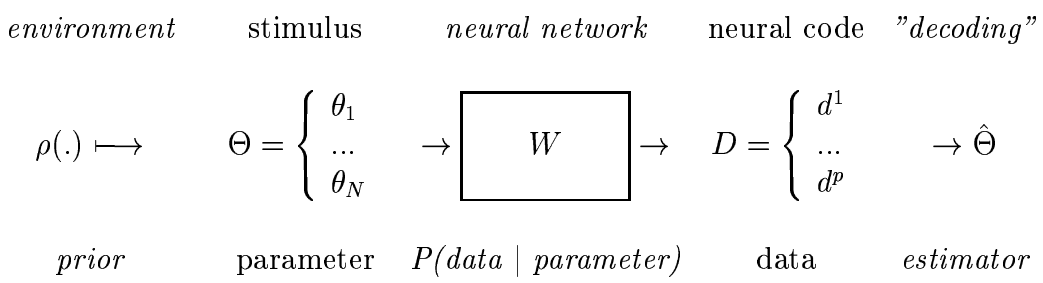

Figure 1: A single framework for neural coding and parameter estimation (see text).

One should note that the use of the name *code* in this context of neural modelling is generally not accepted by information-theorists for whom a code is *defined* as a *deterministic* mapping. The point of view taken in this paper is that a neural representation, or neural code, is the appropriate concept which generalises the engineer's deterministic code to the case of intrinsically noisy systems.

2.2 Parameter estimation

To study a parameter estimation task, one can make use of the very same formalism. Here the quantities D, p and Θ will be given a different interpretation as explained below.

In the context of parameter estimation, one has some data D, a set of p observations $D = \{d^1, ..., d^p\}$. In the simplest setting, one assumes that the probability distribution from which the data have been generated is known, except for some (possibly multidimensional) parameter Θ which is thus the parameter to be estimated from the data. In the general case, the probability distribution $P(D|\Theta)$ may depend on additional parameters W, called *hyper-parameters* if they are considered as additional degrees of freedom useful for modelling the unknown distribution, or *nuisance* parameters if they correspond to parameters of the true model in which one is not interested. In the standard Bayes framework, one considers that the unknown parameter Θ has been chosen by Nature according to some *prior* distribution $\rho(.)$, so that Figure 1 illustrates the Bayes framework when the goal is to compute an estimate $\hat{\Theta}$ of the parameter (another important Bayesian task is to *predict* the next datum, d^{p+1}, but we will here mainly focus on the parameter estimation task). In practice the prior $\rho(.)$ is typically taken as the maximum entropy distribution taking into account all known constraints on the parameter. The neural coding framework presented here can thus be considered as a particular case of Bayesian modelling where there is a *true* parameter (the actual stimulus) and a *true* prior distribution (the environment distribution).

2.3 Other related points of view

The general setting $\rho(.) \longmapsto \Theta \rightarrow D$ introduced here corresponds also to recognition and generative models in data analysis. In a recognition model the data Θ are presented to a network (more generally a learning machine) which produces, say, the independent components (hence in that case Θ plays the role of data, and D the role of independent components). In a generative approach the data D is considered as having been produced by some random variables, Θ. The framework applies as well to signal processing, Θ being the signal, and D the filtered signal.

In this general framework, our interest is in the characterisation of typical and optimal performances making use of information theoretic concepts (for an introduction to Information Theory see Blahut 1988 and Cover and Thomas 1991). This information theoretic approach, and the information theoretic criteria (*infomax*, *redundancy reduction*), are introduced in the next section. Before that let us consider some specific examples of models.

2.4 Specific examples

In this section several specific examples are briefly presented. Each model is introduced by giving the particular definitions of N, Θ, p and D which make the model fit into the framework presented in Figure 1.

First some neural coding models:

(1) early visual system: infomax and redundancy reduction approaches to neural coding in the early visual system (Linsker 1988 and 1993, van Hateren 1992, Atick 1992, Li and Atick 1994a, Li 1995) assume $p \sim N$ large. Here Θ is the set of activities of the photoreceptors, and D the resulting activities of the ganglion cells or of V1 cells.

(2) early visual system: with a *sparse coding* hypothesis for V1 one takes $p > N$, and a code is built such that the number of active output cells for any given input is small (Field 1987, Olshausen and Field 1997, and this book, chapter by Olshausen).

(3) Independent Component Analysis (ICA; for general references on ICA see Jutten and Herault 1991, Comon 1994, and the web site ICA Central, http://sig.enst.fr/~cardoso/icacentral/). ICA can be formulated in various ways, we give here the recognition and generative approaches. Recognition model: Θ is the signal, W the network parameters to be adapted in order to have $\{d^1, ..., d^p\}$ as estimates of the independent components (see Nadal and Parga 1994, Bell and Sejnowski 1995 for the *infomax* approach). In a generative approach, Θ is the set of IC's, and D the observed signal (this corresponds to the maximum likelihood approach to ICA, see Cardoso 1997). Both points of view can be put together if one consider Θ as the set of IC's, D the observed signal and $\hat{\Theta}$ as the estimate of Θ.

(4) *population coding*: Θ is a low dimensional field, e.g. an angle, $N = 1$, and the number p of coding cells is very large, each cell responding with a specific *tuning curve* centred at some particular value of the stimulus; in the simplest models the cells activity are given by Poisson processes. In Seung and Sompolinsky 1993 population

coding is for the first time approached with a parameter estimation point of view (the neural activities D being the data from which the system must extract an estimate $\hat{\Theta}$ of the stimulus). Their analysis is based on the *Fisher information*, whereas the present (Shannon) information theoretic approach is developed in Brunel and Nadal 1998 (see below, section 4.2). For other approaches and references on population coding see also this book, chapter by Zemel and Pillow.

Let us now consider three specific examples in the parameter estimation context:

(1) One is given a data set, $D = \{d^k, k = 1, ..., p\}$, the d^ks being real numbers i.i.d. drawn from a Gaussian distribution with zero mean and unknown variance Θ which one wants to estimate.

(2) Supervised classification task to be performed by a simple perceptron: the data are $d^k = \{\xi^k, v^k\}, k = 1, ..., p$ with ξ^k an N-dimensional pattern, randomly picked with say, some Gaussian distribution, and $v^k = \pm 1$ the binary target. Θ is the N-dimensional perceptron coupling vector which has generated the data; for a deterministic rule, $v^k = \text{sgn} \sum_{j=1}^{N} \Theta_j \xi_j^k$.

(3) Unsupervised learning task: the goal is to estimate an N-dimensional direction Θ from the observations of p N-dimensional vectors, $D = \{d^k = \xi^k, k = 1, ..., p\}$, independently drawn from some probability distribution that depends on the parameter only through the scalar product of Θ with the pattern: $P(d^k|\Theta) = p(\lambda^k \equiv \sum_{j=1}^{N} \Theta_j \xi_j^k)$.

These models (2) and (3) have been intensively studied (see in particular Biehl and Mietzner 1993, Watkin and Nadal 1994, Reimann and Van den Broeck 1996, Van den Broeck 1998, Buhot and Gordon 1998).

In all the following I will make equivalent use of the terminologies associated with either the parameter estimation or the neural coding contexts: hence I will call the output D either the "data" or the "neural code", and the input Θ either the "parameter" or the "stimulus" (or the "signal").

3 Information Theoretic Framework

3.1 The Mutual Information between data and parameter

In all the cases mentioned above (neural coding, parameter estimation...) the main quantity of interest here is the *Mutual Information* $I[\Theta; D]$ between the two random variables, the input Θ and the output D. This Shannon information quantity is defined as the Kullback divergence between the joint and the product distributions of $\{\Theta, D\}$, and can be written as (see e.g. Blahut 1988):

$$I[\Theta; D] = \int d\Theta \rho(\Theta) \int dD P(D|\Theta) \ln \frac{P(D|\Theta)}{P(D)} \tag{1}$$

where $P(D)$ is the marginal probability distribution of D,

$$P(D) = \int d\Theta \rho(\Theta) P(D|\Theta). \tag{2}$$

The mutual information $I[\Theta; D]$ measures the number of bits (if we take base 2 logarithms in 1) conveyed by D about Θ. This is an a priori relevant quantity for neural coding, but also for parameter estimation: intuitively it measures all the information actually available in the data to be used for estimating the parameter. Without specifying in advance any particular estimator, the best possible performance that can ever be achieved in this estimation task is controlled by the mutual information. Indeed, the basic but fundamental information processing theorem tells us that processing cannot increase information (see e.g. Blahut 1988), that is, whatever the estimator (the algorithm for computing $\hat{\Theta}$ from D):

$$I[\Theta; \hat{\Theta}] \leq I[\Theta; D]. \tag{3}$$

The mutual information $I[\Theta; D]$ can also be understood as the Bayes risk in the probability estimation task (for the prediction of d^{p+1}), that is the cumulative entropy loss assuming that the true parameter has been generated by Nature according to $\rho(.)$ (see Haussler and Opper 1997).

The *information channel capacity*, or simply the *capacity*, $C = C(W)$ is defined as the supremum of the mutual information over all possible choices of input distributions $\rho(.)$, the parameters W of the channel being fixed:

$$C = \max_{\rho} I[\Theta; D] \tag{4}$$

In the neural coding context, it is the maximal amount of information that the network can convey, whatever the statistics of the environment.

For parameter estimation, it gives the *minimax risk*, that is the smallest possible worst case risk in the Bayesian framework (see Haussler and Opper 1997 and references therein). It is related to Vapnik's growth function (Vapnik 1995) considered in computational learning theory. To see this, consider the simplest case of a supervised binary classification task by a perceptron (example (2) above). The choice of couplings Θ is equivalent to a choice of a particular dichotomy. The total number Δ of realizable dichotomies is a function of p and N alone - not of the set of patterns to be classified. The mutual information is then upper bounded by the logarithm of the number of possible classifications, that is by $\ln \Delta$ (which is the growth function for the perceptron). The maximum of the mutual information is reached for ρ giving the same weight to all possible classifications, hence

$$C(\text{perceptron}) = \ln \Delta(N, p) \tag{5}$$

(for more details see Nadal and Parga 1993 and 1994a, Opper and Kinzel 1995).

3.2 Infomax, Redundancy and ICA

One specificity of neural coding is the adaptation of the system to a *given* environment $\rho(.)$. A possible criterion for efficient coding is then the maximisation of mutual information over the choice of system (network) parameters W - a criterion called *infomax* after Linsker, 1988:

$$\text{(Infomax)} \quad I_{\max}[\rho] = \max_{W} I[\Theta; D] \tag{6}$$

Optimal coding according to this criterion has been studied by various authors, see in particular Stein 1967, Laughlin 1981, Linsker 1988, van Hateren 1992, Nadal and Parga 1993, Stemmler and Koch 1998.

An alternative is the *minimisation of redundancy* based on the original ideas of F. Attneave (1954) and H. Barlow (1960). The redundancy considered here is the redundancy R in the neural code, that is the mutual information between the p output units. It is thus defined as the Kullback divergence between the joint probability distribution of the code D and the factorial distribution of its components $d^1, ..., d^p$:

$$R = \int \prod_{k=1}^{p} dd^k P(D) \ln \frac{P(D)}{\prod_{k=1}^{p} P_k(d^k)} \tag{7}$$

with $P_k(d^k)$ the marginal probability distribution of d^k. The redundancy is zero iff the probability distribution $P(D)$ is equal to the product of the marginals (except possibly on a set of zero measure),

$$P(D) = \prod_{k=1}^{p} P_k(d^k) \tag{8}$$

in which case one speaks of a *factorial code*. If such a code is achieved, each coding cell is carrying some information statistically independent of the one conveyed by the other cells. In the signal processing language, this is equivalent to stating that the system is performing an Independent Component Analysis (ICA) of the input signal. Barlow's proposal is thus that the efficient coding scheme (again for a given environment ρ) corresponds to adapt the parameters W in order to minimise the redundancy:

$$\text{(Barlow's principle)} \quad \min_{W} R \tag{9}$$

The minimisation of redundancy has been considered for the modelling of the early visual system (Atick and coworkers 1992, Redlich 1993, Li and Atick 1994a and 1994b). The fact that (7) is an appropriate cost function for performing ICA has been also recognised in the signal processing context (Comon, 1994).

In many cases, the output distribution *given the input* is factorised:

$$P(D|\Theta) = \prod_{k=1}^{p} P_k(d^k|\Theta) \tag{10}$$

This is so for the most studied case in parameter estimation, where one considers the data to be statistically independent realizations of the same law ($P_k(.|\Theta)$ is in addition independent of k); this is also the case for multilayer feedforward neural networks, where each output d^k is a (possibly stochastic) function of Θ alone (see the examples given above). Whenever this conditional factorisation (10) holds, one has

$$R = \sum_{k=1}^{p} I_k[\Theta; d^k] \; - \; I[\Theta; D] \tag{11}$$

where $I_k[\Theta; d^k]$ is the mutual information between the kth output unit alone and the input Θ. Since $R \geq 0$, one has then

$$I[\Theta; D] \leq \sum_{k=1}^{p} I_k[\Theta; d^k]. \tag{12}$$

Remark: conversely (11) implies (10) a.s.; it is not difficult to construct examples where (10) does not hold, and with the r.h.s. of (11) strictly negative.

As the above relations (11), (12) suggest, maximisation of the mutual information (infomax) leads to redundancy reduction, hence to ICA (Nadal and Parga 1994b, Nadal et al 1998). Sufficient conditions for this result to hold are: vanishing input noise, and infomax done over the choice of both synaptic couplings and transfer functions. In the particular case where the input is a linear mixture of IC's, and the network a simple nonlinear feedforward network (no hidden units), the infomax criterion is also equivalent to the maximum likelihood approach to ICA (Cardoso 1997), and specific algorithms based on these information theoretic cost functions have been proposed (see e.g. Pham et al 1992, Comon 1994, Bell and Sejnowski 1995).

Remark: In all this paper I am considering memoryless channels: the output D at some time t is a function of a single input Θ. Generalisation to systems with memory, that is where $D(t)$ depends on $\{\Theta(t-\tau), 0 \leq \tau \leq T\}$ for some possibly infinite T, can be done - and in some cases has been done, see e.g. Li 1995 (case of early visual system), Pham 1996 (case of ICA). In particular the relevant mutual informations to be considered in such cases are then $I[D(t); \{\Theta(t-\tau), 0 \leq \tau < T\}]$ for finite T and the limit when $T \to \infty$.

4 Typical behaviour of the mutual information

The typical behaviour of the amount of information conveyed by the (neural) system about the stimulus/parameter, as function of the number of coding cells (the number of observations) p, is shown on Figure 2, for the case of a real valued N-dimensional parameter. One can show (Herschkowitz and Nadal 1999) that for any p the mutual information is upper bounded by a quantity linear in p: this is simply because, at best, every datum conveys some information statistically independent of what is conveyed by the others, in which case the information will be essentially proportional to p. However, it is not always possible to saturate this linear behaviour, and in general for large p the mutual information has a logarithmic behaviour in p. I discuss now in more details these two regimes of particular interest: the small p ($p < N$), and large p ($p >> N$) regimes.

4.1 Small p regime

At small values of p (typically $p < N$, or more generally $p \leq p_c \sim d_{VC}$ where d_{VC} is the VC dimension), the maximal amount of information may saturate the upper

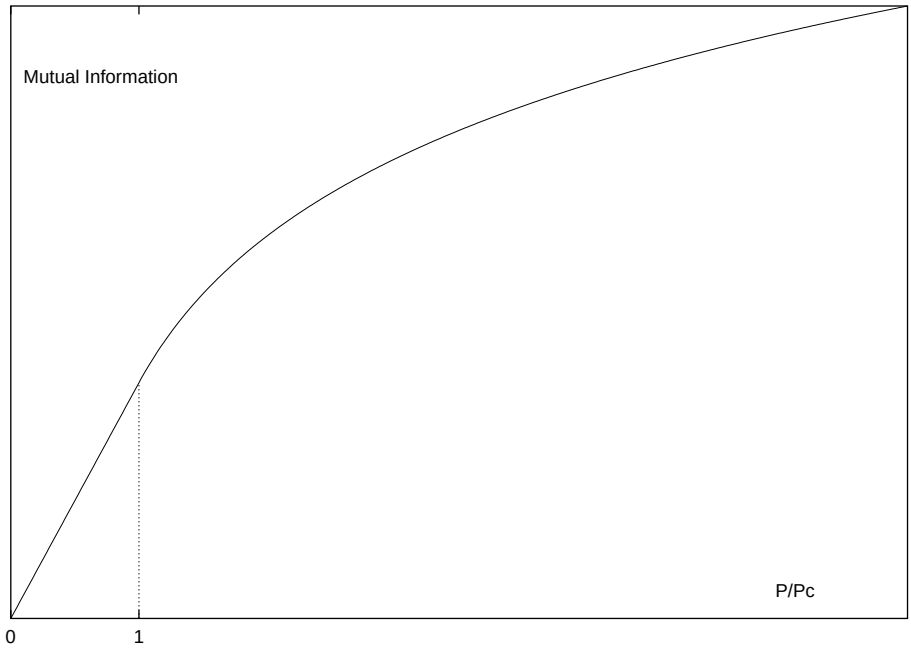

Figure 2: Typical behaviour of the mutual information I between input and output, as a function of the number of coding cells P scaled to the critical value, P_c. For $P \leq P_c$, I is (at best) proportional to P. For $P >> P_c$, I has a logarithmic behaviour.

bound, linear in p. In such a regime, maximal redundancy is achieved or, equivalently, Independent Component Analysis (ICA) is performed by the system.

In the context of learning/parameter estimation, such a linear regime may be observed and is related to what has been called *retarded classification*: in this regime, no estimation of the parameter can be made. This may appear paradoxical: on one side the largest possible amount of information is extracted from the data, and on the other side it is not possible to make any prediction concerning the next observation. These two aspects are in fact intimately related, and we will come back later to this point. Retarded classification has been exhibited in models with particular symmetries, in the asymptotic limit $N \to \infty$ with $\alpha = \frac{p}{N}$ kept fixed (Biehl and Mietzner 1993, Watkin and Nadal 1994). A rigorous proof of the existence of this regime has been derived recently (Herschkowitz and Opper, 1999).

4.2 Large p regime

The other important regime is the asymptotic one, that is for $p \to \infty$, more precisely for $p >> N$ when the VC dimension d_{VC} is finite. This large p limit is the one of interest for population coding, over-complete representations, in particular sparse codes, and it is the standard asymptotic limit considered in statistics (the limit of a large number of available data).

In this regime, the mutual information grows logarithmicly in p, more precisely as

$$\lim_{p>>N} I[\Theta; D] \sim K\ N \ln \frac{p}{N} \tag{13}$$

where the prefactor K depends on the smoothness of the statistical relationship between observations (activities of neural cells) and parameter (stimulus). This logarithmic behaviour is very related to the one obtained for the growth function discussed in Vapnik's framework. Indeed as mentioned in the previous section, in the case of a binary classification by a perceptron, the growth function is an upper bound on the mutual information. One should note however that in (13) N is the dimension of the parameter (more exactly the number of independent degrees of freedom), and *not* the VC dimension of the system.

For a smooth enough distribution the leading behaviour is always $\frac{N}{2} \log \frac{p}{N}$. In such a smooth case, one has in fact a simple exact expression for the mutual information in the asymptotic regime: it is found that it is given in terms of the *Fisher information*, that is (restricting here for simplicity to the scalar case, $N = 1$)

$$I[D; \Theta] = -\int d\theta \rho(\theta) \ln \rho(\theta) + \int d\theta \rho(\theta) \frac{1}{2} \ln \left(\frac{F(\theta)}{2\pi e} \right) \tag{14}$$

where the Fisher information $F(\Theta)$ is defined by:

$$F(\Theta) = \left\langle -\frac{\partial^2 \ln P(D|\Theta)}{\partial\ \Theta^2} \right\rangle_\Theta \tag{15}$$

in which $\langle\ .\ \rangle_\Theta$ denotes the integration over D *given* Θ with the p.d.f. $P(D|\Theta)$. Fisher information is related through the Cramer-Rao bound to the smallest possible variance of an estimator (Blahut 1988): the largest the Fisher information the smallest the variance.

It is quite remarkable that the existence of a direct relationship between a Shannon information and the Fisher information has been known for less than ten years. Indeed, this relation (14) (and its generalisation to arbitrary N) was first derived by Clarke and Barron, 1990, in the case of i.i.d. data; with weaker hypothesis it appears as a side result in the 1996 Rissanen's paper on the Minimum Description Length (MDL) principle; and a direct derivation in the context of neural coding and information processing (including cases of correlated data) is given in Brunel and Nadal 1998.

One should insist that it is only in this asymptotic limit that one has a direct relationship between number of bits and quadratic errors. In particular, for $p \sim N$ and/or with spiking neurons optimisation of the neural code at short times will be different whether the goal is optimal reconstruction or information preservation (Ruderman 1994, Brunel and Nadal 1998).

In the case of non smooth distribution - that is when the Fisher information is infinite -, there is no known general expression for the asymptotic behaviour of the mutual information. However recent results suggest (Haussler and Opper 1997) that the prefactor - the numerical constant K in (13) - can take only a small number of possible values. Specific models have been studied (Nadal and Parga 1993, Opper and

Kinzel 1995) and general bounds derived (Haussler and Opper 1997, Herschkowitz and Nadal 1999), showing that one always has a log behaviour with a rational prefactor. In the particular case of a binary classification by a perceptron, the prefactor is 1 (indeed $\ln \Delta(N,p) \sim N \log \frac{p}{N}$ for large p). This means that when the probability distribution of the data D given the parameter Θ has a discontinuity, the mutual information is asymptotically twice as large than for a smooth distribution.

The optimal performance in the estimation task, in the asymptotic limit of large p, is directly related to the behaviour of the mutual information: the typical discrepancy between the (best possible) estimator and the true parameter will go to zero as an inverse power law in p, with an exponent which depends on the prefactor K (for more details see Opper and Kinzel 1995, Haussler and Opper 1997, Herschkowitz and Nadal 1999). One interesting qualitative aspect is the fact that good performance are obtained in the regime where the mutual information has the slowest (log) growth, in contrast with the poorest performance obtained in the linear regime. Indeed, if after learning p examples one has a good prediction for the next one, d^{p+1}, this means that when the new datum d^{p+1} is actually given, one is only partly "surprised" by its value - precisely because one had already a good idea of what it would be. Hence the new information (the part not already contained in the first p examples) conveyed by this new datum is small: the information growth has to be sublinear. Conversely, a linear growth means that d^{p+1} conveys some totally new information, hence that no prediction at all was possible based on the first p examples.

In the context of neural coding, the asymptotic (log) regime gives a highly redundant code: indeed, the fact that it is possible to have a good prediction of the activity d^{p+1} of the $p+1$th cell, given the activities of the first p cells, means that the amount of information conveyed by the $p+1$th cell is not independent of the information conveyed by the other cells. As we have seen this redundancy cannot be avoided: for a given number N of inputs (sensory receptors), the marginal gain of information obtained by adding a new coding cell becomes smaller and smaller as p increases. However such redundancy may be usefull for various reasons - e.g. noise robustness, or in order to satisfy other constraints such as sparseness, see chapter by Olshausen.

Finally one should add that the log behaviour is obtained whenever one has a real valued parameter belonging to a finite dimensional space. For a parameter with discrete components, the mutual information is upper bounded by the finite entropy of the parameter; this bound is saturated for large p, either at a finite value of p or asymptotically with an exponential convergence rate (Haussler and Opper 1995). When the parameter is a function (which means in particular $N = \infty$), one may find a power law behaviour (Opper, 1999) with an exponent smaller than one - in agreement with the fact that the growth must be sublinear.

5 Beyond sensory coding

I have presented some common and generic properties of information processing in the dual context of neural coding and parameter estimation. To conclude I would like

to make some comments on possible further developments, and on the limitations, of this approach illustrated on Figure 1. I will propose/evoke several possible lines of research which still make use of the information theoretic framework trying to address computing aspects not necessarily obviously related to efficient coding.

First one should say that information theoretic criteria may not be, or not always be, the appropriate ones: as put forward by Z. Li (Li 1998), one may have to contrast optimal coding versus computation. An example of this is already presented in the previous section: for p small optimal coding (via infomax, that is maximisation of information preservation) is not equivalent to optimal reconstruction (minimisation of quadratic error); however, for large p the two criteria become identical. More generally one can say, as we have seen, that the mutual information is the relevant quantity for parameter estimation - if the cost is taken as an entropic loss, and not the quadratic error.

As already said for neural coding, it is not always the case that one has to compute an estimate of Θ. It might be also that one is interested in some function $F(\Theta)$, so that the final output is not $\hat{\Theta}$ but the estimate $\widehat{F(\Theta)}$. If the desired computation task is known, that is $F(.)$ is known, then one is interested in $I[D;F]$ instead of $I[D;\Theta]$. One can note that the input Θ can be considered as a "noisy" representation of $F(\Theta)$ if $\Theta \rightarrow F(\Theta)$ is not invertible.

A different but similar approach would be the one followed in Communication Theory (see e.g. Blahut 1988), when a cost for decoding errors is prescribed in advance. For instance suppose the task is to minimise the quadratic error, $E =< (\hat{\Theta} - \Theta)^2 >$. Then one may ask for the *minimisation* of the mutual information with the constraint that E is smaller than some prescribed level: one is then building a code which conveys just the information directly needed for this particular task.

An other line of research is one followed by Phillips (Phillips, 1995, Phillips and Singer, 1997): one considers the N inputs to be composed of the external stimulus and of what defines the task or context (other external stimuli and/or internal states of the system); an *infomax* cost function is then built which distinguishes the two types of inputs.

An other important aspect is the search for codes preserving topological or geometrical properties (Victorri and Derome 1984, Rao 1998). It is not obvious whether the building of a code showing invariance through particular transformations should be the result of ad-hoc constraints imposed in addition to a coding criterion (see e.g. the approach in Li and Atick, 1994b, where translational and scale invariance is a constraint added to redundancy reduction), or whether such properties must appear as a consequence of efficient coding. I would expect the second hypothesis to be correct, as suggested by works on natural image analysis (Field 1987, Ruderman and Bialek 1994, Turiel and Parga 1999, this book chapter by Piepenbrock and Obermayer).

One should note also that the framework illustrated on Figure 1 applies as well to time dependent activities. Particular cases are very briefly mentionned in this chapter. To give an example, one can consider a single spiking cell responding to a particular stimulus. In our framework p is then the time during which one collects the number of spikes emitted by the cell, and the behaviour of the mutual information between

the cell activity and the stimulus, as a function of time (of p) is the one discussed in this chapter - linear at short times, logarithmic at long times - (Stein 1967, Bialek et al 1991, Brunel and Nadal 1998). The information content of more complex time dependent codes have yet to be studied (for various approaches to the modelling of spiking neural cells, see this book, section on "Neural Models and Implementations").

A last remark is that the framework illustrated on Figure 1 seems to imply a purely feedforward system, where each output d^k is a function of the input Θ. This is not the case. Indeed, our basic tool is the mutual information, which is a symmetric quantity with respect to the two random variables Θ and D (it does not distinguish an "input" from an "output"): it is well defined whenever one can write the joint probability distribution $P(\Theta, D)$. Hence most of what has been said above applies as well for motor coding (D being the internal motor representation and Θ the motor action), and population coding is indeed studied for both sensory and motor representations. The framework applies also whatever the complexity of the network, including with correlated outputs: in particular D might be the activities of an attractor network whose initial state has been, directly or indirectly, imposed by the stimulus Θ. One should note also that mutual information is a relevant tool for characterising the neural code, even when optimal information preservation is not the chosen criterion: for example it would be interesting to study the information content of sparse codes.

The comments in the previous paragraph leads to one of the most interesting problem for which the general framework, Figure 1, may not be fully appropriate, that is the building of neural codes *via* the sensori-motor loop. That is, one would like the system to build its internal representations through *active perception*, which implies that the source distribution, ρ, is not externally given by the environment: the sequence of successive inputs depends on the past history - the past sequence of stimuli, actions. I would suggest that one can stay within the same framework, using the following point of view: a joint sensori-motor representation may be built by considering the mutual information between the neural code D and the pair {perception(t),action(t+1)}; the decision to make a particular action can be considered as trying to choose a particular stimulus, the one which would be the most efficient in the learning process leading to the neural code. This is then the same as *learning by queries* in statistical learning theory (Kinzel and Rujan 1990), a learning strategy in which at each step one picks the training example which provides the largest possible amount of information. The result is then a learning (adaptation) process which can be efficient in terms of convergence rate towards the optimal solution - but it does not affect the nature of the optimal solution (the optimal code).

Acknowledgements

This contribution is based on works done mainly with Nestor Parga, Nicolas Brunel and Didier Herschkowitz. I thank Zhaoping Li, Alexandre Pouget, Sophie Denève, Anthony Bell and Manfred Opper for stimulating discussions during the NIPS meeting. This work has been partly supported by the French grant DGA 96 2557A/DSP.

References

[A92] Atick J. J., Could information theory provide an ecological theory of sensory processing. *NETWORK*, 3:213–251, 1992.

[A54] Attneave F., Informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954.

[B60] 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.

[BKM89] Barlow H. B., Kaushal T. P., and Mitchison G. J., Finding minimum entropy codes. *Neural Comp.*, 1:412–423, 1989.

[BS95] Bell A. and Sejnowski T., An information-maximisation approach to blind separation and blind deconvolution *Neural Computation* 7:1129–1159, 1995.

[Betal91] Bialek W., Rieke F., de Ruyter, van Steveninck R., and Warland D., Reading a neural code. *Science*, 252:1854–57, 1991.

[BM93] Biehl M. and Mietzner A., Statistical Mechanics of Unsupervised Learning, *Europhys. Lett.* 24:421, 1993;

[B88] Blahut R. E., *Principles and Practice of Information Theory*, Addison-Wesley, Cambridge MA, 1988.

[BG98] Buhot A. and Gordon M., Phase transitions in optimal unsupervised learning. *Phys. Rev. E*, 57(3):3326–3333, 1998.

[BN98] Brunel N. and Nadal J.-P., Mutual information, Fisher information and population coding. *Neural Computation*, to appear, 1998.

[C97] Cardoso J.-F., Infomax and maximum likelihood for blind separation *IEEE Signal Processing Letters* 4:112-114, 1997.

[CB90] Clarke B. S. and Barron A. R., Information-theoretic asymptotics of bayes methods. *IEEE Trans. on Information Theory*, 36 (3):453–471, 1990.

[C94] Comon P., Independent component analysis, a new concept ? *Signal Processing*, 36:287–314, 1994.

[CT91] Cover, T. M. and Thomas, J. A., *Information Theory*, John Wiley, 1991.

[F87] Field D., Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am.*, 4:2379, 1987.

[vH92] van Hateren J.H., Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation. *J. Comp. Physiology A*, 171:157-170, 1992.

[HO95] Haussler D. and Opper M., General bounds on the mutual information between a parameter and n conditionally independent observations. In *VIIIth Ann. Workshop on Computational Learning Theory (COLT95)*, pages 402–411, Santa Cruz, 1995 (ACM, New-York).

[HN99] Herschkowitz D. and Nadal J.-P., Unsupervised and supervised learning: the mutual information between parameters and observations *Phys. Rev. E*, 59(3):3344-3360, 1999

[HO99] Herschkowitz D. and Opper M., Retarded Learning in High Dimensional Data Spaces *in preparation*, 1999.

[JH91] Jutten C. and Herault J., Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture, *Signal Proc.*, 24:1–10, 1991.

[KR90] Kinzel W. and Rujan P., Learning by queries, *Europhys. Lett.* 13:473, 1990.

[L81] Laughlin S. B., A simple coding procedure enhances a neuron's information capacity. *Z. Naturf.*, C 36:910–2, 1981.

[Li95] Li Z., A theory of the visual motion coding in the primary visual cortex *Neural Computation* 8(4):705-30, 1995.

[Li98] Li Z., *This Workshop*, and: A neural model of contour integration in the primary visual cortex *Neural Computation* 10:903-940, 1998

[LA94a] Li Z. and Atick J. J., Towards a theory of the striate cortex. *Neural Comp.*, 6:127–146, 1994.

[LA94b] Li Z. and Atick J. J., Efficient stereo coding in the multiscale representation. *Network: Computation in Neural Systems*, 5(2):157-174, 1994.

[L88] Linsker R., Self-organization in a perceptual network. *Computer*, 21:105–17, 1988.

[L93] 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.

[NBP98] Nadal J.-P., Brunel N. and Parga N., Nonlinear feedforward networks with stochastic ouputs: infomax implies redundancy reduction *Network: Computation in Neural Systems* 9(2):207-217, 1998.

[NP93] Nadal J.-P. and Parga N., Information processing by a perceptron in an unsupervised learning task. *NETWORK*, 4:295–312, 1993.

[NP94a] Nadal J.-P. and Parga N., Duality between learning machines: a bridge between supervised and unsupervised learning. *Neural Computation*, 6:489–506, 1994.

[NP94b] Nadal J.-P. and Parga N., Nonlinear neurons in the low noise limit: a factorial code maximizes information transfer *Network: Computation in Neural Systems* 5:565-581, 1994.

[OF97] Olshausen B. A. and Field D. J., Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311-3325, 1997.

[OH95] Opper M. and Haussler D., Bounds for predictive errors in the statistical mechanics of supervised learning, *Phys. Rev. Lett.*, 75:3772-3775, 1995.

[OK95] Opper M. and Kinzel W., Statistical mechanics of generalization. In E. Domany J.L. van Hemmen and K. Schulten, editors, *Physics of Neural Networks*, pages 151–. Springer, 1995.

[P96] Pham D.-T., Blind Separation of Instantaneous Mixture of Sources via an Independent Component Analysis *IEEE Trans. SP* 44(11):2768–2779, 1996.

[PGJ92] Pham D.-T., Garrat Ph. and Jutten Ch., Separation of a mixture of independent sources through a maximum likelihood approach in *Proc. EUSIPCO*, pp 771–774, 1992.

[P95] Phillips W. A., Kay J. and Smyth D., The discovery of structure by multi-stream networks of local processors with contextual guidance *Network: Computation in neural systems* 6:225-246, 1995.

[PS97] Phillips W. A. and Singer W., In search of common foundations for cortical computation *Behavorial and Brain Sciences* 20(4):657-722, 1997.

[R98] Rao R. P. N. and Ruderman D. L., Learning Lie Groups for Invariant Visual Perception, In *Advances in Neural Information Processing Systems 11* (proceedings of NIPS98), Ed. by M. S. Kearns, S. A. Solla and D. A. Cohen (The MIT Press 1999), p. 810-816.

[R93] Redlich A. N., Redundancy reduction as a strategy for unsupervised learning. *Neural Comp.*, 5:289–304, 1993.

[RVdB96] Reimann P. and Van den Broeck C., Learning by examples from a non uniform distribution. *Phys. Rev. E*, 53 (4):3989–3998, 1996.

[Ris] Rissanen J., Fisher information and stochastic complexity. *IEEE Trans. on Information Theory*, 42 (1):40-47, 1996.

[R94] Ruderman D., Designing receptive fields for highest fidelity. *NETWORK*, 5:147–155, 1994.

[RB94] 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.

[SW49] Shannon C. E. and Weaver W., *The Mathematical Theory of Communication.* The University of Illinois Press, Urbana, 1949.

[S67] Stein R., The information capacity of nerve cells using a frequency code *Biophys. Journal* **7** (1967) 797–826.

[SK98] Stemmler M. and Koch Ch., Information maximization in single neurons In *Advances in Neural Information Processing Systems 11*, Ed. by M. S. Kearns, S. A. Solla and D. A. Cohen (The MIT Press 1999), p. 160.

[SS93] Seung H. S. and Sompolinsky H., Simple models for reading neural population codes *P.N.A.S. USA* 90:10749–10753, 1993.

[TP99] Turiel A. and Parga N., The multi-fractal structure of contrast changes in natural images: from sharp edges to textures, To appear in *Neural Computation.*

[VdB98] Van den Broeck C., Unsupervised learning by examples: on-line versus off-line learning. In *proceedings of the TANC workshop* (Hong-Kong May 26-28, 1997).

[VD84] Victorri B. and Derome J.-R., Mathematical model of visual perception, *J. Theor. Biol.* 108:227-260 (1984).

[WN94] Watkin T. and Nadal J.-P., Optimal unsupervised learning. *J. Phys. A: Math. and Gen.*, 27:1899–1915, 1994.