ISAS - INTERNATIONAL SCHOOL
FOR ADVANCED STUDIES

Quantitative methods for analyzing information processing in the mammalian cortex

Thesis submitted for the degree of
"Doctor Philosophiæ"

CANDIDATE

Stefano Panzeri

SUPERVISOR

Alessandro Treves

December 1996

SISSA - SCUOLA
INTERNAZIONALE
SUPERIORE
DI STUDI AVANZATI
TRIESTE
Strada Costiera 11

TRIESTE

Contents

Chapter 1

Introduction

1.1 Motivation

The relevance of information theory for neuroscience ultimately derives from the fact that the nervous system possesses a lot of subsystems that acquire, process and transmit information. Therefore many brain structures can be considered as communication channels, and the more appropriate mathematical framework for the *quantitative* characterization of their performance is information theory (Shannon 1948). In particular, quantities like Shannon's mutual information and channel capacity provide a powerful tool to quantify how neurons represent messages from the external world, and how efficiently they do it.

Since it is well known that the signals from the external world are converted by the nerve cells, already at very early stages in sensory processing, into sequences of spikes, one may wonder why quantifications of how information is represented by the firing of single neurons and of populations of neurons, whose activity is recorded *in vivo*, have been for a long time performed only episodically (*e.g.,* Eckhorn and Pöpel 1975; Eckhorn *et al.* 1976). The main reason for this has certainly been the large amount of data that was required in order to obtain reliable results. A calculation of the Shannon mutual information requires in fact the evaluation of the probabilities of response of a neuron to *e.g.,* a set of external stimuli, and limited sampling produces fluctuations in the estimated probability densities. As a consequence, the resulting estimations of information from a finite number of examples are intrinsically affected by an upward systematic error, occasionally as large as the target quantity itself. That the problem exists can be intuitively understood if one thinks that, even if the response probabilities of the neuron do not depend on the particular set of stimuli, and thus the cell does not carry information about that set, finite sampling would tend to produce differences in the responses to various stimuli, differences that are interpreted in the calculation as containing genuine information. The situation is further complicated by the fact that, if the cell response is quantified by a continuous variable, the limited amount of data requires also a regularization of the neuronal output, *e.g.,* grouping responses into a finite number of classes or bins. One can think of avoiding the finite data size effects by using a regularization strong enough to dump the statistical fluctuations. But in this case the strong regularization reduces the discriminational capacity of the response, leading to a systematic downward bias in the estimation of the information. Given that in practice the size of data sets is limited by experimental constraints, it is evident that one can make effective use of neuronal mutual information measurements only if one is able to control these systematic errors with opposite effects.

Although many empirical procedures to correct information measures for limited sampling have been introduced in the last few years (Optican *et al.* 1991, Chee-Orts and Optican 1993, Hertz *et al.* 1992), the results were not fully satisfactory, as witnessed by the fact that a number of paper appeared, reporting opposite results (see *e.g.,* the discussion in Tovée *et al.* 1993). To shed more light on the problem, we have developed instead a more fundamental approach to the limited sampling problem, that we present as the main result contained in the dissertation, based on a direct evaluation and subtraction of the limited sampling bias (Treves and Panzeri 1995; Panzeri and Treves 1996b). The idea, which has been conceived as early as 40 years ago (Miller 1955), has been developed to be applied to neuroscience, in particular adapted to the regularization procedures that can be used with neuronal data. In our approach, of crucial practical importance is the fact that the subtraction of the finite sampling upward bias allows the choice of milder data manipulations, ultimately decreasing also the downward bias related to response regularization. This lowers the size of the sample required for a given accuracy in the estimate by about an order of magnitude. However an analytical evaluation of the bias does not, in itself, make possible reliable measures of information carried by high dimensional codes with a few trials. A part of the thesis is thus devoted to study, mainly with computer simulations, the range of applicability of the correction and the effectiveness of various regularization techniques in extracting accurate information measures. The statistical method that we present in this thesis is mainly designed for the study of information processing in the mammalian cortex, were the data collection is really a serious constraint, and is less relevant when working with insects and considering systems at the sensory periphery (Bialek *et al.* 1991; Bialek *et al.* 1993; de Ruyter van Steveninck and Laughlin 1996), in which sampling can be extensive. Nevertheless, even in the latter case the use of a procedure based on the same idea as ours can lead to more accurate and assumption free information estimates (Strong *et al.* 1996).

The lack of a systematic study of the information carried by real neurons has infuenced also the theoretical work in neuroscience. In fact, the use of information theory in modelling brain functions has been mainly limited to the sensory periphery, and especially to the early visual system (retina in Atick and Redlich 1990; LGN in Dong and Atick 1995; V1 in Olshausen and Field 1996). In this case, the information theoretical analysis has been not applied to a model of the detailed neuronal circuitry, but the observed properties of receptive fields of the cells have been directly compared to that predicted by some first principle of optimization of information transmission. Instead, when detailed models of neuronal circuitries have been introduced, simpler quantifications of the performance of the system, without a clear experimental counterpart (*e.g.,* in autoassociative memory models, the overlap of a retrieved pattern with the stored one), have been often used.

The last part of the thesis is devoted to address this point, discussing the analytical techniques (essentially derived from statistical mechanics) needed to extract, as a basic index of performance of the network, the amount of information present in the activity of the (model) neurons. For this purpose, we choose to study a biologically plausible model of a part of the hippocampal complex, mainly derived from the previous work of Treves and Rolls (1992; 1994). We show how to define the probability of response of model neurons in terms of quantities with an experimental correlate, and how to extract from it the information carried by the neural network.

The work presented in this thesis, and discussed here, is rather technical. But we want to stress that it is the developement of these techniques that eventually makes comparable, in their full quantitative import, experimental results and analytical modelling results, under the unifying framework of information theory.

1.2 Organization of the thesis

In chapter 2, we study the finite sampling problem in measures of information, discussing previously introduced correction procedures, and describing our own method, based on the analytical evaluation of the average error and its subtraction form raw estimates. The analytical evaluation is carried out for different regularizations of the responses, like pure binning, convolutions with continuous distributions and regularization with neural networks. The last sections of the chapter report the results of computer simulations, which shed light on the relative effectiveness and on the range of validity of our and of other methods.

Chapter 3 is devoted to discuss how to apply information theoretical analyses to neuronal data. In particular two different aspects are considered. The first is the evaluation of the amount of information contained in the firing rates of single cells. We will discuss how to use the high precision of information measures from low dimensional codes to study the contribution of different properties of the neuronal responses, such as noise or the graded nature of responses, to information processing at different time scales. Furthermore, we discuss the important fact that the initial rate at which a neuron transmits information depends only on the mean firing rates (Skaggs *et al.* 1993), and is simply related to the sparseness of the neuronal representation. The second problem considered is the calculation of information from high dimensional response spaces, as the principal components of the spike train, or the response vector of a population of simultaneously recorded neurons. It is shown that the limit of dimensions that can lead to reasonably accurate direct measures is low, 2-3, whereas changes of variables that transforms the response space into the stimulus set, by applying a decoding algorithm that reconstructs a predicted stimulus from the response vector, can give sensible results for higher dimensional codes. This discussion on single cell and multiple cell analysis is corroborated by computer simulations, and by the detailed presentation of original results on the analysis of real data: coding of spatial view by cells in the primate hippocampus, and coding of simple somatosensory stimulations in the rat SI cortical region.

In chapter 4 we present a quantitative model of information processing within hippocampus. We take into account the entorhinal-CA3-CA1 system, focusing on the role of the Schaffer collaterals and the direct perforant path connections from entorhinal cortex to CA1. The model is quantitative in that the relevant details of the biological circuitry are taken into account, and the parameters characterizing the network can be related to experimental quantities. The goal of the chapter is to provide an analytical evaluation of the amount of information present, at the firing rate level, in the model CA1 output about the firing activity of the cells in entorhinal cortex. For this purpose we use standard techniques of statistical mechanics, like mean field theory and the replica trick. We discuss also how to use experimental data to validate some of the building hypotheses of the model, and to check its quantitative predictions.

Chapter 2

The limited sampling problem in measures of information.

Quantifying the relation between neuronal responses and the events that have elicited them is important for understanding the brain. One way to do this in sensory systems is to treat a neuron as a communication channel (Shannon 1948; Cover and Thomas 1991) and to measure the information conveyed by the neuronal response about a set of stimuli presented to the animal. In such experiments (*e.g.,* Gawne and Richmond 1993; McClurkin *et al.* 1991; Tovée *et al.* 1993) a set of S sensory (*e.g.,* visual) stimuli is presented to the animal, each stimulus being presented for N_s trials. After the neuronal response is quantified in one of several ways, e.g., the number of spikes in a certain time interval or a descriptor of the temporal course of the spike train, the transmitted information (mutual information between stimuli and responses) is estimated. This approach is useful for investigating issues such as the resolution of spike timing (Heller et al. 1995), the effectiveness of encoding for stimulus sets (Optican and Richmond 1987; McClurkin *et al.* 1991; Rolls *et al.* 1996c,d), or the relations between responses of different neurons (Gawne and Richmond, 1993). Nevertheless, extracting information from real neurons, whose activity is recorded *in vivo*, is so ridden with subtleties that in practice important questions such as the type of neural coding used by different systems in the mammalian brain, or the speed of information processing, have been most easily approached qualitatively from a theoretical point of view, rather than quantitatively from experimental observations. The main problem in quantifying information carried by neuronal spike trains is the limited sampling: calculation of mutual information from limited samples is affected by a systematic upward bias. The bias can be, if the trials available are few, much larger than the true information values themselves. An intuitive explanation for this is that fluctuations due to finite sampling tend, on average, to emphasize differences among distributions corresponding to different stimuli, differences that can be interpreted in the calculation as carrying genuine information.

To discuss how the limited sampling distorts measures of information, and how to correct for this systematic error, let us consider a concrete situation in which we wish to measure the amount of information, in bits, that some variable r, associated with the response of one or more neurons, conveys about a stimulus s, presented to the animal. We take s to belong to the discrete [1] set $\mathcal{S}$ of S elements. We wish to measure both the (average) conditional information transmitted when s is presented,

$$I(s) = \int drP(r|s)\log_2 \frac{p(s|r)}{p(s)} = \int drP(r|s)\log_2 \frac{P(r|s)}{P(r)} \tag{2.1}$$

and its average across stimuli, *i.e.*, the mutual information

$$I = \sum_{s\in\mathcal{S}} p(s) \int drP(r|s)\log_2 \frac{P(r|s)}{P(r)} \ . \tag{2.2}$$

We assume that only N stimulus-response pairs (s,r) are available, instead of the full probabilities $p(s), P(r)$ and $P(s,r)$ (the last two are, in general, probability densities rather than probabilities, and are thus denoted with capital letters). For $N \to \infty$, individual (s,r) pairs are expected to occur with frequencies tending to match the underlying probabilities, but for N finite, use of the experimental frequencies $p_N(s), P_N(r)$ and $P_N(s,r)$ directly in the formulae above leads to systematic error. That the problem exists, can be seen by considering uncorrelated stimuli and responses, such that $P(s,r) = p(s)P(r)$: a finite-N evaluation of the mutual information, which is zero by definition, will almost certainly yield a positive result, which therefore indicates a systematic error. The problem, neglected in the early literature on neuronal information processing (Optican and Richmond 1987), has been studied during the last years by several authors.

The first procedure, introduced in the context of neuronal data analysis, to correct for the error, was suggested by Optican *et al.* (1991). It follows from considering the case of uncorrelated pairs: it involves generating a *shuffled* probability distribution by randomly pairing stimuli and real responses, calculating the *shuffled information* contained in the real responses about the randomly paired pseudostimuli and finally subtracting a fraction of the shuffled information from the raw value of measured information. This random shuffling procedure, often called *bootstrap* because it uses the data to correct the data themselves, is flawed in several ways. First, and most evidently when responses are discrete, the shuffled information may be a strong *overestimation* of the bias, for reasons to be clarified below, and then it is wrong to subtract from the raw estimate the correction derived from random shuffling. Furthermore, the shuffling procedure is applicable only to measures of mutual information (2.2) and not to measures of conditional information

[1] Here we specialize to the case of discrete sets of stimuli, which is mostly relevant to the case of experiments in mammalian memory and higher sensory areas. There is, however, a lot of work based on invertebrates (Bialek 1991; Bialek *et al.* 1991; de Ruyter van Steveninck and Laughlin 1996; Rieke *et al.* 1993; Theunissen *et al.* 1996), where experimental data is obtained by presenting a stimulus which is a continuous quantity. This opens up the possibility of using notions such as the linearized limit, the Gaussian approximation, etc., that do not apply to our situation with a discrete nonmetric set of stimuli, and which when applicable can further alleviate finite sampling effects.

(2.1), since the shuffling mixes responses occurring to different stimuli. Finally, when the responses are regularized before being used to measure information, the regularization can affect the raw and shuffled measures to different degrees, as will again be clear below.

More sophisticated is the procedure suggested by Hertz *et al.* (1992), based on strong regularization of the input-output distribution by means of a neural network used to estimate the probability of each input s given the output r. The neural network is trained so as to maximize the probability that a stimulus is correctly recognized, *i.e.*, that the stimulus estimated to be most probable is the actual one. This method appears to have small finite size effects (Kjaer *et al.* 1994). What appears to be unsatisfactory in the network regularization is that, while *any* regularization results in information loss and information measures relative to *that* regularization, the regularization produced by the artificial network is particularly complex and data dependent and it is hard to assess the relation between the target information and the regularized measure one obtains. This is made evident in the paradoxical result of Kjaer *et al.* (1994) where occasionally codes that are by definition more rich in information (retaining more principal components of the responses) appear to carry less information, after they have been squeezed through the artificial network.

The limited sampling error is a statistical problem common to many different fields, whenever one tries to estimate, from a finite sample, a function of a full probability distribution. Several authors have addressed it, outside the domain and the peculiarities of computational neuroscience, *e.g.,* focusing on probabilities given on discrete sets. Wolpert and Wolf (1995) (see also references therein) propose the calculation of the function (in our case, *e.g.,* , I) of the true probabilities *given* the experimental frequencies. This, which is in fact the original aim (and which is obviously different from calculating the function of the frequencies, our I_N) is feasible, however, only by making an assumption as to the *a priori* probability distribution. It is then difficult to see how to use this conceptual appealing approach in cases, such as ours of stimulus-response pairs, when no reasonable assumption on the prior is self-evident.

We have developed an alternative, non parametric, approach, based on the analytical calculation of the *average error* as an asymptotic expansion in inverse powers of the sample size. The calculation has been first performed by assuming the response space be discrete, or at least discretized (Treves and Panzeri 1995), and then extended (Panzeri and Treves 1996b) to more general regularization procedures of neuronal responses, such as convolution with continuous distributions, or neural network regularization, which can be particularly useful when multidimensional codes or multi-cell recording are concerned. We have found that the leading (in $1/N$) contribution to the bias, depending very smoothly on the underlying probabilities, and easily computable from the data, yields most of the error and can thus be subtracted to correct raw estimates. Successive terms of the expansion are of little use: either they are negligible in comparison to the first term, or when N becomes very small, they explode quickly (see also Strong *et al.* 1996), signalling that data are so scarce that the expansion is meaningless beyond the first term. We have also adapted our procedure not only to mutual information but also to conditional information (*i.e.,* relative to a given stimulus). Moreover, our direct evaluation of the bias allows a better understanding of the role of the shuffled information in correcting for limited samples, and a more appropriate choice of the data regularization for a given problem and data size.

An evaluation of the sampling error called *Jackknife* technique (Efron 1982) was introduced, in the context of non-parametric statistic. This technique allows unbiased estimates of generic functions of probabilities by means of recalculating the function after deleting different data subsets. This method, which was until recently (Theunissen *et al.* 1996) neglected in the neuronal data analysis concerned here, is based on the assumption that only the leading $1/N$ term of the bias is important, and allows to obtain unbiased estimates avoiding the problem of direct evaluation of the bias. The jackknife correction is however less worthwhile for information quantities, where the average error, easily computable from data, can be calculated using only very weak assumptions on the underlying probability distributions. Moreover, the jackknife has the disadvantage that it involves calculation and subtraction of quantities of order N. This latter fact leads both to computational problems (CPU time and numerical fluctuations) and, when compared to our method, to a slower convergence with N towards the unbiased result.

The aim of this chapter is to present and discuss the analytical correction for the average error that we have performed, and then to use analytical results, as well as computer simulations, to establish the range of validity and the relative effectiveness of the various methods described before. The organization of the chapter is as follows: in the first section we address the finite sampling problem by means of the calculation of the average error as an asymptotic expansion in inverse powers of the data sample size. Results are obtained for different regularizations of the responses often used with neural data. In the second section we report the outcome of a similar analysis for the case in which stimuli are not drawn at random from a multinomial probability distribution, but in contrast the experimental frequency of presentation of stimuli does not fluctuate. The third section is devoted to test, by means of computer simulations, the validity of analytical results. In section 4 we summarize the results found and finally, in the last section, we discuss how the analytical evaluation of the bias can help with the choice of the best (or most effective) regularization for a given problem and set of data.

2.1 The average error

In this section we present our evaluation of the bias, i.e. the average error, when different regularization procedures are applied to the raw data. We take the stimuli s to have been drawn at random (with a multinomial probability distribution) from a discrete set $\mathcal{S}$ of S elements. Note that when the experimental frequency of presentation of stimuli is, instead, set exactly equal to its probability and does not fluctuate, one finds slightly different correction terms, as will be discussed separately in section (2.2).

Let us initially consider the more general case in which the (raw) neuronal response is a real (possibly multidimensional) variable [2]. It is clear from the formula for the mutual information (2.2), that if one is measuring a continuous output variable, in order to obtain an estimate of the mutual information from a finite set of N data a regularization of the raw data is always necessary; otherwise the finite number of responses will almost certainly be all different from each other, therefore each response will uniquely identify its stimulus ($p_N(s|r)$ will be either 1 or 0) and, as a result one will obtain only a measure of the entropy of the stimulus set, and not of the mutual information. Moreover, the response space is usually quantized anyway, because one needs to evaluate the expressions for I and $I(s)$, in practice, by performing a sum rather than an integral. Furthermore, many authors, for several reasons, prefer to use manipulations of data different from a pure discretization of the response space.

In the following subsections, we shall consider four important cases of regularization: pure discretization; convolution with a continuous distribution and discretization; neural network fitting of the conditional probabilities; without discretization of response space.

We shall report in Appendix A the explicit calculations leading to our expression of the bias in the second case only; but, for the sake of generality, we shall also discuss how to retrieve the results presented when the other data manipulations are applied. More details of the procedure are reported in Panzeri and Treves (1996b).

2.1.1 Pure discretization of the response space

Let us consider in this section the case in which real responses have been binned into R different intervals [3] $[m_{j-1}, m_j], j = 1, \cdots, R$, by just assigning each response to the interval it falls in. In this case, the binning procedure satisfies an *independence* condition, i.e., the number of times a given bin is occupied depends only on the underlying probability of the given bin, and not on the occupancy of other bins (this condition is violated by the prior regularization of the responses, as in the cases to follow).

Within this binning procedure, from N experimental trials available, one can obtain a raw estimate of the information:

$$I_N^D(s) = \sum_{i \in \mathcal{R}} p_N(i|s) \log_2 \frac{p_N(i|s)}{p_N(i)}; \qquad I_N^D = \sum_{s \in \mathcal{S}} p_N(s) I_N^D(s) \; . \tag{2.3}$$

In (2.3) the p_N's are the experimental frequency-of-occupancy tables, *e.g.*, $p_N(i) = n(i)/N$, or $p_N(i|s) = n(i|s)/N_s$, where $n(i|s)$ is the number of times response i occurred when stimulus s was presented, $n(i)$ the number of times response i occurred across all stimuli, and N_s is the number of experimental presentations of stimulus s. For large N

[2] We write all the formulae in the manner appropriate to a one-dimensional response space, but the generalization to higher dimensions, as well to the case in which the original response is discrete (*e.g.*, the number of spikes in a given time window) is straightforward.

[3] We stress that R is the total number of response bins, independently of what is the underlying dimensionality, if any, of the raw response space. If e.g. the raw responses are the firing rates of two cells, which are then discretized into R_1 and, respectively, R_2 bins, we set $R = R_1 \times R_2$.

the experimental frequencies $p_N(i)$ tend to the corresponding probabilities $p(i)$, which are simply related to the original continuous underlying probability distribution by an integration over each response bin. Similarly, as N increase, the estimate of transmitted information tends to the information carried by the discretized probabilities:

$$I^D(s) = \sum_{i \in \mathcal{R}} p(i|s) \log_2 \frac{p(i|s)}{p(i)}; \qquad I^D = \sum_{s \in \mathcal{S}} p(s) I^D(s) \; . \tag{2.4}$$

By temporarily restricting ourselves to the total transmitted information, it is important to note that the value of the information obtained *after* quantization, or regularization, is less than the value of information carried by the continuous responses, and in general information measures are dependent on the binning procedure adopted, and most importantly on the number of bins R. There is no way to estimate the difference between the unregularized and regularized values of the mutual information from first principles, but a good strategy to control these discrepancies can be to quantize the responses by successively increasing the value of R until the finite N measure, *after* the correction we are discussing, does not change very much. However, when the size of the data sample is small, a reasonable choice for R is a compromise between trying to keep the loss of information due to discretization as small as possible, which would require R large, and the need to control the finite-size distortion, which, as we shall see below, can require R small.

Of course, the difference, or bias, between I_N^D and I^D fluctuates depending on the particular outcomes of the N trials performed. We can estimate the average of the difference, however, by averaging ($< \ldots >$) over all possible outcomes of the N trials, keeping the underlying probability distributions fixed. We have obtained an expression for the bias as a series expansion in inverse powers of the sample size N:

$$< I_N^D > - I^D = \sum_{m=1}^{\infty} C_m^D \tag{2.5}$$

where C_m represents successive contributions to the asymptotic expansion of the bias (the term C_m is proportional to N^{-m}; see the Appendix A for the details of the calculation). Here we report just the leading term, whose expression is:

$$C_1^D = \frac{1}{2N \log 2} \left\{ \left(\sum_s \tilde{R}_s \right) - \tilde{R} - S + 1 \right\}, \tag{2.6}$$

where $\tilde{R}_s$ denotes the number of "relevant" response bins for the trials with stimulus s, which are the response bins with non-zero probability to be occupied during the presentation of the stimulus S. In the same way, $\tilde{R}$ is the number of response bins with non-zero occupancy probability across all stimuli. In the case in which each response bin i has a non-zero probability of being occupied for every stimulus s, we recover the simpler expression reported in Treves and Panzeri (1995)

$$C_1^D = \frac{(S-1)(R-1)}{2N \log 2} \; . \tag{2.7}$$

At the end, to correct for the finite size problem we have to evaluate the correction term in Eq. (2.6), which depends on the underlying probabilities solely through the $\tilde{R}_s$ parameters, and thus in a much weaker way than the mutual information, which depends on the full distributions. Therefore, even though the parameters $\tilde{R}_s, \tilde{R}$ have to be, in turn, estimated from the data, this procedure is much more accurate than a direct estimate of the information.

To understand how one can estimate the number of "relevant" bins, we note that the number of relevant bins differs from the total number of bins allocated because some bins may never be occupied by responses to a particular stimulus. As a consequence, if $\tilde{R}_s$ is calculated using for each stimulus the total number of bins R, then the C_1 term, which is in this case equal to (2.7), turns out to overestimate the systematic error, whenever there are stimuli that do not span the full response set. On the other hand, the number of relevant bins differs also from the number of bins actually occupied, R_s, for each stimulus (with few trials), because more trials might have occupied additional bins. Again, it turns out that using the number of actually occupied bins R_s for calculating C_1 leads, when few trials are available, to an underestimate of the systematic error (the underestimation becoming negligible for $R/N_s << 1$ because R_s tends to coincide with $\tilde{R}_s$ for all stimuli).

It is clear that when N_s is small, more sophisticated procedures, such as Bayesian estimation, are needed to evaluate the quantities we are interested in. As mentioned above, Wolpert and Wolf (1995)[4] show how to calculate any function of the probabilities *given* the experimental frequencies, using Bayes rule. This requires some knowledge, or some assumption, on the *a priori* probability distributions of the probabilities. Since we do not have any knowledge of the prior, we do not see how to use this approach to estimate the mutual information itself, quantity which depends on the full details of the probability tables. Nevertheless, *a correction* to the mutual information depending on a few parameters, such as $\tilde{R}_s, \tilde{R}$, is likely to be well estimated also with a crude hypothesis about the prior probability functions. As an example, in the following we introduce a very simple procedure, based on the idea of using Bayes's theorem to reconstruct the true probabilities, supposing they are non-zero into $\tilde{R}_s$ intervals, and then choose an $\tilde{R}_s$ such that the expected number of occupied intervals (which can be calculated as a function of the Bayes estimates of the probabilities) matches the experimentally observed value.

Let us first recall some terminology from Bayes theory (Wolpert and Wolf 1995). If, for example, we want to measure a function $G(\{P(r|s)\})$ of the set of probabilities $\{P(r|s)\}$, and we know the prior probability distribution of the probabilities $\mathcal{P}(\{P(r|s)\})$, then the Bayesian estimate of the function $G(\{P(r|s)\})$ has the following expression as a function of the set of experimental data $\{n(r|s)\}$:

$$\widehat{G}(\{n(r|s)\}) = \int \left(\prod_r \mathrm{d}P(r|s)\right) \mathcal{P}(\{P(r|s)\}|\{n(r|s)\}) G(\{P(r|s)\}) \qquad (2.8)$$

where $\mathcal{P}(\{P(r|s)\}|\{n(r|s)\})$ is the "posterior" conditional probability of the underlying probabilities

[4] But see also the very recent paper by Bialek *et al.* (1996).

given the experimental outcome which is calculated with Bayes theorem:

$$\mathcal{P}(\{P(r|s)\}|\{n(r|s)\}) = \frac{\mathcal{P}(\{n(r|s)\}|\{P(r|s)\})\mathcal{P}(\{P(r|s)\})}{\mathcal{P}(\{n(r|s)\})} \tag{2.9}$$

where

$$\mathcal{P}(\{n(r|s)\}) = \int \left(\prod_r \mathrm{d}P(r|s)\right) \mathcal{P}(\{n(r|s)\}|\{P(r|s)\})\mathcal{P}(\{P(r|s)\}) \tag{2.10}$$

and the 'likelihood' probability distribution is binomially distributed:

$$\mathcal{P}(\{n(r|s)\}|\{P(r|s)\}) = N_s! \prod_r \frac{P(r|s)^{n(r|s)}}{n(r|s)!} \tag{2.11}$$

The procedure we use here to evaluate $\widetilde{R}_s$, for each stimulus s, is the following:

- We first pick for $\widetilde{R}_s$ one of the allowed values, $R_s \le \widetilde{R}_s \le R$.
- We construct, by using (2.8) the Bayes estimate $\widehat{P}(r|s)$ of the true probabilities given the experimental frequencies. The prior probability function $\mathcal{P}(\cdot)$ is chosen constant among the R_s non-empty bins, and for the other $\widetilde{R}_s - R_s$ empty bins is a different constant, fixed by requiring that the probability of that bin being empty is h_s times larger than the probability of being occupied, where $h_s = \frac{N_s}{R_s}$. This last requirement simply reflects the fact that when the responses are concentrated into a few bins (i.e. high $\frac{N_s}{R_s}$), the probability in the empty bins should be less than the probability assigned by a prior function constant on all the $\widetilde{R}_s$ bins. We want to emphasize that we use the constant ansatz for the prior probability distribution only because this is the simplest one. Of course, if, in particular cases, some reasonable assumption on the prior probabilities is available, this more detailed assumption can be used, and Bayes approach is expected to give better results.
- We pick other values for $\widetilde{R}_s$, and we finally choose as an estimate for $\widetilde{R}_s$ the value of $\widetilde{R}_s$ which gives the expectation value of the number of occupied bins:

$$< R_s >= \sum_r \left[1 - (1 - \widehat{P}(r|s))^{N_s}\right] \tag{2.12}$$

  closest to the experimental value of R_s.
- The procedure is the same for the evaluation of $\widetilde{R}$, the only difference being that the Bayesian estimate for $\widehat{P}(r)$ should be calculated from N, and not N_s, trials.

This estimation, although based on a very simple ansatz on the prior distributions, is sufficient to give, as we shall see below, good results even up to relatively small values of N_s.

The reason of this good estimation, in our opinion, is in the fact that only the parameters $\widetilde{R}_s$ have to be estimated based on the arbitrary ansatz, and the information I depends on them only in the correction terms.

The observation that the leading bias term (2.6) is, in general, probability dependent leads to a better understanding of the effectiveness with which the shuffled information can correct for limited samples. The probability dependency of (2.6) shows that the leading correction term is different for the true and the shuffled probabilities, and so subtracting the bootstrap correction does not cancel the leading $1/N$ contribution to the average error. If, for example, we have many zero-probability bins, the shuffling obviously overestimates the number of occupied bins, which implies that, in this case, the shuffled information is a (possibly high) overestimation of the bias, whereas the C_1^D term (2.6) continues to give a good estimate. Therefore, even when restricting to mutual information and discrete responses, there is no way, valid for all probability distributions, of relating the value of shuffled information to the value of the bias, as originally proposed by Optican *et al.* (1991) (but see also the discussion in Toveé *et al.* (1993)).

By considering the conditional information, we can give again an asymptotic expansion for the bias:

$$< I_N^D(s) > - I^D(s) = \sum_{m=1}^{\infty} C_m^D(s) \tag{2.13}$$

and the leading correction term is now:

$$\begin{aligned} C_1^D(s) &= \frac{1}{2N\log 2}\widehat{\sum}_i \langle \frac{1}{p_N(s)} \rangle [1 - p(i|s)] \\ &+ \frac{1}{2N\log 2}\widehat{\sum}_i \left\{ \frac{-p(i|s) + 2p^2(i|s)}{p(i)} - p(i|s) \right\} \end{aligned} \tag{2.14}$$

where the hat on the sum over response bins i denotes that only intervals of non-zero occupancy probability are to be considered, and in calculating explicitly the average of $< p^{-1}(s) >$ the instances with $p_N(s) = 0$ must be excluded. Estimating this expression (2.14) for the bias directly from real data is likely to lead, as for the C_1^D term, to undercounting if N is small. However, the dependence of $C_1^D(s)$ on the probabilities is not as simple as for C_1^D, and therefore a Bayesian estimate of $C_1^D(s)$ is more complicated and, without some knowledge on the prior, is not expected to work as well.

All the analytical results and considerations presented here are fully confirmed by computer simulations (Treves and Panzeri 1995; Golomb *et al.* 1996; Panzeri and Treves 1995b; Panzeri and Treves 1996b). The results of the simulations will be presented at various stages in this dissertation.

2.1.2 Convolution with continuous kernels and discretization

Let us now consider the case in which the regularization of the data is performed by first convolving the responses with a continuous kernel function and then discretizing the output space into R intervals $[m_{j-1}, m_j]$, $j = 1, \cdots, R$. With this data manipulation, smoothing (denoted by a tilde) followed by discretization, we obtain, from the N available stimulus-response pairs, a raw estimate of the information:

$$\tilde{I}_N^D(s) = \sum_{i \in \mathcal{R}} \tilde{p}_N(i|s) \log_2 \frac{\tilde{p}_N(i|s)}{\tilde{p}_N(i)}; \qquad \tilde{I}_N^D = \sum_{s \in \mathcal{S}} p_N(s) \tilde{I}_N^D(s) \ , \tag{2.15}$$

where the $\tilde{p}_N(\cdot)$'s are the experimental frequency tables, obtained by convolving the actual experimental responses r_j with some kernel distribution $K(r, r_j, \sigma)$ (*e.g.*, a gaussian one) and then integrating out the obtained probability density over the response intervals:

$$\tilde{p}_N(i|s) \equiv \frac{1}{N_S}\sum_{j=1}^{N_s} E_i(r_j;\sigma) \qquad \tilde{p}_N(i) \equiv \sum_{s\in\mathcal{S}} p_N(s)\tilde{p}_N(i|s) \tag{2.16}$$

where $E_i(r_j;\sigma)$ is the integral (over the i-th interval) of the kernel function centered in r_j:

$$E_i(r_j;\sigma) = \int_{m_{j-1}}^{m_j} drK(r,r_j,\sigma)\;. \tag{2.17}$$

The sum over j in (2.16) is performed over all the actual responses to stimulus s and the function K can depend on some parameter σ (such as the width in the case of a Gaussian convolution) which can be a function of the data distribution[5] itself: $\sigma = \sigma(s, r_j)$. For large N the raw response distributions approach the underlying ones and thus we can write:

$$\tilde{p}(i|s) = \int drE_i(r;\sigma)P(r|s) \qquad \tilde{p}(i) = \sum_{s\in\mathcal{S}} p(s)\tilde{p}(i|s) \tag{2.18}$$

Similarly, the estimate of the transmitted information tends to the information carried by the smoothed underlying probabilities:

$$\tilde{I}^D(s) = \sum_{i\in\mathcal{R}} \tilde{p}(i|s)\log_2\frac{\tilde{p}(i|s)}{\tilde{p}(i)}; \qquad \tilde{I}^D = \sum_{s\in\mathcal{S}} p(s)\tilde{I}^D(s)\;. \tag{2.19}$$

. Again, information values are in general dependent upon the smoothing and binning procedure adopted and, most importantly, upon the number of bins R and, now, upon the smoothing width. It is worth emphasizing that smoothing produces a further loss of information on the top of loss due to discretization alone, and if the rationale for smoothing is only to better control the finite sampling error, it is important to understand whether much better control can indeed be achieved.

For the leading terms in the bias

$$<\tilde{I}^D_N> - \tilde{I}^D \simeq \tilde{C}^D_1 \qquad <\tilde{I}^D_N(s)> - \tilde{I}^D(s) \simeq \tilde{C}^D_1(s) \tag{2.20}$$

we now find the expressions

$$\tilde{C}^D_1 = \frac{1}{2N\log 2}\left\{\widehat{\sum}_i \left[\left(\sum_{s\in\mathcal{S}}\frac{\tilde{q}(i|s)}{\tilde{p}(i|s)}\right) - \frac{\tilde{q}(i)}{\tilde{p}(i)}\right] - (S-1)\right\} \tag{2.21}$$

$$\begin{aligned}\tilde{C}^D_1(s) &= \frac{1}{N\log 2}\widehat{\sum}_i\left\{< p_N^{-1}(s) > \frac{\tilde{q}(i|s) - \tilde{p}^2(i|s)}{2\tilde{p}(i|s)} + \frac{\tilde{p}^2(i|s)\tilde{q}(i|s)}{\tilde{p}(i|s)}\right\} \\ &+ \frac{1}{2N\log 2}\widehat{\sum}_i\left\{\frac{\tilde{q}(i)\tilde{p}(i|s) - \tilde{p}(i|s)\tilde{p}^2(i)}{\tilde{p}^2(i)}\right\}\end{aligned} \tag{2.22}$$

[5] In the following, in evaluating averages, we assume that the regularization parameters do not fluctuate depending upon the outcome. When data-dependent parameters are used, we suppose that the fluctuations in information measures due to variations in the parameters are subleading with respect to those due to fluctuations of $P_N(\cdot)$.

where $\tilde{q}(\cdot)$ are evaluated from the underlying probability distributions as follows:

$$\tilde{q}(i|s) \equiv \int dr P(r|s) E_i^2(r|s) \qquad \tilde{q}(i) \equiv \sum_{s\in\mathcal{S}} p(s)\tilde{q}(i|s) \; . \tag{2.23}$$

The correction terms (2.21) and (2.22) are now dependent upon both the underlying probability and the chosen regularization. The first dependence raises, as in the discrete case, the problem of how to estimate the corrections (2.21) and (2.22) from the data, and, in particular, how to avoid undercounting the bins with non-zero probability over which to take the sums in (2.21) and (2.22). If one convolves the responses with an infinite range distribution, such as the Gaussian, no interval remains strictly empty after the convolution, and then the potential underestimation of the correction is less important than in the discrete case. Even with a Gaussian convolution, however, some undercounting might occur because of numerical truncation. If we suppose that the typical smoothing width is small compared with the typical bin length, we can take the smoothing to have significant effects only in the nearest intervals. In this case, an approximate form of the averaged underestimation can be worked out [6]:

$$\begin{aligned} \tilde{C}_1^D - <(\tilde{C}_1^D)_N> \;&\equiv\; \Delta(\tilde{C}_1^D) \\ &= \frac{1}{2N\log 2}\left\{\sum_{s\in\mathcal{S}} \widehat{\sum}_i [1-\tilde{p}(i-1|s)-\tilde{p}(i|s)-\tilde{p}(i+1|s)]^{N_s}\right\} \\ &- \frac{1}{2N\log 2}\left\{\widehat{\sum}_i [1-\tilde{p}(i-1)-\tilde{p}(i)-\tilde{p}(i+1)]^{N}\right\} \end{aligned} \tag{2.24}$$

This approximate form for the underestimation of $\tilde{C}_1^D$ captures just the fact that, when the smoothing width is small with respect to the typical bin length, in a bin the smoothed probability $\tilde{p}(i|s)$ can be considered null only if we do not have outcomes in the nearest bins. In this case, $\Delta(\tilde{C}_1^D)$ can be added to $\tilde{C}_1^D$ to marginally improve the estimation of the bias.

As for the validity of the bootstrap procedure, the fact that the correction terms (2.21) and (2.22) are now also regularization dependent, further complicates the analysis. If the convolution width is not too large, we can expect that the procedure will tend to overestimate the response range for some stimulus (due essentially to the same mechanism which appears in the discrete case) and then to overestimate the bias in the case in which one observes very different response ranges to different stimuli. Thus, in this situation the shuffled information might be larger than the bias. On the other hand, when the convolution width is large and data dependent (for example, determined by the standard deviation of the responses to each stimulus, as in Optican and Richmond (1987)), or, in general, when the regularization is data dependent (and then different for the actual

[6] An expression for ΔC_1^D can also be derived for the discrete case (in fact, a simpler and exact expression). However, in that case, it gives typically large contributions which are themselves difficult to estimate from the data, so that in the discrete case it is much better to use the Bayesian algorithm to estimate $\tilde{R}, \tilde{R}_s$ instead.

and the shuffled responses), the shuffled information might not be an upper bound to the bias, but it might easily underestimate the bias, when reflecting a stronger regularization. Thus, in this situations it is not safe to rely on the bootstrap procedure, either to correct the raw estimate by subtraction, or to conclude, when the shuffled information is very small, that the average bias itself must be small.

2.1.3 Neural network regularization

In this subsection we briefly review the neural network regularization introduced in Hertz *et al.* (1992) and Kjaer *et al.* (1994), and we discuss how the bias can be calculated in a similar fashion.

The idea of Hertz and coworkers is to use a two-layer network trained by back-propagation to classify the neuronal responses according to the stimuli that elicited them. The network uses sigmoidal activation for the nodes in the hidden layer and exponential activations for the nodes in the output layer, with the sum of the outputs normalized to one after each step. The input to the network is the quantified output of the biological neuron: the spike count, the first n principal components or both. There is one output unit for each stimulus, and, after training, the value of output unit number i is an estimate $\hat{P}(i|r)$ of the conditional probability that response r was elicited by stimulus i. Summing over responses belonging to the same stimulus s, one finally obtains the conditional probability that a stimulus s is recognized as the i-th:

$$\tilde{p}_N(i|s) \equiv \frac{1}{N_s}\sum_{j=1}^{N_s} E_i(r_j;\omega) \qquad \tilde{p}_N(i) \equiv \sum_{s\in\mathcal{S}} p_N(s)\tilde{p}_N(i|s) \tag{2.25}$$

where

$$E_i(r;\omega) = \frac{\exp\left[\sum_l (W_{il}H_l + B_l)\right]}{\sum_{j=1}^{S}\exp\left[\sum_l (W_{jl}H_l + B_j)\right]} \tag{2.26}$$

and the hidden unit activation function is given by:

$$H_l(r) = \tanh\left[\sum_{m=1}^{Q} \omega_{lm} r_m + b_l\right] . \tag{2.27}$$

In (2.26),(2.27), H_l depends on Q variables r_m chosen to describe the raw neuronal response, whereas W, ω, b, B are parameters for the neural network, selected according to a certain optimization procedure (see below). After this regularization, the output space becomes an S-dimensional discretized set, equivalent to the stimulus set, which could be called the set of 'posited stimuli', and the conditional probability $\tilde{p}(i|s)$ (2.25) can be interpreted as the conditional probability with which a response elicited by stimulus s may be attributed to stimulus i.

The parameters of the algorithm are controlled by cross-validation. The data are divided into training and test sets, and, for each division, the training is stopped when the test error, defined as

$$E = -\sum_{\mu} \log_2 \hat{P}(s^{\mu}|\mathrm{r}^{\mu}), \tag{2.28}$$

(the negative log-likelihood or crossed-entropy) reaches a minimum. In Eq.. (2.28) the index μ labels the trials in the test set, and s^{μ} is the stimulus that actually evoked the response r^{μ} observed in that trial. Since parameters are adjusted on training data, and information is calculated on test trials only, in the context of evaluating the finite size bias, the parameter N is the number of test stimulus-response pairs. Without going further into details of the procedure [7], is is sufficient, for our purposes, to remark that the form of regularized probability distributions (2.25) is the same as in (2.16), except that $E_i(r)$ is no longer evaluated simply by integrating a continuous kernel over the i-th bin, but with the more complicated rule (2.26). This does not affect the result for the bias, which are therefore the same as in subsection 2.1.2, with the only difference that $E_i(r)$ must be computed from (2.26) instead of (2.17).

2.1.4 Convolution with continuous kernels

Finally, let us consider the case in which raw responses are manipulated by convolving them with a continuous kernel function, as before, but without a subsequent discretization of the output space. The raw information estimates now read:

$$\tilde{I}_N(s) = \int dr \tilde{P}_N(r|s) \log_2 \frac{\tilde{P}_N(r|s)}{\tilde{P}_N(r)}; \qquad \tilde{I}_N = \sum_{s \in \mathcal{S}} p_N(s) \tilde{I}_N(s) \ , \tag{2.29}$$

where the $\tilde{P}_N$'s are the experimental distributions, obtained by convolving experimental responses r_j with a continuous kernel function $K(r, r_j, \sigma)$:

$$\tilde{P}_N(r|s) \equiv \frac{1}{N_S} \sum_{j=1}^{N_s} K(r, r_j, \sigma) \qquad \tilde{P}_N(i) \equiv \sum_{s \in \mathcal{S}} p_N(s) \tilde{P}_N(r|s) \tag{2.30}$$

The sum over j in (2.30) is performed over all the actual responses to stimulus s. As N increases, the raw response distributions approach the underlying ones:

$$\tilde{P}(r|s) = \int dr_1 P(r_1|s) K(r, r_1, \sigma) \qquad \tilde{P}(r) = \sum_{s \in \mathcal{S}} p(s) \tilde{P}(r|s) \tag{2.31}$$

and the raw estimates of information tend to:

$$\tilde{I}(s) = \int dr \tilde{P}(r|s) \log_2 \frac{\tilde{P}(r|s)}{\tilde{P}(r)}; \qquad \tilde{I} = \sum_{s \in \mathcal{S}} p(s) \tilde{I}(s) \ . \tag{2.32}$$

[7] However, it should be noted that the mutual information defined in Kjaer *et al.* (1994) (see also eq. (3.8)) is not, in our opinion, fully equivalent to the mutual information carried by the regularized probabilities (2.19)

. The expressions we find in this case are:

$$\begin{aligned}\tilde{C}_1 &= \frac{1}{2N\log 2}\left\{\int dr\left[\left(\sum_{s\in\mathcal{S}}\frac{\tilde{Q}(r|s)}{\tilde{P}(r|s)}\right)-\frac{\tilde{Q}(r)}{\tilde{P}(r)}\right]-(S-1)\right\} && (2.33)\\ \tilde{C}_1(s) &= \frac{1}{N\log 2}\int dr\left\{< p_N^{-1}(s) > \frac{\tilde{Q}(r|s)-\tilde{P}^2(r|s)}{2\tilde{P}(r|s)}+\frac{\tilde{P}^2(r|s)-\tilde{Q}(r|s)}{\tilde{P}(r|s)}\right\}\\ &+ \frac{1}{2N\log 2}\int dr\left\{\frac{\tilde{Q}(r)\tilde{P}(r|s)-\tilde{P}(r|s)\tilde{P}^2(r)}{\tilde{P}^2(r)}\right\} && (2.34)\end{aligned}$$

where:

$$\tilde{Q}(r|s)=\int dr_1 P(r_1|s)K^2(r,r_1,\sigma)\qquad \tilde{Q}(r)=\sum_{s\in\mathcal{S}}p(s)\tilde{Q}(r|s) \tag{2.35}$$

In the continuous case, the problem of underestimation of the correction terms (2.33) and (2.34), when calculated form data. is not important, since this problem is intrinsically related to the discretization of the output space. This continuous case is rather academic anyway, as in practice one usually performs the required integrals on the computer by first discretizing and then taking sums. It remains true, however, that one is close to the continuous limit, and the simple expressions above hold, whenever the discretization is sufficiently fine with respect to the width of the kernel.

2.2 The bias with fixed number of trials per stimulus

In the previous section we studied the finite size distorsions when the stimuli are drawn at random from a discrete set. Here we present the result valid when, instead, the experimental frequency of presentation of stimuli does not fluctuate, but it is set exactly to its probability: $p_N(s) \equiv p(s)$. The calculation of the bias is very similar to that presented for the previous case, but with the obvious difference that, in evaluating averages as in (A.4)-(A.6), one has to average over responses in the same way as detailed in appendix A, but *not*, as before, over $p_N(s)$ with the multinomial distribution.

We report only the results for the case of convolution with a kernel $K(r, r_j, \sigma)$ and discretization into R intervals:

$$\begin{aligned}\tilde{C}_1^D &= \frac{1}{2N\log 2}\left\{\widehat{\sum}_i\left[\sum_{s\in\mathcal{S}}\left(\frac{\tilde{q}(i|s)}{\tilde{p}(i|s)}+\frac{p_N(s)\tilde{p}^2(i|s)}{\tilde{p}(i)}\right)-\tilde{q}(i)-\tilde{p}(i)\right]-S\right\} && (2.36)\\ \tilde{C}_1^D(s) &= \frac{1}{N\log 2}\widehat{\sum}_i\left\{< p_N^{-1} > \frac{\tilde{q}(i|s)-\tilde{p}^2(i|s)}{2\tilde{p}(i|s)}+\frac{\tilde{p}^2(i|s)-\tilde{q}(i|s)}{\tilde{p}(i)}\right\}\\ &+ \frac{1}{2N\log 2}\widehat{\sum}_i\left\{\frac{\tilde{p}(i|s)\tilde{q}(i|s)}{\tilde{p}^2(i)}-\sum_{s'}\frac{p(s')\tilde{p}^2(i|s')\tilde{p}(i|s)}{\tilde{p}^2(i)}\right\} && (2.37)\end{aligned}$$

where the notations is the same as in section 2.1. The results corresponding to the other regularizations considered in the previous section can be easily derived by taking the appropriate limits, as explained in Appendix A.

2.3 Tests of the range of validity of analytical results

To support these analytical results, and to compare different correction procedures, we perform explicit numerical simulations. First we study the effects of two regularizations often used in data analysis, pure discretization and convolution with Gaussians, on both finite size effects and loss of information due to regularization. In addition, we compare the effectiveness of subtracting C_1 terms with that of the bootstrap procedure. At the end of the section, we use another set of simulated data to compare the jackknife correction and the analytical subtraction. The neural network regularization will be considered in next chapter, together with other decoding procedures.

Let us start by choosing as "test" underlying probabilities Poisson distributions, which are fair simple models of the spontaneous activity of neurons under certain conditions (Abeles *et al.* 1990; Levine and Troy 1986; Scobey and Gabor 1989). We generate the distribution of mean firing rates $\bar{r}(s)$ corresponding to each stimulus s by selecting a random variable x from a flat distribution in the interval $[0, 1)$, and then setting

$$\bar{r} = -\log\left(1 - \frac{x}{2a}\right) \quad \text{if} \quad x < 2a, \quad \bar{r} = 0 \quad \text{if} \quad x > 2a, \tag{2.38}$$

The parameter a is, on average, the sparseness (Treves 1990) of the firing rate distribution. The number of spikes n recorded on each trial over a period t (t = 500 msec. in the present simulations) follow the Poisson distribution

$$P(n|s) = \frac{[\bar{r}(s)t]^n \exp -[\bar{r}(s)t]}{n!}. \tag{2.39}$$

To measure, from N trials, the information carried by the firing rates generated in this way, we use the following regularization procedure: the range of responses is discretized into a preselected number R of bins, with the bin limits selected so that each bin contains the same number of trials within ± 1 (equipopulated bins). A smoothing procedure is applied by convolving the individual values with a gaussian kernel. The smoothing width has an overall multiplicative parameter γ (successively increased in the simulations to test how different convolution widths influence the finite size effect) and is proportional to the square root of each value (the proportionality factor is set such that on average the smoothing widths match $\gamma\sigma_s$, where σ_s is the standard deviation of the firing rate of each stimulus).

Figures 2.1 and 2.2 show, for different sample sizes, how our correction procedure improves both on raw estimates of mutual information and on the bootstrap procedure of subtracting the shuffled information; moreover, the figures illustrate the effect of smoothing the responses on the accuracy of information estimates. When no smoothing is applied (fig. 2.1), the asymptotic value of discretized information (dashed line) is only a few percent below the "true", or unregularized, value (the full line). The finite sampling bias in raw information estimates becomes of similar size to the loss due to discretization, and roughly compensates for it, only if as many as 256 trials per stimulus are available. The bootstrap procedure reduces the bias to similar levels earlier, at roughly 100 trials per stimulus (but note that the remaining bias is also downward and does not compensate for the regularization loss). Our correction procedure using Bayesian estimates for $\tilde{R}_s, \tilde{R}$ allows the same precision already for $N_s \sim R$ (in this case, $R = 16$). In contrast, using correction terms based on the number of response bins actually occupied, or on the total number of bins, is not much more effective than the bootstrap or even raw estimates. When a weaker (fig. 2.2,top) or stronger (fig. 2.2,bottom) smoothing is applied before discretizing the responses, the loss of information due to regularization becomes much larger and more important than finite sampling errors. Nevertheless, the latter are still controlled effectively by our correction procedures. Although the procedure is less refined than in the discrete case, convergence to the asymptote is faster ((but the asymptote is strongly downward biased, particularly in fig. 2.2,bottom). The conclusion appears to be that smoothing with Gaussians does more damage than good, although we note that (i) there may be other reasons for smoothing with Gaussians (e.g. avoiding edge effects), and (ii) when it is known that the smoothing width is small with respect to the relevant differences in the responses, smoothing may induce much smaller loss than in our examples, with possibly faster convergence with sample size.

Figure 2.3 shows the value of subtracting the $\tilde{C}_1^D(s)$ term in the case of conditional information. In this case, no shuffling of the stimulus-response pairs would be applicable, whereas it is evident that our subtraction yields reasonable results, bringing the corrected values within the narrow range spanned by the difference between real and regularized information values.

After having studied the effectiveness of the analytical and bootstrap corrections, let us look at the jackknife procedure. The jackknife, introduced by Quenouville (1949), is a non-parametric method which allows estimates of the bias of generic probability functions. Let us briefly review this technique, by restricting ourselves to the case of mutual information. For the more general case, and for an overview on non-parametric estimators, we refer to the excellent review written by Efron (1982).

Quenouville's method is based on sequentially deleting experimental responses r_j, and recomputing the mutual information I from $N-1$, instead of N, data. Denoting by $I_{N-1;(r_j)}$ the value of information obtained using all data points but r_j, and introducing the following quantity:

$$\hat{I}_{N-1} \equiv \frac{1}{N}\sum_{j=1}^{N} I_{N-1;(r_j)} \ , \tag{2.40}$$

the Quenouville estimate of the bias has the following expression:

$$< I_N > - I = (N-1)\left(\hat{I}_{N-1} - < I_N >\right) \tag{2.41}$$

It is not difficult to see that the estimate (2.41) is based essentially, like our estimate, on the assumption that only the $1/N$ contribution to the bias is important. In fact, on

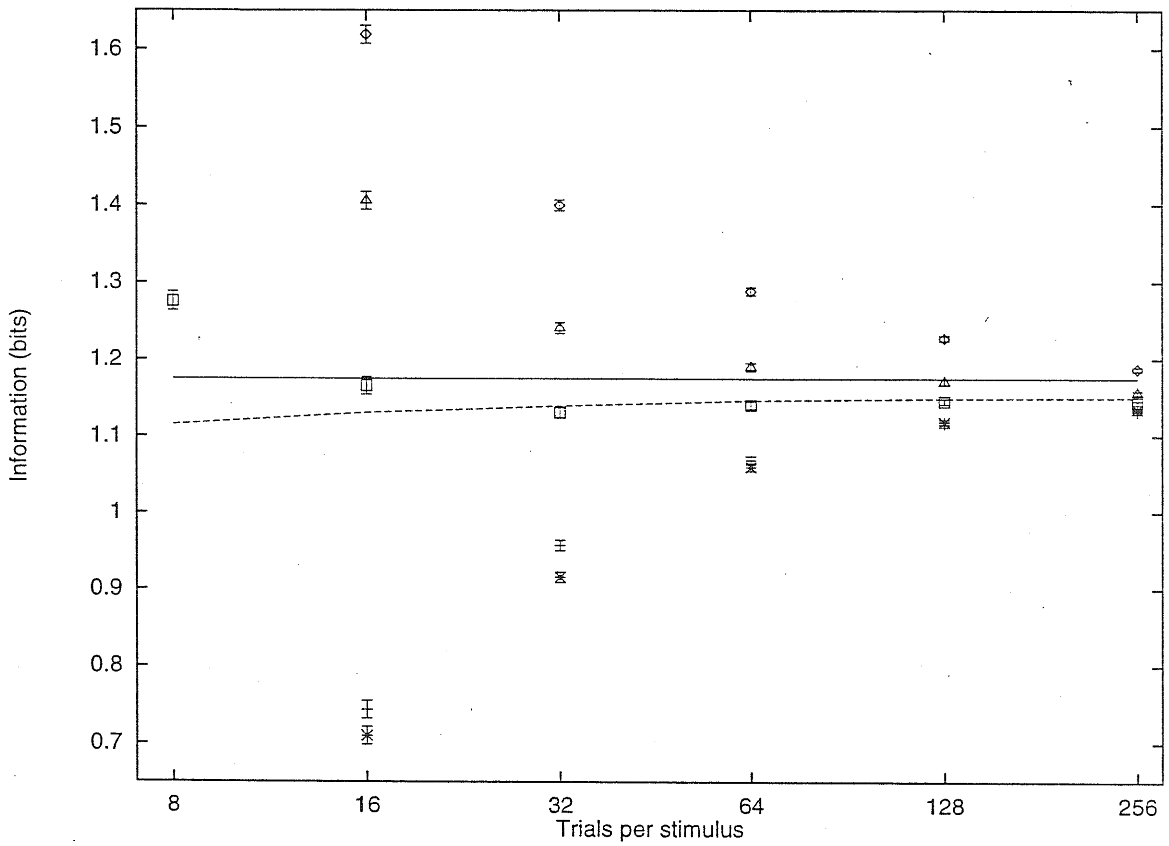


Figure 2.1:

Mutual information values for the distribution of stimuli and Poisson responses described in the text (the sparseness of the mean firing rates is $a = 0.4$), with $S = 16$ and $R = 16$ and different values of N_s. This panel corresponds to pure discretization ($\gamma = 0$). The full line is the real value of the information in the distribution and the dashed line is the *regularized* value, that could be extracted from an infinite sample of data, after the prescribed regularization of the responses. Compared to these reference values are, for each N, the raw estimate ($\diamond$), the estimates corrected by subtracting the C_1 term calculated by estimating the relevant bins by counting the number of actually occupied ones ($\triangle$), estimating the effective bins with the Bayesian procedure described in the text ($\square$), taking all bins to be relevant ($\widetilde{R}_s = R = 16$) ($+$) and the estimate corrected by the bootstrap method ($\star$). Each value is plotted with the standard deviation of the mean of 100 measurements. Note that the N_s axis is on a logarithmic scale.

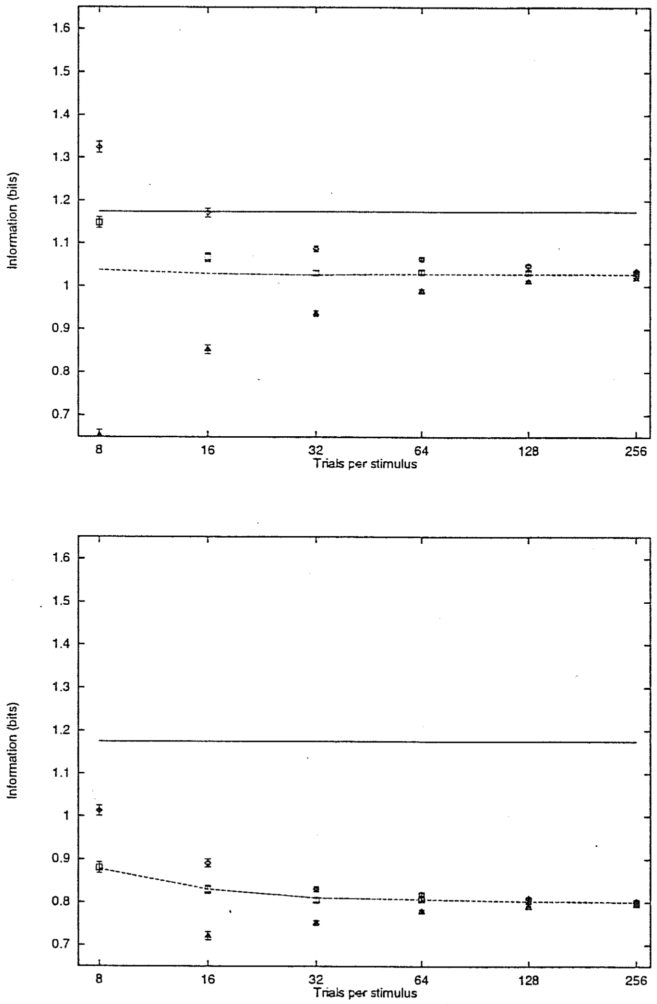


Figure 2.2:

Mutual information values for the distribution of stimuli and Poisson responses, as in the previous figure. The two panels correspond now to Gaussian convolution with (top) $\gamma = 0.5$; (bottom) $\gamma = 1.0$. The symbols are the same as in previous figure.

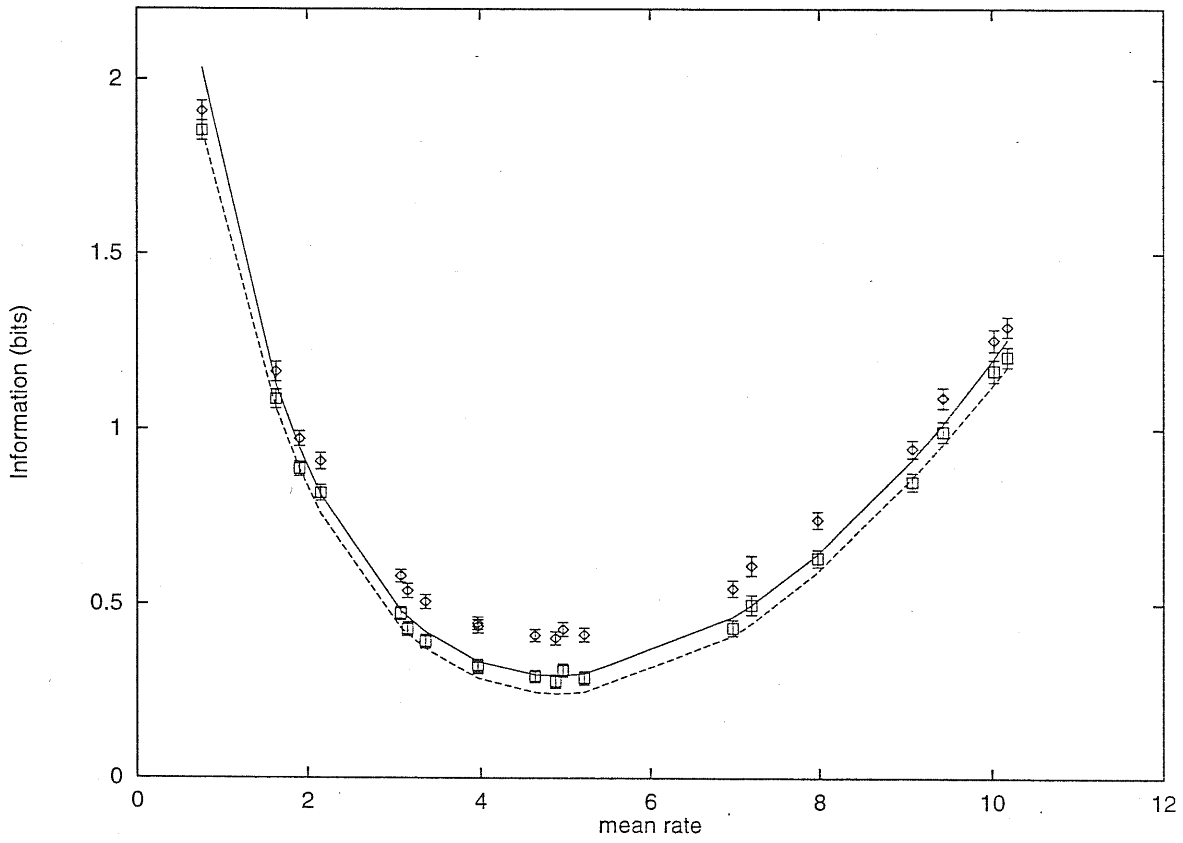


Figure 2.3:

Values for the information conditional to which of $S = 20$ (simulated) stimuli was presented, plotted against the mean rate $\bar{r}(s)$ to each stimulus (on arbitrary scale). The firing rates are distributed with sparseness $a = 0.7$. Again, the full curve indicates the real and the dashed one the regularized information values; and the symbols indicate raw and subtracted measures, each with standard deviation of the mean over 100 measures. Here $R = 10, N = 300, \gamma = 0.33$.

one hand we know that $< I_N >$ can be expanded in powers $1/N$, eq. (2.5), and the coefficients (namely C_m) do not depend on N. On the other hand, $\hat{I}_{N-1}$ can be expanded in the same way (and with the same coefficients) in powers of $1/(N-1)$. Using the two expansions, one can easily verify that the quantity obtained by subtracting to $< I_N >$ the bias estimate (2.41), is biased only $O(1/N^2)$, compared to $O(1/N)$ for the naive estimator.

To test the effectiveness of the jackknife versus the analytical correction, we use here the firing rate distributions obtained by a model of the response of parvocellular and magnocellular LGN cells to a set of 32 visual stimuli. This model, introduced by Golomb *et al.* (1994), is explained in more detail in section 3.1.1. What is interesting here is to compare the asymptotic value of the information, carried by the underlying probabilities, to the value one can obtain from a finite number of samples. Here we regularize the responses by pure discretization, choosing $R \sim N_s$, to be at the limit of the region where the correction procedure, as we have seen before, is still expected to work. In this case the downward bias produced by discretization is small, because each response, quantified as the number of spikes over 250 ms after the stimulus onset, is just an integer ranging from 0 to the maximal number of spikes (denoted in the following as $\mathrm{NOS}_{\mathrm{max}}$, and equal to 25 for the parvocellular cell and 34 for the magnocellular cell). If $N_s > \mathrm{NOS}_{\mathrm{max}}$, it is thus enough to fix R equal to $1 + \mathrm{NOS}_{\mathrm{max}}$. In fig. 2.4 we compare the true information values to the estimate obtained from limited samples with both our procedure and the jackknife. We see that the two corrections give similar results, as expected, but our method converges more rapidly to the asymptotic information value.

The effects of neural network regularization will be separately discussed in next chapter (sec. 3.1), since it is a regularization procedure, and not a correction for finite sampling.

2.4 Comparison among different correction procedures

In this section we summarize the results we have obtained, with the work presented in this chapter, about the relative value of various correction procedures.

2.4.1 Bootstrap

This procedure is flawed in several ways:

- Subtracting the bootstrap correction does not cancel the leading $1/N$ contribution to the average error, and typically the bootstrap strongly overestimates the bias.

- Data regularization can affect the raw and shuffled information to different degrees.

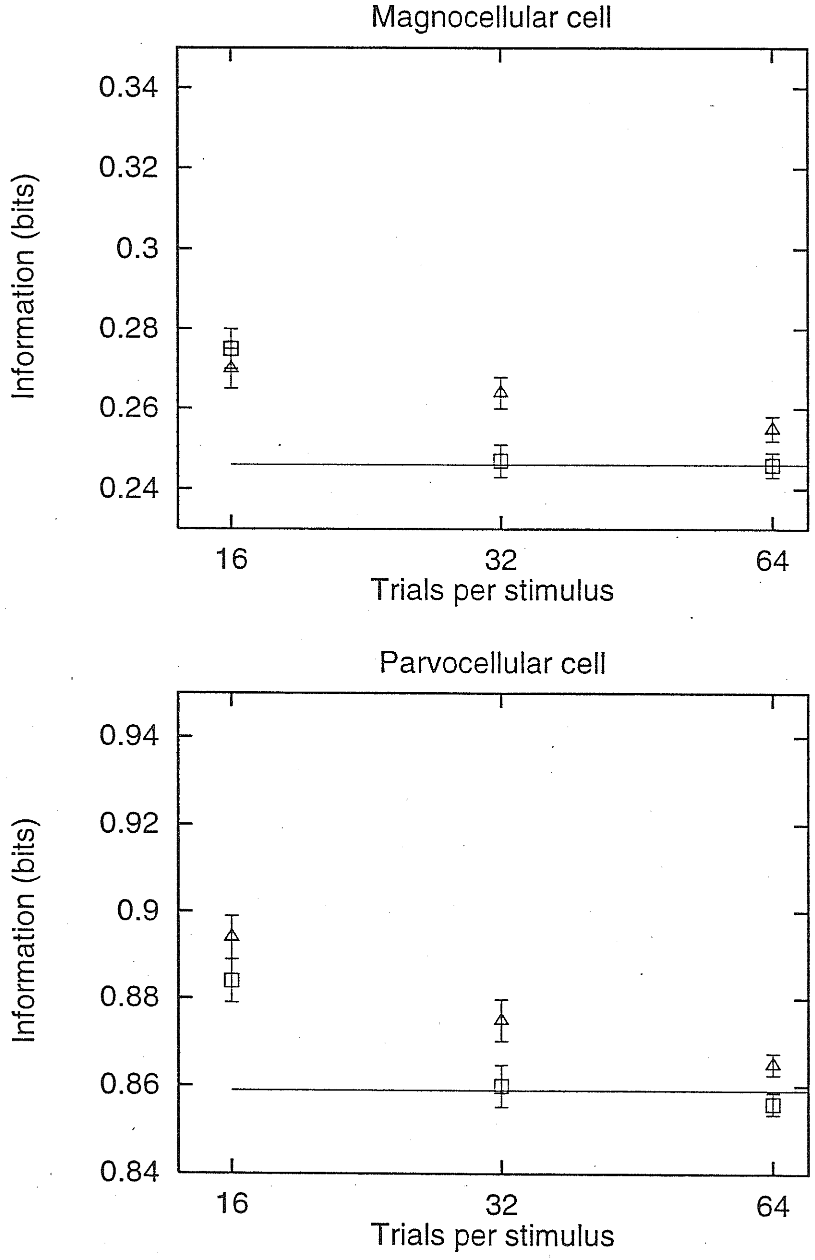


Figure 2.4:

Mutual information values extracted from the firing rates distributions of LGN magnocellular (top) and parvocellular (bottom) cells, in response to a set of 32 Walsh pattern. The simulated responses are generated as explained in section 3.1.1. The full line is the real value of the information contained in the distributions. Compared to this reference value are, for each N_s, the estimates corrected by subtracting the C_1 term (□) and the jackknife estimator (△), each with standard deviation of the mean over 100 measures.

In conclusion, there is no way, valid for generic probability distributions, to relate the value of the shuffled information to the bias, and we do not recommend the use of this method anymore.

2.4.2 Neural Network

This method, instead of correcting for finite size effects, squeezes the experimental responses through an artificial neural network, trained so as to maximize the probability that a stimulus is correctly recognized. So, this procedure is more properly classifiable as a "decoding", or regularization, instead of a "bias correction", technique. Computer simulations indicate that the regularization induced by the network is strong enough to dispose of the finite sampling bias, at the price of underestimating information values, especially for higher dimensional codes, which are more strongly regularized. This strong regularization is not necessary with regard to low dimensional and simple codes, like single unit firing rates, which can be studied with high precision using a simple "binning and correcting" technique. When high dimensional codes are concerned, the network is still able to control finite size effects, but better estimates can be often achieved with milder decoding procedures coupled with finite size corrections. These decoding procedures will be introduced and discussed in the next chapter.

2.4.3 Analytical estimator

When responses are regularized by discretizing into R bins, our corrections works even down to $N_s \simeq R$. When convolutions with continuous distributions are applied before discretization, the convergence of the finite sample correction is even faster, at the prize of a further loss of information due to regularization.

2.4.4 Jackknife

The jackknife is essentially based on the same assumptions leading to our analytical corrections, and allows estimation of the bias while avoiding the explicit calculation of the latter. Results obtained with the jackknife seem to be fairly similar to those that can be obtained by subtracting the C_1 term. The main problems with the jackknife estimator are:

- It involves recalculation of $N + 1$ information quantities. This make the procedure inapplicable to the study of information carried by large populations of cells. In fact in this case, due to the large dimensionality of response space, information has to be extracted through time consuming decoding procedures, and each single recalculation of information can require a lot of time.

- It involves subtraction of two quantities of order N, and thus it suffers from the same imprecision of any algorithm that determines a quantity as the result of the subtraction of two large and nearly equal terms. This ultimately leads to a slower converge with N towards the asymptotic result.

However, this correction, unlike the bootstrap one, is based on meaningful assumptions and can lead to unbiased estimates.

2.5 How best to choose the number of bins?

In previous sections we have discussed the possible problems arising when convolving data with continuous distributions, and established the range of reliability of our correction, which in the discrete case works even down to $N_s \simeq R$. Given the effectiveness of the binning procedure, we recommend limiting the regularization to simple binning. unless motivated by other considerations (*e.g.,* that presented in sec 2.3, or, when the dimensionality of responses is so big, that it is impossible to distribute, among the various dimensions, the allowed bins in some meaningful way). The important question, of the "optimal choice" of the number of bins for an experiment with S stimuli and N_s trials per stimulus, still remains. A reasonable answer to this question can be to choose $R \sim N_s$, to be at the limit of the region where the correction procedure is expected to work, and thus still be able to control finite sampling, while minimizing the downward bias produced by binning into too few bins [8]. This choice should effectively minimize the combined error due to regularization and finite sampling. In fig. 2.5 the information estimates obtained by choosing $R = N_s$ are compared, for different values of N_s, to the full, unregularized, value of the information carried by the Poisson distribution of responses introduced in sec. 2.3. It can be noticed that, in this situation, results appear to be a reasonable estimate of the full value of the information in the whole N_s range explored.

The subtraction procedure based on binning indicates how to work out the minimum number of trials which should be used in experiments. The analytical correction functions reasonably up to $N_s \simeq R$, and the minimum number of response bins which may, if the appropriate code is used, not throw away information, is just the same as the number of stimuli, $R = S$. Therefore a minimum of $N_s = S$ trials per stimulus is a fair demand to be made on the design of experiments from which information estimates are going to be derived.

[8] The choice of the number of bins in each dimension, for a multidimensional code, remains however somewhat arbitrary, and particularly subtle when comparison among codes of different dimensionalities are concerned. This problem will be discussed in next chapter (sec. 3.1)

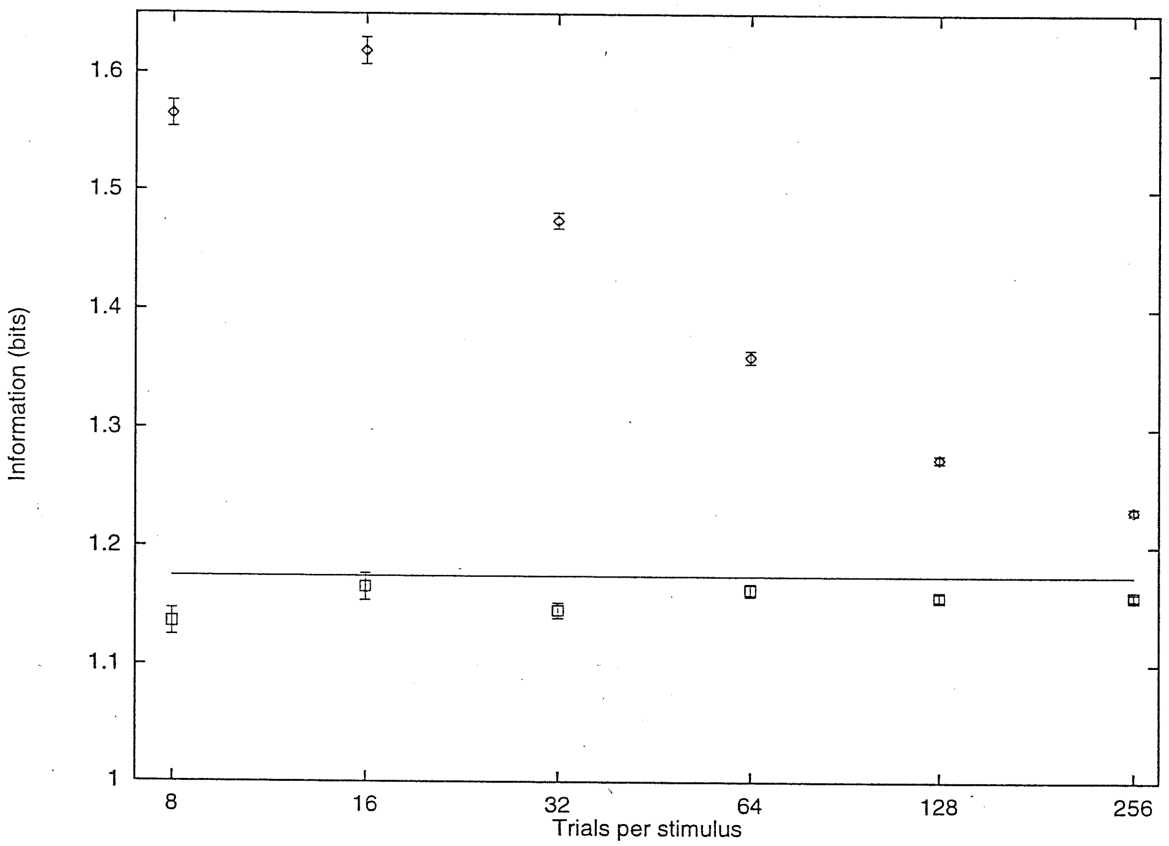


Figure 2.5:

Mutual information values for the distribution of stimuli and Poisson responses described in the text. Here $a = 0.4$, $S = 16$ and $R = 16$ and the response space is purely discretized. The symbols have the same meaning as in fig. 2.1. Note that for $N_s = 16$, also $R = 16$ and the result is the same as shown in fig. 2.1. For higher values of N_s, results approach the unregularized value of the information, whereas in fig. 2.1 they approached the value regularized with $R = 16$ bins.

Chapter 3

Applications of information theoretical analyses to neuronal data

In chapter 2 we have studied how to minimize the (downward) bias due to regularization in information measures, while being still able to control finite sampling. In this chapter we wish to discuss, by examples, how to apply our results to information theory based experiments, involving different types of neuronal codes. The degree of accuracy to which each problem can be studied will lead us to discuss which kind of questions regarding neural codes can be reliably investigated, and to introduce information-theoretic quantities that can help in understanding the modalities of neuronal representation.

This chapter is organized as follows. In the first section, we discuss, using detailed simulations of responses of LGN cells to visual stimuli, the degree of accuracy to which we can quantify the role of temporal modulation of responses in information processing. The problem is made particularly difficult as it involves comparison of codes of different dimensionalities (*e.g.,* firing rates versus a few principal components of the spike train). We find that one dimensional codes can be studied by the simple binning with a very high degree of accuracy, whereas higher dimensional codes (like the three first principal components) can be studied only within 5-10% of the correct values. Raw measures (Optican *et al.* 1987), or measures corrected with more empirical procedures (Optican *et al.* 1991) give strongly biased (and thus useless) results. The second section is specialized to the case of information contained in the firing rates of single cells. We will discuss how to use the high precision of these information measures to study the contribution of different properties of the neuron, such as noise or the graded nature of responses, to the information processing at different time scales. Furthermore, we show that the initial rate at which a neuron transmits information depends only on the mean firing rates, and is simply related to the sparseness of the neuronal representation. Section 3.3 is devoted to the extraction of information from large population of cells, a problem which is particularly interesting, since data from simultaneous recording of groups of neurons are becoming available, and is particularly difficult, due to the large dimension of the response space. Since in this case the simple binning alone is no more enough, we introduce, discuss and test different decoding procedures.

In each section we provide examples of how our techniques could be used to analyze neuronal responses. In particular, we report original results of a study of the neural coding strategy used by single units, and by populations of neurons, in the primate hippocampus and in the rat somatosensory cortex.

3.1 Information contained within the spike train of single cells

One of the most basic questions about the neuronal code is what is the relevant parameter for information processing at the single cell level. An answer to this question involves a comparison between the amount of information contained in the firing rates and in the full temporal structure of the spike train. A complete description of the latter could be given by dividing the recorded time window into temporal bins smaller than the refractory period, and assigning a binary value (1 or a 0) to every bin in which there is, or there is not, a spike. A precise extraction of information from this complete characterization of the spike train as a binary string has been carried out in studies of the sensory periphery of the fly nervous system (Strong *et al.* 1996), but it requires so much data that it cannot be performed for the mammalian cortex. In the latter case however, principal component analysis can be used to compress the response space with minimal information loss (Optican and Richmond 1987). If a few principal components (PCs) are enough to reconstruct with fidelity the spike train, or at least only a few PCs contribute significantly to information processing, the experimental study of the role of temporal modulation in information processing can be performed by comparing the amount of information carried by firing rates and by those PCs of the spike train. The difference between these two entities is however difficult to quantify. In fact it involves calculation of informations which are very differently biased, as they are carried by responses with different dimensionalities. The difficulty of overcoming this problem has led, in the last few years, to a series of paper reporting opposite results (see *e.g.,* Optican and Richmond 1987; Optican *et al.* 1991; Tovée *et al.* 1993; Kjaer *et al.* 1994).

In this section, we test procedures designed to overcome this problem, to make it clear which results present in the literature should be considered correct. We perform the test by comparing mutual information calculations from large data sets with those calculated from smaller ones. The only way to get such large data sets is through simulations. We use here the database of artificial spike trains created, by David Golomb, by means of a model of the response of lateral geniculate nucleus (LGN) neurons (Golomb *et al.* 1995). The model is based on experimentally-measured spatiotemporal receptive fields of LGN neurons (Reid and Shapley 1992). We apply the procedure of binning-and-correcting, discussed in chapter 2 (here binning means that we regularize the response space by pure discretization). We discuss how to perform the discretization when a comparison of codes of different dimensions is concerned. Furthermore, we use the same set of simulated data to test also the neural network method of estimating the conditional probabilities (Heller *et al.* 1995, Hertz *et al.* 1992, Kjaer *et al.* 1994). We find that both methods yield accurate results for one-dimensional codes, even for a relatively small number of samples. Moreover, estimates of the extra information carried in three-dimensional codes are also reasonable, within 0.05-0.1 bits (about 10%) of the correct values, although the network appears to give results more biased downward than those obtained with a very simple binning procedure. Raw measures (Optican *et al.* 1987), or measures corrected with more empirical procedures (Optican *et al.* 1991) give instead strongly biased results.

3.1.1 Methods

Producing simulated data

The spike trains are created using a model of the response of parvocellular and magnocellular LGN cells, as described in Golomb *et al.* (1994). In brief, the spatiotemporal receptive fields $R(\vec{r},t)$ of the two cells types in response to an impulse in space and time were measured (Reid and Shapley 1992; data presented in Fig. 1 of Golomb *et al.* 1994). The set of stimuli $\{\sigma_s\}, s = 1 \cdots S(=32)$ includes 4×4 flashed Walsh figures in space and their contrast reverse:

$$\sigma_s(\vec{r},t) = u_s(\vec{r})\Theta(t) \tag{3.1}$$

where $u_s(\vec{r})$ are the spatial Walsh figures and $\Theta(t)$ is the Heavyside function ($\Theta(t) = 1$ if $t > 0$ and is 0 otherwise). The ensemble-average response to the sth figures $Z_s(t)$ is calculated by centering it on the receptive field center, convolving it with the spatiotemporal receptive field, adding the constant baseline Z_0 corresponding to the spontaneous firing, and rectifying at zero response:

$$Z_s(t) = \Theta\left[Z_0 + \int d\vec{r} \int_{-\infty}^{t} dt'\, R(\vec{r}, t-t')\ \sigma(\vec{r}, t')\right] \tag{3.2}$$

The response $Z_s(t)$ for Walsh figures is shown in Fig. 5 of Golomb *et al.* (1994).

Realizations of spike trains are created at random with inhomogeneous Poisson statistics, using the average response as the instantaneous rate. The probability density of obtaining a spike train $\Lambda_s(t)$, with k spikes at times $t_1 \ldots t_k$ during a measurement time T, is

$$P\left(\Lambda_s(t)\,|\sigma_s\right) = P\left(t_1 \ldots t_k|\sigma_s(\vec{r},t)\right) = \frac{1}{k!}\left[\prod_{i=1}^{k} Z_s(t_i)\right] \exp\left(-\int_0^T Z(t')\,dt'\right) \ . \tag{3.3}$$

A set of 1024 simulated responses for each of the 32 stimuli is used for testing the information calculation procedures. The asymptotic estimate of transmitted information is calculated using 10^6 trials per stimulus.

Response representation

The neuronal response to a stimulus as represented by the spike train is quantified by several variables. One is the number of spikes (NOS) in the response time interval, taken here to be 250 ms. The others are the projection of the spike train into the n PCs (Richmond and Optican 1987; Golomb *et al.*, 1994). We concentrate here on the first principal component (PC1) and on the first three principal components (PC123).

Information estimation

We describe here briefly the methods that we have used for estimating information from neuronal responses. We simulate an experiment in which a set of S stimuli is presented at random. Each stimulus is shown N_s times; here N_s is the same for all the stimuli. The total number of visual stimuli presented is $N = SN_s$. Different methods have been used to evaluate asymptotic ("true") results (calculated using 10^6 trials per stimulus), and to calculate information from small data samples:

Summation over the Poisson distribution. The asymptotic value of the transmitted information carried by the number of spikes NOS can be calculated directly by summing over the distribution. For each stimulus here, the number of spikes NOS is Poisson distributed with an average $\overline{\text{NOS}} = \int_0^T Z_s(t)dt$. The transmitted information (2.2) becomes

$$I(S;\text{NOS}) = -\sum_{\text{NOS}} P(\text{NOS})\log_2 P(\text{NOS}) + \frac{1}{N_s}\sum_s \sum_{\text{NOS}} P(\text{NOS}|s)\log_2 P(\text{NOS}|s) \quad (3.4)$$

This sum is discrete and is calculated using the Poisson probability distribution $P(\text{NOS}|s)$. The sum over NOS from 1 to ∞ is replaced by a sum from 1 to $\text{NOS}_{\text{max}} = 36$; taking a higher NOS_{max} has only a negligible effect on the result. Using this method the mutual information can be calculated exactly, but only when the firing rate distribution is known.

Straightforward binning (Golomb *et al.* 1994). This method is used to evaluate asymptotic values of information contained in principal components. Since the true underlying probabilities cannot be written explicitly, as in the NOS case, we perform the calculation by means of a simulation with a very large number of trials per stimulus (10^6). As we shall discuss, this allows us to keep negligible both the systematic errors due to finite sampling and regularization.

The principal components used here are calculated from the covariance matrix $C(t,t')$ formed over all responses in the set under study

$$C(t,t') = \frac{1}{SN_s}\sum_{s=1}^{S}\sum_{\mu=1}^{N_s}\left[\Lambda_{s,\mu}(t) - \bar{\Lambda}(t)\right]\left[\Lambda_{s,\mu}(t') - \bar{\Lambda}(t')\right] \quad , \quad (3.5)$$

where $\Lambda_{s,\mu}(t)$ is the μth realization of the response to the sth stimulus and $\bar{\Lambda}(t)$ is the average response over all the stimuli and realizations

$$\bar{\Lambda}(t) = \frac{1}{SN_s}\sum_{s=1}^{S}\sum_{\mu=1}^{N_s}\Lambda_{s,\mu}(t) \quad . \quad (3.6)$$

The eigenvalues of the matrix C are labeled according to a decreasing order; the corresponding eigenvectors are $\Phi_1(t), \Phi_2(t) \ldots$ The expansion coefficients of the neuronal response $\Lambda_{s,\mu}(t)$ are given by

$$a_{s,\mu,m} = \frac{1}{T}\int_0^T dt\, \Lambda_{s,\mu}\Phi_m(t) \tag{3.7}$$

Each response is then quantified using the coefficients of the first n principal components, and these are used as the response representation. The number n of coefficients used for quantifying the response is referred here as the code dimension. The maximal and minimal values for each component are found, and the interval between the minimum and the maximum of the mth component is divided into $R(m)$ bins. The mutual information is calculated from the discrete distribution obtained. The n-dimensional response space is therefore divided into $R = \prod_{m=1}^{n} R(m)$ n-dimensional bins. For PC123, we choose $R(1) = 36$, $R(2) = 20$, $R(3) = 20$. The mutual information carried by the first principal component only, PC1, is calculated in a similar way with $R(1) = 36$. The downward bias due to the finite binning is expected to be small when using such a high number of bins. For example, with PC1 and our simulated data set, using $R(1) = 36$ results in underestimating the mutual information by ~ 0.01 bit in comparison to $R(1) = 300$.

Binning with finite sampling correction.

Mutual information is calculated from the frequency table and then the bias correction is subtracted. The method was explained in detail in chapter 2, but we remark a few facts, just to explain how the technique can be adapted to the problem.

I. I. For each dimension, equipopulated bins are used. For a one-dimensional code (NOS, PC1), the bin-size varies across the response dimension, with non-equal spacing, so that each bin gets on average the same number of counts. For a three-dimensional code (PC123), the equipopulated binning is done for each dimension separately. The use of equipopulated bins is an attempt to minimize, keeping fixed the number of bins, the information loss due to discretization.

II. The choice of the number of bins $R(m)$ in each dimension for an experiment with S stimuli and N_s trials per stimuli remains somewhat arbitrary. Here we choose $R \sim N_s$, to be at the limit of the region where the correction procedure is expected to work, and thus still be able to control finite sampling, while minimizing the downward bias produced by binning into too few bins. For the number of spikes, NOS, however, each response is just an integer ranging from 0 to the maximal number of spikes (NOS_{max}, 25 for the parvocellular cell and 34 for the magnocellular cell), so even if we allocate more bins than this maximum, the extra ones will stay empty. For a multi-dimensional code (*e.g.*, PC123), we allocate a number of bins $R(m)$ in the m-th direction in relation to the amount of mutual information carried by this principal component alone, as shown in Table 3.1. When differences between different codes are calculated, we use the same numbers of bins in the relevant dimension. When PC1 and NOS are compared, we use the same number, R as for PC1; in this case, many bins for NOS stay empty. For comparing PC123 and NOS we use the same number of bins for NOS as for the first principal component, which is the richest in information among the three (*e.g.*, 8 for $N_s = 128$). In this way we compare quantities calculated in a homogeneous way.

Neural network. (Hertz *et al.* 1992; Kjaer *et al.* 1994)

A two-layer network is trained by back-propagation to classify the neuron's responses according to the stimuli that elicited them, as explained in more detail in section 2.1.3. The input to the network is the quantified output of the biological neuron: the spike count, the first n principal components or both. There is one output unit for each stimulus, and, after training, the value of output unit number s is an estimate $\hat{P}(s|\mathbf{r})$ of the conditional probability $P(s|\mathbf{r})$. The mutual information is then calculated from these estimates using the following formula:

$$I(S;R) = \left\langle \sum_s P(s|\mathbf{r}) \log_2 \left[\frac{P(s|\mathbf{r})}{P(s)} \right] \right\rangle_{\mathbf{r}} . \tag{3.8}$$

The average over the response distribution in (3.8) is estimated by sampling over randomly chosen data points.

N_s	16	32	64	128
R for NOS	16	$1 + \mathrm{NOS}_{max}$	$1 + \mathrm{NOS}_{max}$	$1 + \mathrm{NOS}_{max}$
R for PC1	16	36	63	128
$R(1) \times R(2) \times R(3)$ (PC123)	$4 \times 2 \times 2$	$6 \times 3 \times 2$	$7 \times 3 \times 3$	$8 \times 4 \times 4$

Table 3.1:
Number of bins R used for the various codes and numbers of trials N_s.

A.

$R = R(1) \times R(2) \times R(3)$	$7 \times 6 \times 3$	$8 \times 4 \times 4$	$10 \times 4 \times 3$	$15 \times 3 \times 3$
Parvocellular cell	0.145	0.135	0.125	0.114
Magnocellular cell	0.289	0.289	0.291	0.289

B.

hidden units, learning rate	6 , 0.0003	6, 0.001	10, 0.001	6 , 0.003
Parvocellular cell	0.121	0.127	0.123	0.118
Magnocellular cell	0.219	0.201	0.219	0.206

Table 3.2:
Difference in mutual information $I(S;\mathrm{PC123}) - I(S;\mathrm{NOS})$ (bits) for $N_s = 128$ and: A. Several binning schemes; B. Several network schemes. The standard deviations of the difference are about 0.01.

3.1.2 Results

We calculated the information carried about a set of 32 Walsh patterns by the simulated neuronal response quantified by the number of spikes $I(S;\text{NOS})$, the first principal component $I(S;\text{PC1})$, and the first 3 principal components $I(S;\text{PC123})$ (Fig. 3.1) The arrows at the right side of the panels in Fig. 3.1 represents the asymptotic values calculated from simple binning using 10^6 trials per stimulus (for the parvocellular cell: $I(S;\text{NOS}) = 0.857$, $I(S;PC1) = 1.003$ and $I(S;PC123) = 1.038$; for the magnocellular cell, $I(S;NOS) = 0.246$, $I(S;PC1) = 0.456$ and $I(S;PC123) = 0.535$). These figures show the estimated transmitted information for $N_s = 16, 32, 64, 128$, and for two sample cells: magnocellular and parvocellular. The parvocellular cell has sustained activity over an interval of 250 ms, whereas the magnocellular cell is active mostly over the first 100 ms. The magnocellular cell has more phasic responses (Golomb et al, 1994). Thus, we expect the multidimensional codes to capture a larger proportion of the information in its responses. A simple rate code is more likely to be an acceptable zeroth order description of the parvocellular cell. The results we obtain (compare panels A and C, D and F, respectively, in Fig. 3.1) bear this expectation out.

All of the calculations show that the first principal component is more informative about the stimulus than the spike count. As expected, this effect is especially strong for the magnocellular cell, as the first PC weighting function suppresses contributions from the spikes after the first 100 ms, which are mainly noise.

Fig. 3.1 shows that the raw binning is strongly biased upward (see above). The difference between the estimates made with raw and corrected binning almost does not vary with N_s, for PC1 and PC123. This is because, as discussed above, the first-order correction term (Eq. 2.6) is approximately proportional to R/N_s, which we choose to keep roughly constant in our calculation. As mentioned above, for NOS there is no point choosing R above $1+\text{NOS}_{max}$, hence the correction term, and with it the raw estimate, decreases as N_s is increased.

Both the corrected binning method and the network method tend to underestimate the information in PC1 and PC123 (the only counter-example is shown in Fig. 3.1E, where the binning method overestimated it). This is because both methods involve a regularization of the responses (explicit in the binning, and done implicitly by the network), and, as described above, a regularization always decreases the amount of information present in the raw response. For example, the corrected binning method underestimates the information whenever the number of bins is too small to capture important features of the probability distribution of the responses. Therefore, the effect is strong for PC123 when the number of bins in the direction of the first PC is not large enough, and the bias downward decreases with increasing N_s, because the number of bins increases too. The underestimation does not occur with NOS because the maximal number of spikes in our examples is around 30, and there is no meaning to using finer binning. In general, the underestimation due to regularization is more prominent for the higher-dimensional code (PC123) because the effect of adding more bins in each dimension is stronger when the number of bins is small.

In Fig. 3.2 we present the differences $I(S;\mathrm{PC1}) - I(S;\mathrm{NOS})$ between the information carried by PC1 and NOS, and $I(S;\mathrm{PC123}) - I(S;\mathrm{NOS})$, between PC123 and NOS. For all the cases considered here the network yields a value for the extra information that is biased downwards. This shows that the network automatically regularizes responses, and apparently the regularization is stronger for the higher-dimensional code. The corrected binning technique, on the other hand, gives both downward- and upward-biased values for the extra information, in this instance downwards for the parvocellular cell. However, results obtained with the simple corrected binning procedure appear in general more accurate, both for lower and higher dimensional codes.

Since the choice of number of bins along each dimension is somewhat arbitrary for a multi-dimensional code, we checked the effect of using different binning schemes for $N_s = 128$. The results are summarized in table 3.2A. The information differences for the parvocellular cells are in a 30% range; the information differences for the magnocellular cells are all nearly the same, no matter which method is used. Thus, even in the least favorable case, information differences using different binning schemes remain in the range of the remaining (downward) systematic error, about 30%. In a similar way, we varied the parameters of the network: the number of hidden units and the learning rate. The difference in information varies within 10% for both cells, indicating that changes in these parameters are less important than the downward bias due to the regularization. The test error for the various network parameters is quite similar, with differences within 0.4% for both cell types. Thus, it is difficult to determine the best result of the network just from this number.

3.1.3 Discussion

The results of the simulated experiment presented here illustrate well the problems occurring when estimating information from neuronal activity, and in particular they show that (i) information measures are specific to the stimulus set considered, but also they are specific to the quantity/ies chosen to quantify neuronal responses; (ii) information measures are affected by limited sampling, which results typically in an upward bias, but also, since it is always necessary to regularize continuous responses, they may be affected by the regularization, which results in a bias downward; (iii) the introduction of a technique which eliminates (or at least reduces) the finite sampling error leads to a choice of milder regularizations, and thus also to a decreasing of the relative downward bias.

When a simple binning of the responses was used, raw information measures were strongly biased upward, and thus it was necessary to apply a correction for limited sampling. If one follows this procedure, the only parameter that has to be set is R, the number of response bins, but results are strongly dependent on the choice of R. If R is chosen too large, subtracting the term C_1, Eq. (2.6). will not be enough to correct for limited sampling (see also Treves & Panzeri, 1995); while if it is chosen too small, a strong regularization will be imposed and information will be underestimated. The

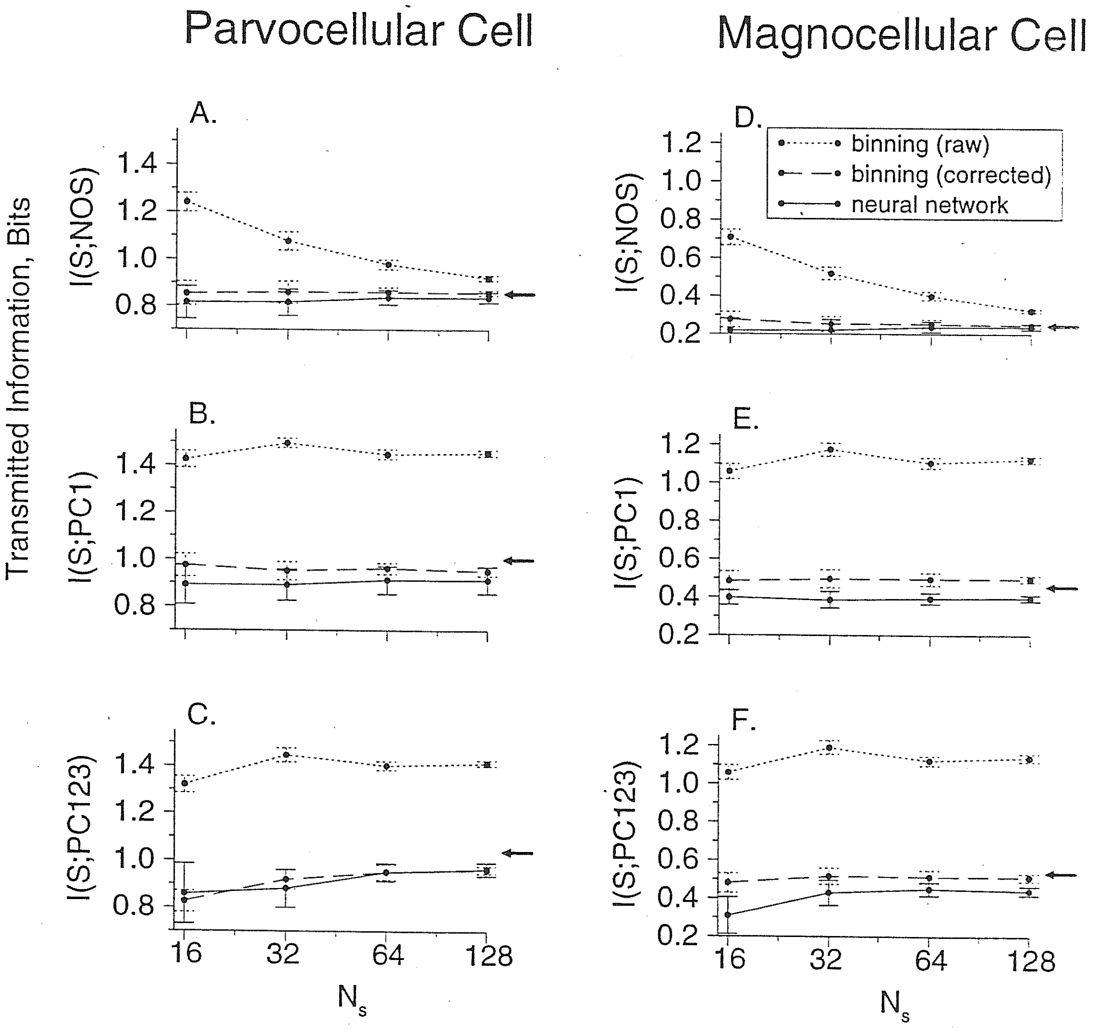


Figure 3.1:
The information about a stimuli set of 32 Walsh figures conveyed by the neuronal response of model parvocellular (A-C) and magnocellular (D-F) cells, as estimated by various methods. The response is quantified by the number of spikes (A,D), the first principal components (B,E) and the first three principal components (C,F). The mutual information is estimated by straightforward equipopulated binning (dotted lines), equipopulated binning with finite sampling correction (dashed lines) and a neural network (solid line). The numbers of bins for each code and N_s are shown in Table 3.1. The neural networks has 6 hidden units and a learning rate of 0.003. The arrows at the right indicate an asymptotic value (very good approximation for the "true" value) obtained with equi-spaced binning with $36 \times 20 \times 20$ bins and 10^6 trials per stimulus.

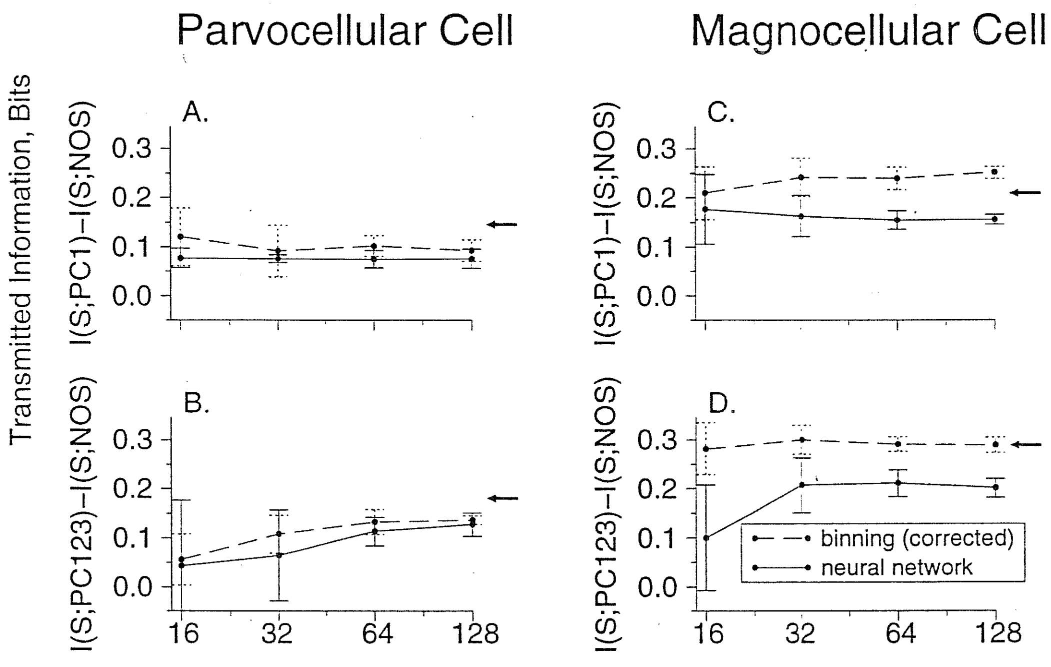


Figure 3.2:

Differences between measures used for quantifying the response of a parvocellular cell (A,B) and a magnocellular cell (C,D). We show in A and C the difference between the information carried by the the first PC and that carried by the number of spikes, and in B and D the difference between the information in the first three PC's and that in the number of spikes. Dashed lines indicate results obtained from normalized equipopulated binning; solid lines indicate those from the neural network.

In (A,C) the number of bins is 16 for $N = 16$ and 36 for $N \geq 32$. In (B,D) the number of bins in the first three principal components is as given in Table 3.1, and the number of bins for NOS is equal to R(1) in that table, *i.e.*, to the number of bins for the first PC (4,6,7 and 8 respectively). The arrows at the right indicate an asymptotic value obtained with equi-spaced binning with 36 x 20 x 20 bins and 10^6 trials per stimulus.

present results indicate that it makes sense to set the number of response bins at roughly the number of trials per stimulus available, $R \simeq N_s$. C_1 is inversely proportional to the number of trials available and, roughly speaking, directly proportional to the number of response bins, and this choice approximately balances the upward bias due to finite sampling with the downward one due to the regularization. Choosing the number of bins for each of the first three principal components in PC123 was more delicate that for NOS or PC1, and tended to yield a stronger downward bias in information values. This suggests that the use of the binning procedure alone becomes insufficient for higher dimensional codes, when it could be impossible to distribute among the various dimensions the R allowed bins in a meaningful way. One useful procedure for the computation of information contained in high dimensional codes is to use *decoding* to extract the relative probabilities of the stimuli from the responses and thus to reduce the original set of responses to the size of the stimulus set (Gochin *et al*, 1994, Rolls *et al*, 1996d). Decoding procedures based on stimulus reconstruction will be discussed in section 3.3.

We have shown that the amount of information carried in the firing of a single neuron during a certain time interval about a set of stimuli, and quantified by a certain code, can be estimated with a reasonable accuracy (within 10% if the dimensionality of the code is 1-3), even using a simple binning and correcting procedure. It is evident that the introduction of an analytical correction for finite sampling makes the predictions more reliable by about an order of magnitude, when compared with very empirical (Optican and Richmond 1986; Optican *et al.* 1991) statistical techniques.

Considering the results obtained with the network, we note that for all the cases considered, the network *underestimated* the mutual information, and also the extra information in the temporal response in comparison to the number of spikes. This indicates that the regularization induced by the network is enough to dispose of the finite sampling bias, at the price of underestimating information values, especially for higher dimensional codes, which are more strongly regularized. Information values generally increase weakly with N, which indicates that the regularization induced has decreasing effects as N becomes large. Since the underestimation is stronger for $I(S;\text{PC123})$ than in the case of unidimensional codes, estimates of the extra information in the second and third principal components (Fig. 3.2) are also biased downward. Several parameters need to be set when the network is used. Some (e.g. the number of iterations) can be set by cross-validation. Others (e.g. the learning rate) have little effect on the results across a broad range of values, as indicated in Table 3.2. An interesting aspect of the network procedure is that it effectively incorporates a decoding step, and as such can be immediately applied to high dimensional, e.g. multiple single-unit, data. Nevertheless, as we shall see in sec. 3.3, better estimates can be achieved with milder (and more transparent) decoding procedures, coupled with finite sampling corrections. These decoding procedures are expected to give, unlike the neural network, estimates of increasing precision as the number of cells in the sample increases.

In this section, principal components are used for quantifying the data with a low-dimensional code, because the first principal components carry most of the difference among the responses to different stimuli (Golomb *et al.* 1994). However, principal component analysis is nonessential to our procedure for handling finite sampling problems: any kind of n-dimensional response extracted from the neuronal firing patterns can be used, *e.g.*, PC123s, or the firing rate vector of a population of neurons. Our procedures of finite-sampling corrections were demonstrated here on stimuli with a sharp onset in time. They are applicable, however, also to continuously changing stimuli (after a suitable discretization), as long as the response to each stimulus is measured during a fixed time interval T, and stimuli are either discrete or have been discretized.

The network procedure needs a long computational time. A typical calculation for $N_s = 128$ and 32 stimuli runs for about 7 CPU hours on an SGI-ONYX computer. The binning-and-correcting technique is much faster, and most of the CPU time is taken up by sorting responses in order to construct equipopulated bins. For $I(S;\mathrm{NOS})$, which involves no sorting, a calculation with $N_s = 128$ runs for less than 1.8 seconds on an HP-Apollo computer.

3.2 The kinetics of the information conveyed by firing rates

Although temporal modulation can be relevant in certain situations, there is evidence that the rate at which it emits spikes is, for a neuron, an important way of coding information (*e.g.*, Tovée *et al.* 1993; Rolls *et al.* 1996a). Moreover, most of the neural network models introduced to understand some of the brain functions, are based on firing rates (Amit 1989). The aim of this section is to discuss how a quantitative experimental analysis of information contained in the neuronal firing rates, which can be performed very accurately for single cells, can help in using neural network models at a more quantitative level.

A most prominent correlate of the average amount of information in the firing rate of a neuron is the sparseness of the distribution of mean rates to each of the stimuli (Skaggs *et al.* 1992). Sparse firing, with only a small fraction of the stimuli evoking substantial responses (an example being place-related firing in the rat hippocampus), carries little information, whereas a more even use of its own firing range allows the neuron to transmit more information (as e.g. in the monkey temporal visual cortex, Rolls and Tovée (1995b)). Relating information and sparseness is made particularly interesting by the fact that the effects of sparseness have been analyzed in several models, for example in associative memory networks, in which sparse coding, while reducing the information content of each stored pattern, increases the storage capacity of the system (Tsodyks and Feigel'man 1988; Treves and Rolls 1991).

At fixed sparseness, there are at least two more aspects of the firing which are important in determining how informative it is. The first is how variable, or noisy, are responses to the same stimulus. The second is how close the distribution of mean rates is to being binary or bimodal, or conversely how graded is the response of the cell. Both aspects are often neglected in the construction and analysis of theoretical models, making the correspondence between such models and real neurons less direct. Many simple models, and a lot of common sense intuition, are based on noiseless binary variables. Deviations from this idealized case have opposite effects: noise always reduces information transmission, whereas a more graded response enhances it. The net effect depends on the time the neuronal activity is sampled for. Here we discuss how to analyze the firing of real cells, on different time scales, in these terms. In particular, we introduce a simple formula for information transmission rates valid for short recording times, and relate it to the sparseness of the neuronal representation. Further, we provide examples, from single cells recorded in the rat somatosensory system, in the primate temporal visual cortex, and in the primate hippocampus, of the role of noise and of the graded nature of responses in single unit information processing. At the end we discuss how to extrapolate these results to the multiunit case.

We believe that a systematic analysis of neuronal activity done in these terms could provide useful insights for quantitative models.

3.2.1 Time-derivatives of the information.

The neuronal responses considered here are the rates, r, recorded from a cell in correspondence with a given stimulus, s. Rates are measured simply by counting the number of spikes in a given time window $[t_0, t_0 + t]$, and hence are just positive integers, (except they are divided by t itself). The specific information about each stimulus, and the average transmitted information, can be written, according to the notations of chapter 2, as a function of response probabilities and of the recording time t:

$$I(s,t) = \sum_r p(r|s) \log_2[p(r|s)/p(r)]. \tag{3.9}$$

$$I(t) = \sum_{s\in S} \sum_r p(s,r) \log_2 \frac{p(s,r)}{p(s)p(r)} \tag{3.10}$$

In the case of spike count, which is a discrete response variable, the natural regularization, when measuring (3.9,3.10) on the basis of N events, is the discretization into R response bins. When rates are computed from a time window short enough that the maximal number of spikes recorded is not too high, the responses are already binned, no regularization is necessary, and, provided finite sampling effects can be controlled, one measures in fact the "true" underlying information. In particular, as the window shrinks to zero, the number of bins eventually reduces to just two (one spike or none), which implies that a) responses are binary and b) naive estimates can be easily corrected even with just a few trials per stimulus, subtracting a small C_1 (2.6) correction.

To study the initial rate at which information accumulates from time t_0, one can also consider directly its time-derivatives at t_0, which can be calculated by approximating

$I(s,t)$ by the Taylor expansion

$$I(s,t) = t\, I_t(s) + \frac{t^2}{2}\, I_{tt}(s) + \ldots \tag{3.11}$$

where $I_t(s), I_{tt}(s)$ are the first two time-derivatives of $I(s,t)$ calculated at t_0. The first derivative[1] is universal, i.e. independent of firing statistics, while the second takes a very simple expression under the assumption that the firing of the cell is purely Poissonian (note that it can be singular, instead, in other cases). To first order in t, if the spike train of the cell is a stationary random variable, so that the expected firing rates $r(s)$ are well defined, the probability $p(n|s)$ of emitting n spikes in the time window is determined solely by the mean rate r_s to each stimulus s. The latter statement is also true to the second order in t, but only with the further assumption of Poisson statistic. To second order in t we have

$$\begin{aligned} p(0|s) &\simeq 1 - tr_s + \frac{(tr_s)^2}{2} \\ p(1|s) &\simeq tr_s(1 - tr_s) \\ p(2|s) &\simeq \frac{(tr_s)^2}{2} \\ p(n > 2|s) &\simeq 0. \end{aligned} \tag{3.12}$$

Denoting with $\bar{r} = \sum_s p(s) r_s$ the grand mean rate to all stimuli, and with

$$a = \bar{r}^2 / \sum_s p(s) r_s^2 \tag{3.13}$$

the sparseness (Treves and Rolls 1991) of the rate distribution, we get

$$I_t(s) = r_s \log_2 \frac{r_s}{\bar{r}} + \frac{\bar{r} - r_s}{\ln 2} \tag{3.14}$$

and for Poisson statistics

$$I_{tt}^{Pois}(s) = r_s^2 \log_2 a + \frac{\bar{r}(2r_s - \bar{r})(1-a)}{a \ln 2}. \tag{3.15}$$

It can be easily seen that $I_t(s) \geq 0$, while $I_{tt}^{Pois}(s)$ can be positive or negative (but $I_{tt}^{Pois} \equiv (\bar{r}^2/a \ln 2)[\ln a + (1-a)] < 0$, implying that for Poisson statistics the rate of information transmission always slows down after the first spike).

The simple formulas for $I_t(s)$ and $I_{tt}^{Pois}(s)$ are remarkable because they require a measure of only the mean rates r_s, and not of the full distribution of rates to each

[1] Since only two spiking events (zero or one spike) are relevant to the first order, expressions (3.14,3.16) are in fact the first derivatives of the information carried by the full spike train, and not only from the rates.

stimulus $p(r|s)$. This translates into a clear advantage for measuring information, when data is scarce [2].

Several interesting relationships should be appreciated (Panzeri *et al.* 1996a). First of all, $I_t(s)$ itself represents, apart from a rescaling by the overall mean rate, a universal U-shaped curve which gives, whatever the rate distribution, the initial speed of information acquisition as a function of the rate. In recordings of primate temporal cortical cells, it was often found that the dependence of $I(s,t)$ on r_s for finite t (50 to 500ms) closely reproduced that predicted at $t \rightarrow 0$, i.e. that associated with the time-derivative $I_t(s)$ (Rolls *et al.* 1996c).

Dividing (3.14) by the overall mean rate and taking an average across stimuli one has

$$\Phi = \sum_s P(s)\frac{r_s}{\bar{r}}\log_2\frac{r_s}{\bar{r}}, \tag{3.16}$$

which has the meaning of *mean information per spike*. It is easy to show that in general. for any distribution of rates,

$$0 < \Phi < \log_2(1/a). \tag{3.17}$$

and for the distributions of rates that are close to binary, with one of the peaks at zero. $\Phi \approx \log_2(1/a)$, while if they are nearly uniform, or strongly unimodal, $\Phi \ll \log_2(1/a)$.

Extensive recordings of rat hippocampal and neocortical cell activity (Skaggs and McNaughton 1992; Skaggs *et al.* 1993), indicate that Φ and a (or equivalently $\log_2(1/a)$) could be used almost interchangeably to characterize firing rates distributions: one parameter turns out to be an excellent predictor of the other. As an example, we plot in Fig. 3.3 the information per spike Φ carried by a set of 20 different cells, recorded in the primate hippocampus, in response to a set of 16 different views[3], versus the sparseness of the cell, calculated from firing rates sampled in time windows 25 msec long. It can be noted that sparseness and information per spike are almost in one-to-one correspondence. and that, since these cells seldom fire more than one spike in 25 msec, $\Phi \approx \log_2(1/a)$.

We note that when S stimuli are presented each with equal frequency, the information per spike Φ is simply related to the breadth of tuning

$$H = -\frac{1}{\log_2 S}\sum_s \frac{r_s}{S\bar{r}}\log_2\frac{r_s}{S\bar{r}} \tag{3.18}$$

(Smith and Travers 1979), which was used to characterize the distribution of responses e.g. to gustatory stimuli in the monkey (Rolls *et al.* 1990). In fact

$$\Phi = (1-H)\log_2 S \tag{3.19}$$

and when the cell respond to only one stimulus, $H = 0$ and $\Phi = \log_2 S$ (extreme selectivity), whereas when it responds equally to all stimuli $H = 1$ (broad tuning) and $\Phi = 0$ (no information).

[2] A small systematic error occurring when calculating (3.14,3.16) from a finite sample is anyway present, but it can be easily calculated with standard error propagation

[3] Details of the recording procedure are reported in section 3.3.5

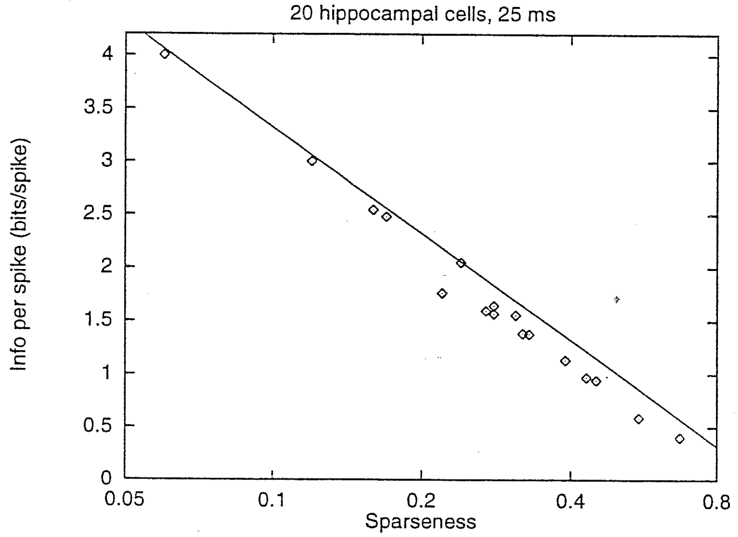


Figure 3.3:

Information per spike (◇) versus sparseness, calculated from the mean firing rates (sampled in time windows 25 msec long) of 20 primate hippocampal cells. The stimulus set consists of 16 different views. See chapter 3.3.5 for details. The solid line is $\Phi = \log_2(1/a)$. The sparseness axis is on a logarithmic scale.

3.2.2 Real responses and their idealization: the role of noise and grading

The figures 3.4 and 3.5 provide examples of the way the firing rate of real cells, when measured over increasing time, conveys information about those stimuli. The information in the actual rates is compared with the information present in binarized responses, and with that available from an ideal binary unit (Panzeri *et al.* 1996a).

Responses are binarized by taking as '1' all responses above a certain threshold, and as '0' all others, with the threshold chosen, in each window, to optimize the amount of information transmitted. Note that this binarization preserves at least part of the original trial-to-trial variability, but results in an apparent sparseness different from the true value.

Ideal binary responses are simply those of a unit operating at the same grand mean rate and sparseness as the real unit (in each window), but with zero noise, i.e. mean rates as well as the rates on individual trials are taken to be zero for a fraction $(1 - a)$ of the stimuli, and $\bar{r}/a$ for the remaining fraction a. Thus the information conveyed by this ideal binary unit is just:

$$I_{id} = -a \log_2(a) - (1 - a) \log_2(1 - a) \tag{3.20}$$

In addition, the time-derivative I_t is shown, as calculated for actual responses from the shortest window considered (2 msec for the two cortical cells in Fig. 3.4; 25 msec

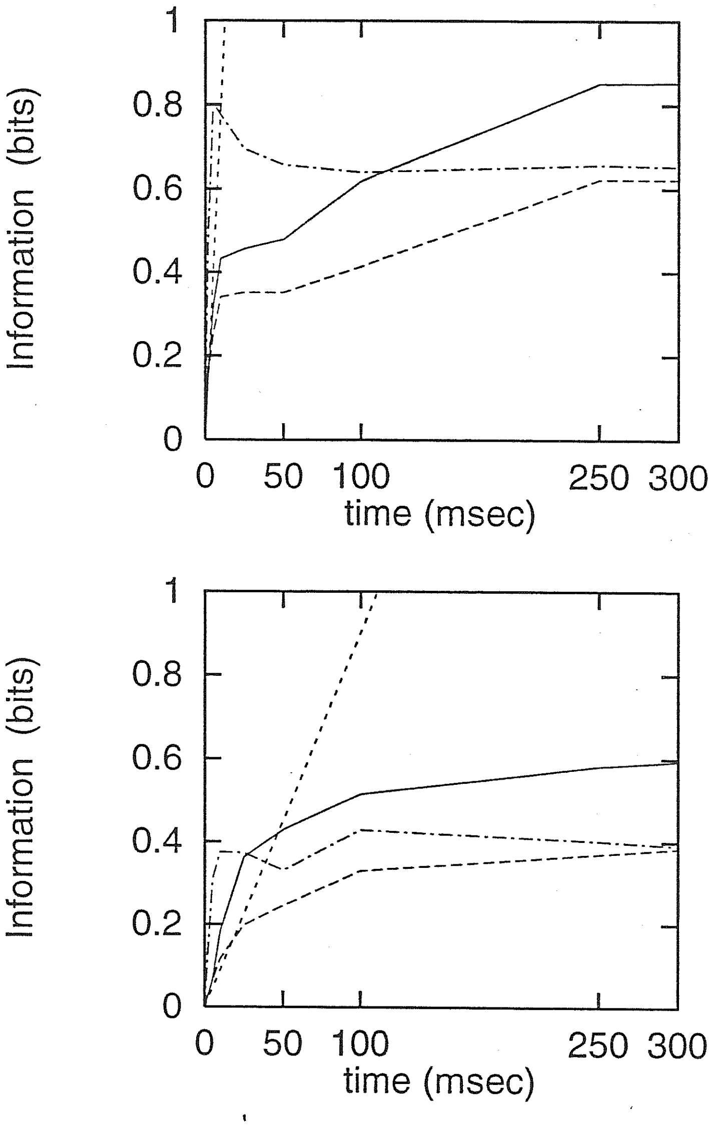


Figure 3.4:

(Top): Information in the number of spikes emitted in response to 4 electrical stimuli (——) by a cell in the rat SI cortex, in the binarized responses (- - -) and in the noiseless responses of an ideal binary unit (— · —). The initial slope at t_0 is also indicated (– — — —). (Bottom) Information in the responses to 20 face stimuli by a cell in the monkey IT cortex, with the same notation.

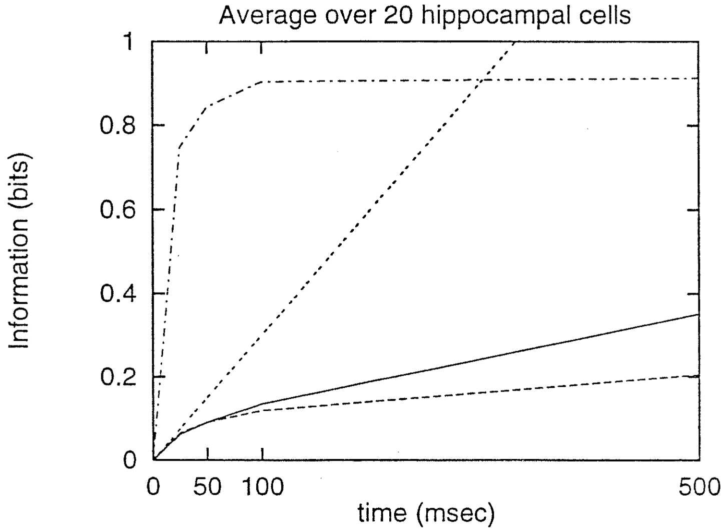


Figure 3.5:
Average over population of the information in the number of spikes emitted in response to 16 different views (——) by 20 hippocampal cells, in the binarized responses (– – –) and in the noiseless responses of an ideal binary unit (— · —). The initial slope of the information processing is also indicated (– – – –).

for the hippocampal cells in Fig. 3.5)). For binarized responses the derivative is the same, because for very short windows the responses are already binary; whereas for ideal binary units the derivative is higher, as it is clear from the figure and Eq. 3.17. Note, though, that for the cell of Fig. 3.4b the time derivative was almost constant, even when computed from the distribution of mean rates over longer windows (whereas for that of Figures 3.4a and 3.5 it progressively decreased).

Figure (3.4, top) refers to a cell in the rat somatosensory cortex, responding to electrical stimulation of 4 different intensities. The information in the actual rates rises steeply from t_0 (15 msec post stimulus onset) and then slows down and saturates when t is of order a few hundred *msec*. The initial rate of transmission is very high, about 80 bits/sec. Binarized responses convey less information, but even for long windows only by a factor of about 3/4. An ideal binary unit with the same sparseness would yield almost instantly all the information it can convey. This levels off around 0.6 bits, which is above the value for the real cell, thanks to the lack of variability, at short times, $t \leq 100$ msec; but it is inferior for longer times, when the binary output becomes limiting if contrasted with actual graded output.

Figure (3.4, bottom) refers to a cell in the primate visual cortex, responding to 20 face stimuli. t_0 is 100 msec post stimulus onset, near the peak of the response. The information in the actual rates accelerates almost instantly from a lower initial slope $I_t \approx 10$bits/sec, and slows down only later, having reached values twice above those for the binarized responses. In this case the time-derivative provides a poorer indication of the information available in the full response. The positive second derivative at t_0 reflects a limited trial-to-trial variability in this cell, much less than for Poisson statistics. This may be related to the operation of recurrent circuits. The ideal binary unit with the same sparseness would again be more informative at short times, but now the time, after which the advantages of a graded response take over, is shorter, $t \approx 25$ msec.

Figure 3.5 refers to the information carried (on average) by a set of 20 primate hippocampal 'view' cells (Rolls *et al.* 1995a), responding in relation to where the monkey is looking in space. The information about the spatial environment has been quantified by dividing the space into 16 discrete 'stimuli'. During recording the monkey was freely behaving, and data was collected by taking a fixed length of record whenever the eyes were still looking at a given part of the environment (more details are reported in section 3.3.5). In this case thus t_0 is not well defined. We see that for the set of hippocampal cells under analysis the initial rate of information processing is much lower ($\approx$ 3 bits/sec) than that of cortical cells in Fig. 3.4, and the characteristic time scale of information processing appears also slower. Binarized responses convey less information, but only by a negligible factor until $t \approx 100$ msec, and by a factor of about 30% for longer periods. An ideal binary unit with the same sparseness would yield almost instantly all the information it can convey, which is much higher than the information carried by real responses of the cells. It is important to note that these results for the hippocampal cells could be affected by the crude discretization of the environment into 16 parts. However, that discretization was the finest one compatible with the size of the data set.

3.2.3 Time scales and the role of grading in the information conveyed by a network

At this point, an important question is how to generalize these results, valid in the single cell case, to the information processing in a whole network of neurons. In the latter case, as it will be clarified in sec. 3.3, it is not possible to experimentally address the problem with the same degree of precision available in the single cell case. Anyway, a few considerations can shed light on this issue.

Let us imagine to have an ensemble of C cells, with a mean response population vector $\vec{r}_s$ during the presentation of a stimulus s ($\vec{r}_s$ is a vector with one element (or component) for each of the C cells considered; these components are labeled as $r_{s;c}$, $c = 1, \cdots, C$), and let us suppose that the cells fire *independently.*

In this case, the only events with non-zero probability to the first order in t are:

$$\begin{aligned} p(\vec{0}|s) &= 1 - t \sum_c r_{s;c} \\ p(\vec{e}_c|s) &= t\, r_{s;c} \ , \end{aligned} \tag{3.21}$$

where $\vec{e}_c$ is the response vector with one spike in the c-th cell component and zero in the other ones. From eq. (3.21) we can immediately calculate the first derivative of the information carried by the population:

$$I_t(s) = \sum_{c=1}^{C} I_t^c(s) \; , \tag{3.22}$$

where $I_t^c(s)$ are the time derivative of each of the C single cells, as obtained from eq. (3.14).

From eq. (3.22), it appears clear that, to the extent that cells are really independent, the characteristic time scale for information processing in a population is just C times smaller than the average time scale for single cells. Therefore we can conclude that, if the size of the population is big enough, and the correlations among cells are small enough, most of the information carryed by the network can be already extracted from time windows so short that response of individual cells are all binary (zero or one spike).

3.3 Information available in the responses of an ensemble of neurons

The informational properties of single cells are a building block for a quantitative understanding of brain functions. Nevertheless, for many functions controlled by the brain or variables represented in the brain, the relevant unit is the neural population rather than individual cells (Georgopoulos *et al.* 1986; Georgopoulos *et al.* 1993; Abbott *et al.* 1996; Rolls *et al.* 1996d).

In fact, many cells, in a given cortical area, are found to code any given feature, with relative broad tuning. The way that the network sharpens the tuning curve and reduces noise depends crucially on the nature of the neural code. In particular, representational capacity is extremely sensitive to how independently messages are distributed across a population of coding neurons. If each stimulus is represented by the response of a single unit ("grandmother cell"; Barlow 1961), the number of stimuli that can be represented grows linearly, and the information about a given set of stimuli logarithmically, with the number of coding cells. Whereas, if the mutual information about stimuli is distributed across the full network, the number of stimuli that can be represented grows exponentially, and the information linearly, with the number of coding neurons. Of course intermediate and redundant strategies are possible.

The properties of single cell responses are not enough to establish the existence of truly distributed representations. Distributed coding requires in fact that each neuron has a distinctive response profile across stimuli (in order to minimize redundancy), and that the differences in the broadly tuned responses are not masked by their trial-to-trial variability. Experimental measures of how information that can be extracted from a population depends on the number of coding neurons are thus necessary to clarify the nature of neuronal representation.

The problem of extracting information from a large population is difficult because of the high dimensionality of the response space (at least one dimension per cell). The binning procedure alone becomes insufficient for high-dimensional codes. In this case a *decoding* procedure can be used to extract the relative probabilities of the stimuli from the responses, and thus to reduce the original set of responses to the size of the stimulus set. This is a drastic reduction, but is appropriate because the minimum number of response bins that may not throw away information, if the appropriate code is used, is just the same as the number of stimuli, $R = S$. In addition, one can optimize the choice of decoding, trying to maximize the extraction of information, or can model the decoding to neurophysiologically plausible algorithms that can be used by the downstream network to read off information.

In this section we discuss [4] how to extract information present in responses of population of neurons, which can be recorded simultaneously, or sequentially. We introduce also the quantities relevant for the characterization of the ensemble performance, namely the percentage of correct decoding and two different information quantities reflecting different aspects of the quality of the encoding. In the discussion we address questions at the level of firing rates, but the procedure is easily generalizable to other codes. The section is organized as follows. First we review the basic technical steps of decoding: cross-validation, definition of information in terms of probabilities of correct decoding, and the algorithms for estimation of response likelihood. Then we test this decoding method by means of analytical considerations and numerical analysis. Finally, we provide explicit examples of neuronal data analysis which can be performed in these terms.

3.3.1 Decoding and cross-validation procedures

In estimating the information carried by the responses of several cells the basis of the analysis discussed here, is the construction of population response vectors $\vec{r}_s$. occurring. during the presentation of a stimulus s, in what are labeled as *test* trials ($\vec{r}_s$ is a vector with one element (or component) for each of the C cells considered; these components are denoted as $r_{s;c}, c = 1, \cdots, C$). If the population is recorded simultaneously, the vectors should be constructed by using measures of cell activities (e.g. discharge rate within a given time window) recorded at the same time. Otherwise, if the population is recorded sequentially, only pseudosimultaneous response vectors can be constructed. Each response vector is compared to the mean population response vector to each stimulus, as derived from a different set of *training* or reference data, in order to estimate, by means of one of several decoding algorithms, as described below, the relative likelihoods for each (s') of the possible stimuli to be the current one, $p(s'|\vec{r}_s)$. Summing over different test trial responses to the same stimulus s, one could extract the probability that by presenting stimulus s the neuronal response would be interpreted as having been elicited by stimulus s',

$$p(s'|s) = \sum_{\vec{r}_s} p(s'|\vec{r}_s) P(\vec{r}_s|s) \ , \tag{3.23}$$

[4] The procedure we discuss here has been mainly developed by Treves (Treves 1996) and Rolls (Rolls *et al.* 1996d).

and from that one obtains the resulting measures of percent correct identification and of the information decoded from the responses. Separating the test from the training data is a frequently used procedure in parametric statistics, called cross-validation, and is needed to prevent overfitting. In the analysis presented here we have performed cross-validation as follows. One of the available trials for each stimulus is used for testing, and the remaining trials for training. The resulting probabilities that s is decoded as s' are however averaged over all choices of test trials, thus alleviating finite sampling problems. This cross-validation procedure is particularly convenient when the number of trials per stimulus N_s is very low. When N_s is higher, one could use more than one test trial (*e.g.*, a given fraction).

3.3.2 Information extraction from stimulus reconstruction

Having estimated the relative probabilities that the test trial response had been elicited by any one stimulus, the stimulus $s' = s_p$ for which this likelihood is maximal can be said to be the stimulus *predicted* on the basis of the response. In general s_p will not coincide with the true s and the accuracy in the decoding can be quantified by the percentage of correct decoding (or the corresponding fraction f_{cor}), or alternatively by the mutual information in the joint probability table $q(s, s_p)$,

$$I_{ml} = \sum_{s,s_p \in \mathcal{S}} q(s, s_p) \log_2 \frac{q(s, s_p)}{p(s)q(s_p)} , \tag{3.24}$$

where $q(s, s_p)$ is constructed from the fraction of times an actual stimulus s elicited a (test) response that led to a predicted (most likely) stimulus s_p. Thus I_{ml} measures the information in the predictions based on *maximum likelihood*, and as such it does not only reflect, like percent correct, the number of times the decoding is exact, but also, beyond percent correct, the distribution of wrong decodings. A further quantity is the mutual information

$$I_p = \sum_{s,s' \in \mathcal{S}} p(s, s') \log_2 \frac{p(s, s')}{p(s)p(s')} \tag{3.25}$$

obtained from the *probability* $p(s'|s)$ of confusing s with s', which is given by averaging $p(s'|\vec{r}_s)$ over the responses to s, eq (3.23).[5] This second information measure reflects, unlike the first, also the degree of certainty with which each single trial has been decoded, and it thus sheds light on a further aspect of the quality attained in decoding. Both information quantities suffer from limited sampling distortions, but the second much less than the first, in the sense that, with the limited sampling correction procedures we have

[5] The difference between the table $q(s, s_p)$ and $p(s, s')$ can be appreciated by noting that each vector comprising a given trial contributes to p a set of numbers (one for each possible s') whose sum is 1, while to q it contributes a single 1 for s_p and zeroes for all other stimuli. (Obviously each contribution is normalized by dividing, in both cases, by the total number of test trials available.) As a consequence, I_{ml} must be corrected with the correction term corresponding to the 'quantized' case, eq. (2.6), whereas I_p must be corrected with the term derived for the 'smoothed' case, eq (2.21).

developed, I_p can be estimated accurately even with few trials per stimulus, while I_{ml} requires more trials.

Several different, and simple, decoding algorithms can be written for estimating from the recorded response the likelihood of each stimulus. The goal of decoding is to reconstruct the correct Bayesian probabilities from the data, extracting from the data itself as much information as is possible. Since the true *a priori* probabilities are unknown, a good strategy could be to do an extensive analysis by means of a number of different, and plausible, decoding algorithms, and then choose the "optimal" decoding, which gives, after finite sampling corrections, the best estimate of the unregularized values of information and percent correct. In other words, the strategy consists in finding, among a set of possible algorithms, the decoding that *minimizes* the downward bias due to regularization. Another very interesting approach is based on the idea of emulating the processing that could be performed by neurons receiving the output of the neuronal population recorded, thus extracting that portion of the information theoretically available that could be extracted with simple neurophysiologically plausible operations by receiving neurons (Rolls *et al.* 1996d).

3.3.3 Algorithms for likelihood estimation

Now we describe the algorithm that extract $p(s'|\vec{r}_s)$ from an estimate of the probability $P(\vec{r}_s, s')$ of a stimulus-response pair, by normalizing so that $\sum_{s'} P(s'|\vec{r}_s) = 1$. When neurons in the sample were recorded one at a time, trial to trial fluctuations are unlikely to be correlated between different neurons, and it is meaningful to write $P(\vec{r}_s, s')$ as a product of probabilities of individual neurons. In this case, the probability $P(\vec{r}_s, s')$ can be estimated for this purpose as:

$$P(\vec{r}_s, s') = p(s') \prod_c P(r_{s;c}|s') \ , \tag{3.26}$$

and finally, $P(r_{s;c}|s')$ is derived from the responses of cell c in the training trials. Otherwise, when neurons were recorded simultaneously, the independence assumption (3.26) has to be verified *a posteriori* by analyzing the data. If the within trial correlations among different cells are strong, the decoding (3.26) may fail, destroying correlations and leading to larger systematic errors. Therefore in the case of strong trial to trial correlation among the cells, we recommend to use decoding for the estimate of $P(\vec{r}_s, s')$, not merely based on single cells probabilities like (3.26) (for example, one can use the algorithm based on the distance between test and average response vectors, eq. (3.28)).

Different decoding procedures differ in the estimate of $P(\vec{r}_s|s')$. Here we describe the algorithms that we have used in the analysis reported in the next subsections.

Gaussian fitting

The single cell probabilities $P(r_{s;c}|s')$ are fitted with a Gaussian distribution whose amplitude at $r_{s;c}$ gives $P(r_{s;c}|s')$, except when $r_{s;c} = 0$. In this case, if there are training trials with zero firing, $P(0|s')$ is estimated as the fraction of training trials yielding zero firing. Otherwise, $P(0|s)$ is calculated integrating the negative tail of the Gaussian distribution.

Poisson-like fitting

The probability to have n spikes in a given time window is fitted, for every cell, to the following formula:

$$P(r_{s;c} = n|s') = \alpha_{s';c}\delta_{n,0} + (1 - \alpha_{s';c})\frac{(< r >_{s';c})^n \exp(- < r >_{s';c})}{n!} \tag{3.27}$$

where $\alpha_{s';c}$ and $< r >_{s';c}$ are, respectively, the fraction of trials with zero firing and the average number of spikes for a given cell c and stimulus s'. The deviation from the pure Poisson distribution is introduced because usually the number of trials with zero spikes was more than that predicted by the Poisson law. (The same applies to the Gaussian fit described above).

Distance between test and average response vectors

This algorithm computes the probabilities $P(\vec{r}_s|s')$ by means of the euclidean distance between the test response vector $\vec{r}_s$ and the average response training vectors $< \vec{r} >_{s'}$ to each stimulus s'. This distance is divided by a measure σ of the variability within responses (*e.g.,* the average, over the population, of the standard deviation of cell's responses, as calculated from training trials) and then exponentiated (with negative sign). The resulting formula is then:

$$P(\vec{r}_s|s') \propto \exp - \left(\frac{|\vec{r}_s - < \vec{r} >_{s'}|^2}{2\sigma^2} \right) \tag{3.28}$$

The most likely stimulus s' is that whose mean response vector $< \vec{r} >_{s'}$ is closest (in the euclidean sense) to the actual one. It is important to note that this algorithm is not based on the independency assumption (3.26).

Dot product between test and average response vectors

To understand how much of the information present in neuronal responses can be read off by the brain , it is useful to compare the amount of information extracted with the 'optimal' procedure with the information that is extracted by an algorithm that could be easily implemented by neurons receiving the population's output.

For this purpose, an algorithm called Dot Product (DP) decoding has been introduced (Rolls *et al.* 1996d). The DP algorithm computes the normalized dot products between the current firing vector $\vec{r}_s$ on a test trial and each of the mean firing rate response vectors in the training trials for each stimulus s'. (The normalized dot product is the dot or inner product of two vectors divided by the product of the length of each vector.) The highest dot product indicates the most likely stimulus that was presented, and this is taken as the best guess for the percentage correct measures.

For the calculation of I_p, eq. (3.25), it is desirable to have a graded set of probabilities for which of the different stimuli was shown, and these were obtained from the dot products as follows. The S dot product values were cut at a threshold equal to their own mean plus one standard deviation, and the remaining non-zero ones were normalized to sum to 1. It is clear that in this case each operation could be performed by an elementary neuronal circuit (the dot product by a weighted sum of excitatory inputs, the thresholding by activity-dependent inhibitory subtraction, and the normalization by divisive inhibition). The resulting relative probabilities are cruder estimates than those obtained with the algorithms designed to optimize information extraction, and a precise quantitative assessment of the price paid for using a simpler and neurophysiologically plausible algorithm can be derived from a comparison of the amounts of information obtained in both cases.

3.3.4 Tests of the accuracy of decoding procedures

In this subsection we test the effectiveness of the decoding procedure presented before, both by using simulated data and by calculating, in the limit of short processing time, analytical estimations of the information quantities (3.24,3.25).

Simulation results

We use here the response probability distributions obtained by a model of the response of parvocellular LGN cells to a set of 32 visual stimuli, and discussed in section 3.1.1. What is interesting here is to compare the asymptotic value of the information, carried by the underlying probabilities, to the value one can obtain, through the decoding procedure and after bias subtraction, from a finite number of samples. Moreover, we test the accuracy of information estimation of the two different information quantities that one can define in terms of the decoded probabilities (3.23): the one extracted from *maximum likelihoods* I_{ml}, eq. (3.24), and the one extracted from *probabilities*, I_p, eq. (3.25).

Fig. 3.6 shows, for different sample sizes, how good are the decoding procedures (after finite sampling correction). Two different quantification of the response of parvocellular unit are chosen: the spike count and the first three principal components[6] , in order to test the algorithms with simulated responses of different dimensionalities. Raw estimates (not shown in the figure) are much more biased for the maximum likelihood information I_{ml}, but anyway the finite sampling bias appears to be under control in the whole N_s range explored, as one can see from the fact that corrected estimates remain fairly constant by varying N_s. The decoding procedure chosen for this plot is the Gaussian one, even

[6] Each of three components can be considered, for the purpose of our simulation, as a different 'cell'. Notice that these principal components turn out to be only weakly correlated (in fact, linearly, but not fully, independent).

if the fit based on 'distance between test and mean response vectors', eq. (3.28), gave slightly better results. One can see that the decoding is efficient, in the sense that I_{ml} is fairly close to the true information values. This means that the efficiency of the decoding procedure in reconstructing the stimuli yields an information that is fairly close to that encoded in the true, unregularized probabilities.

The 'probability' information I_p gives instead poorer estimates of the information contained in the true distributions. This is not related to an 'inefficiency' of the decoding procedure, as explained before, but is more related to the the very definition of I_p, which is constructed to reflect also the degree of certainty with which each single trial has been decoded. Anyway, if one analyzes responses of a set of cells which are really independent (and thus adding up units to the population leads to a substantial improvement in stimulus discrimination), the two quantities I_{ml} and I_p are generally found to become closer as the size of the population increases, perhaps because the contribution to $p(s'|s)$ of wrong decodings becomes less important.

Analytical results

There is no way, in general, to estimate analytically the downward bias due to regularization occurring when using decoding procedures, as it is strongly dependent on how well the functional form chosen for $P(\vec{r}_s|s')$ fits the true response probabilities. Nevertheless, the fact that the first time derivative of the information depends only on the mean firing rates, and not on all the details of the distributions, allows one to obtain estimates of how well the decoding algorithm can estimate the rate of information transmission. This subsection is devoted to the calculation of these estimates.

Let us first consider the case of the information I_p, obtained from the probabilities $p(s'|s)$ of confusing s with s'. It can be easily shown, by using eqs. (3.23) and (3.21), that, to the first order in t $p(s'|s)$ can be written as:

$$p(s'|s) = p(s') \left[1 + t \sum_c \frac{(< r >_c - < r >_{s;c})(< r >_c - < r >_{s';c})}{< r >_c} \right] + O(t^2) \; , \qquad (3.29)$$

where $< r >_c = \sum_s P(s) < r >_{s;c}$. By substituting eq. (3.29) into the definition of I_p (3.25), it follows that the first derivative of I_p is *always* vanishing:

$$I_p(s,t) \approx O(t^2) \; . \qquad (3.30)$$

This means that I_p cannot estimate information transmission rates, and it gives poor estimates of information for small times.

In the same way, one can easily calculate the initial rate of transmission of the maximum likelihoods information I_{ml}, and find that it can be a very accurate estimate of the 'true' rate of information transmission. In fact, it is straightforward to show that the first derivative of I_{ml} coincides with that calculated from the true probabilities, eqs. (3.14,3.22), if each one of the $C+1$ events with non zero probabilities to first order in t, eq. (3.21), predicts a different stimulus.

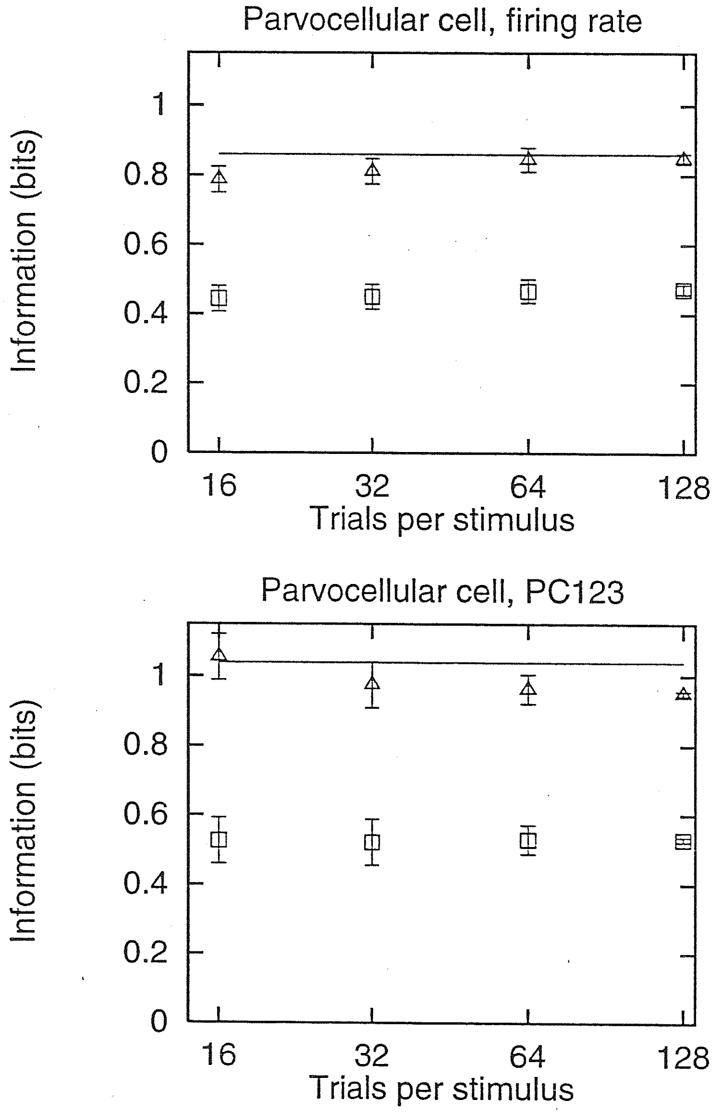


Figure 3.6:

(Top): Mutual information in the number of spikes emitted in the (simulated) responses to 32 visual stimuli by a parvocellular unit in the LGN. The full line is the 'true' value of the information in the distribution. Compared to this reference value are, for each N_s, the *maximum likelihood* information I_{ml} ($\triangle$), calculated after finite sampling corrections, and the *probability* information I_p ($\square$), again corrected for limited sampling. The N_s axis is on a log scale. (bottom) As before, but the response of the LGN cell is quantified with the three principal components instead of the spike count.

3.3.5 An example: information about spatial view in an ensemble of primate hippocampal cells

After the detailed discussion of decoding procedures reported above, we present here an application of information theoretical analysis to the study of the encoding of spatial views in the primate hippocampus. In particular, we test the idea that the firing rate (and not the temporal modulation of the spike train) is the relevant variable at the single cell level. Furthermore, we try to understand how distributed are the messages carried by the neurons in the (sequentially recorded) population of neurons under analysis.

Spatial view cells in primate hippocampus

Hippocampal function was analyzed by Rolls, Robertson and Georges-Francois (1995a: 1996b) by making recordings from hippocampal pyramidal neurons in monkeys actively walking in a rich spatial environment (the open laboratory). In this study they were able to find "spatial view" cells which responded when the monkey looked at one part of the environment, but not when he looked at another. These responses occurred relatively independently of where the monkey was in the testing environment, provided that he was looking towards a particular part of the environment. Eye position recordings with the monkey stationary confirmed that these neurons fired when the monkey looked at a particular part of the spatial environment, and not in relation to where it was. It has been also shown that these neurons respond in relation to *where* the monkey is looking in space, and not to the head direction *per se* or to eye gaze angle *per se*. For these reasons the cells were named by Rolls, Robertson and Georges-Francois (1995a; 1996b) "spatial view" cells, and not "place cells", like those in the rat hippocampus.

Here we calculate how the information about the spatial environment is represented by a population of 20 spatial view cells recorded one at time. For this purpose, we describe the walls of the laboratory as "stimuli". Because of the limited number of trials, the spatial environment (*i.e.,* the 'set of stimuli') has been discretized into up to 16 different views, each wall of the laboratory being discretized into up to 4 parts. We note that dividing the walls into 16 "stimuli" means that the information required to decode correctly where the monkey is looking at is 4 bits. Provided that this "ceiling" is not reached by the information available from one cell or the ensemble of cells, it is not really necessary to divide the space into more "stimuli", in that not much more information would be measured in the neuronal responses.

The firing rate is the relevant response characterization for information processing in single hippocampal neurons

As explained in section 3.1, to test the idea that the *temporal modulation* of the spike train, and not the *firing rate*, is the relevant parameter for information processing at the single unit level, one can compare the information carried by the firing rates and by the first few principal components of the time course of the responses.

We have performed this analysis on hippocampal view cells. Data was collected by taking a fixed length record (usually 500 ms long or more) whenever the eyes moved and then remained still (within 2 degrees) looking at a particular part of the environment. This data collection procedure has enabled us to define a natural onset in time for the stimulus, i.e. the part of the environment the monkey is looking at. The time of the stimulus onset can in fact be taken as the time when the eyes stop moving and remain still looking at the given part of the space. Then we have calculated both firing rates and principal components of each response, together with where the monkey was looking at during the record. The covariance matrix, eq. (3.5), was calculated by discretizing the sampling time window into 20 intervals. In order to have enough data to perform a meaningful analysis, we have discretized the environment into just four stimuli (four walls). In this way we have obtained from 10 to 50 repetitions for each stimulus presentations. Finally, we have calculated the information about this set of views carried by the different neuronal codes under study, by using the decoding procedure introduced before. After the decoding, the mutual information was quantified using the *maximum likelihood* information (3.24), which can give fairly good estimates of the unregularized values. We have chosen to fit the relative likelihoods by means of the algorithm based on distance between test and average response vectors (3.28), which provided better fits and (after finite sampling corrections) higher values for both information and percent correct. In table 3.3 we report our results for the average over the ensemble of single cells of the mutual information carried by firing rates and the first few principal components when responses are sampled in 500 ms (starting 100 ms after the stimulus onset) [7].

Firing Rates	0.141
PC1	0.145
PC12	0.150
PC123	0.158

Table 3.3:
Information content (averaged over the population) of the firing rates and first three principal components of the spike train. Details of the calculation reported in the text.

One can see that the information reflected by the firing rate accounts for most of the information (89 %) derived from the first three principal components. We have checked that adding more principal components leads to only negligible further increases. When one considers time windows shorter than 500 ms, the role of temporal encoding quantified in this way is even less important. In fact, when the time windows are shorter than 100-

[7] To check the effectiveness of the decoding procedure in this particular case, we have compared the estimate of information carried by unidimensional quantities (firing rates, PC1, PC2, PC3) obtained by the decoding algorithm and by direct estimation from (discretized) responses. The loss of information due to decoding remains within 10-15 % for each code considered.

200 ms,we have found not only that the majority of information is contained in the firing rates, but is already present in the 'binarized' firing rates (see fig. 3.5). The latter result is due to the fact that, for the set of hippocampal pyramidal neurons under analysis, the mean firing rate are so low that cells only seldom emit more than one spike in 100 ms.

The information available in the responses of single units

Once established that the firing rate contains the majority of the information carried by the single units, we analyze whether the information about views is carried by responses very finely tuned on each view, or instead the encoding is rather distributed.

For the analysis described here, data was collected by taking a fixed length of record (*e.g.,* 500 ms) whenever the eyes were still looking at a given part of the environment. The difference with respect to the data collected for the principal component analysis is that now, since we are interested only in the information conveyed by the firing rates, we do not need to determine the time onset of the stimulus, and thus we can collect even more than one trial if the eyes remain still, and the animal alert, for a longer period. In this way we have obtained enough data for binning each wall into four parts, and thus calculated the information about a set of 16 spatial views. This is relevant for the analysis described in the following, in that if we consider poorer discretization (*e.g.,* 4 spatial bins), the information extracted form the population approaches the ceiling (2 bits in the case of 4 bins). At the end we have collected 20 or more trials for each of the 16 views and for each of the 20 cells in sample. This data collection has been used also for the calculation of the information contained in the firing rates of a population of hippocampal neurons presented below.

From this data, we have calculated both the mutual and stimulus specific information, by using a pure discretization of the response space into 10 bins. For most of the cells, even for time windows as long as 500 ms, the number of bins was equal or larger than the maximum number of spikes. Therefore the amounts of information obtained in this way are expected to be, after the finite sampling correction, unregularized and quite accurate estimates of the information present in the cell firing rate.

A histogram showing the values of $I(\mathcal{S},\mathcal{R})$ (the average information in the responses of a cell about the stimulus set within 500 ms periods) for each cell is provided in Fig. 3.7. Most of the neurons had values for $I(\mathcal{S},\mathcal{R})$ in the range 0.15 - 0.55 bits, with the average across the population of neurons being 0.35 bits (see also Table 3.4). All these neurons show a reasonable amount of information available in the firing rates in a 500 ms period about spatial views, even though the firing rates of the neurons were low, with a mean peak response to the most effective spatial location of 13.1 spikes/s (compared to a spontaneous rate of 0.5 spikes/s). Although $I(\mathcal{S},\mathcal{R})$ may not appear to be high, it should be remembered that this neuronal information measure is the average, over stimuli, of the information contained in the responses to the individual stimuli. If many of the stimuli (walls) evoke a similar neuronal response, then the average information from the neuronal response about which stimulus was being looked at is low. If *e.g.,* the

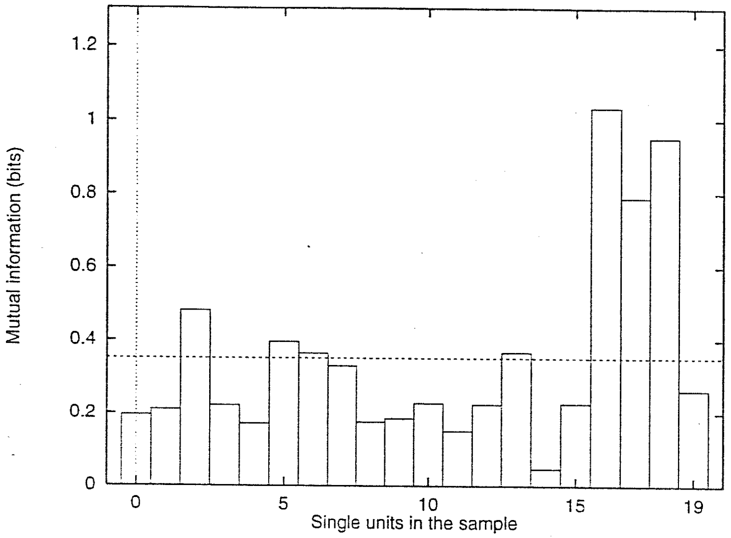


Figure 3.7:
Information in the number of spikes emitted in response to 16 different views by the 20 different hippocampal cells in the sample. The broken line (- - -) indicates the average mutual information across the population. The responses are recorded over 500 ms.

neuron responded to one of the stimuli (one of the quarter of walls), and not to any other (grandmother cell encoding), then the stimulus specific information contained when that effective stimulus was shown would be 4 bits, and, when any of the other stimuli, close to 0 bits (in fact $\log_2(16/15) = 0.093$ bits). In this case the mutual information would be 0.33 bits.

In order to understand the actual representation of individual stimuli by our set of individual cells, the information $I(s)$ available in the neuronal response about each of the stimuli (indexed by s) in the set of stimuli S has been calculated for each unit. The maximum information values I_{max} of the different neurons about any one stimulus are reported in Table 3.4, again calculated for 500 ms periods of the neuronal response. The majority are in the range 0.5 - 1.5 bits. The mean value of I_{max} for the different cells was 1.35 bits. As one can see from table 3.4, in general the neurons have different centers ("preferred view") for their view fields, but there is partial overlap among the view fields of some of the cells. It is evident from this analysis that the encoding is much more distributed and noisy than that of a set of ideal 'grandmother' units, although the total amount of mutual information is similar in the real and the ideal 'grand mother cells' case.

Nevertheless, from this single cell analysis, one cannot establish whether the encoding is truly distributed or not. As explained before, this point can be better studied by considering the information extracted from the population response vector.

Redundancy versus independence across different cells

Evidence on the nature of population encoding and on the representational capacity in the hippocampus can be obtained by examining the response properties of the population of the 20 view cells. We address the problem at the level of the firing rates (without considering more complex neuronal codes), because this appears to be the relevant characterization of the responses of single units. We analyze data by calculating the information about spatial view carried by (randomly chosen) subsets of cells of any size, from single neurons to the entire set recorded, and then by averaging over subsets with the same size. In this way we measure how the information scales with the number of neurons.

To approach the problem of the independence of the messages carried by different cells, we note that, given a population of C cells, a set of stimuli S, and the relative values of single cells informations $I_c, c = 1, \cdots, C$, the information that is present in the population (and in subpopulations) has its range limited from two bounding values:

- If the messages of single cells are *fully redundant*, then, from each population, one can extract only an information equal to the maximum single cell information in this ensemble. The information from C cells obtained in this way is denoted with $I^r(C)$. This quantity $I^r(C)$ (the superscript r denotes 'redundancy') is the information that we would obtain in the case of full redundancy among cells. It is important to note that, when calculating how information grows with the population size, due to the averaging over different subsamples with the same size, even in this case of full redundancy, one can obtain an apparent linear increase of information with the number of neurons (*e.g.,* when only a few cells carry information, see the result of our analysis of responses of cells in the rat primary somatosensory cortex to painful stimulations, reported below).

- In the case of full independence of messages, the information extracted from the ensemble is just the sum of the information carried by single units:

$$I^i(C) = \sum_{c=1}^{C} I_c \tag{3.31}$$

  where $I^i(c)$ denotes the information in the case of full independence. In the case of full independence, the information increase linearly with the number of neurons, irrespective of the actual values of information carried by single cells.

In real cases, the information carried by the ensemble of cells is in between these two interesting theoretical bounds.

We calculate both the *probability* (3.25) and *maximum likelihood* (3.24) information. For comparison, we repeat the analysis for longer (500 ms) and shorter (100 ms) time windows. In fig. (3.8, top) and (3.9,top), we report the way that the *maximum likelihood* information, defined in (3.24), grows with the population size. For comparison, we plot the two theoretical bounds corresponding to full redundancy and full independence, as calculated from the single cell values (computed as maximum likelihood information extracted from the same decoding procedure). In Fig. (3.8, bottom) and (3.9, bottom), we report the results obtained in the same way, but using the *probability* information (3.25) instead of the maximum likelihood one.

By looking at figures 3.8 and 3.9, it is evident that the estimate of information via the decoding algorithm is in the case of small populations not accurate enough, as it can be seen also from the fact that the information extracted form the population is sometime higher than the theoretical upper bound derived from the estimates of information carried by single cells. Since the exact values of upper and lower bounds are strongly dependent on decoded single unit information values, they should be taken only at a qualitative level. The problem in decoding the information contained in small populations can be understood if one examines the average value of information carried by single cells, as estimated directly from the responses (Table 3.5). The estimates of single cell information obtained with I_p appear much more regularized than those obtained with I_{ml}, but the latter are affected by heavier sampling errors and larger fluctuations (remember that with 20 trials per stimulus and 16 stimuli our correction procedure is at the limit of its effectiveness).

All those problems are much less important at higher number of cells, where the decoding is more reliable, both fluctuations and sampling errors are smaller, and the two quantities I_{ml} and I_p become closer. The comparison between I_p and I_{ml} as functions of the number of neurons is reported in fig. 3.10.

Even if the results are not very clear, and bigger samples of neurons are probably needed to shed more light on the problem, one can anyway point at two facts:

- We see that in all the cases the information extracted from the population grows much faster than the "fully redundant" information I^r, indicating that the growth of information with the number of coding neurons in not an artifact due to the average over subsamples, but is a genuine effect due to population encoding of the messages.

- Another important point is that at higher number of cells, when decoding is likely to be more reliable, both I_p and I_{ml} are increasing fairly linearly with the size of the population. In this sense the results are at least compatible with a distributed representation of spatial views in the primate hippocampus, and with a discrimination capacity exponentially increasing with the number of coding cells.

3.3.6 Another example: redundant coding of somatosensory stimulations in the rat somatosensory pathway

Among the possible efficiency principles evoked to explain the nature of neuronal sensory processing, one of the most interesting is the *principle of reduction of redundancy* (Atick 1992).

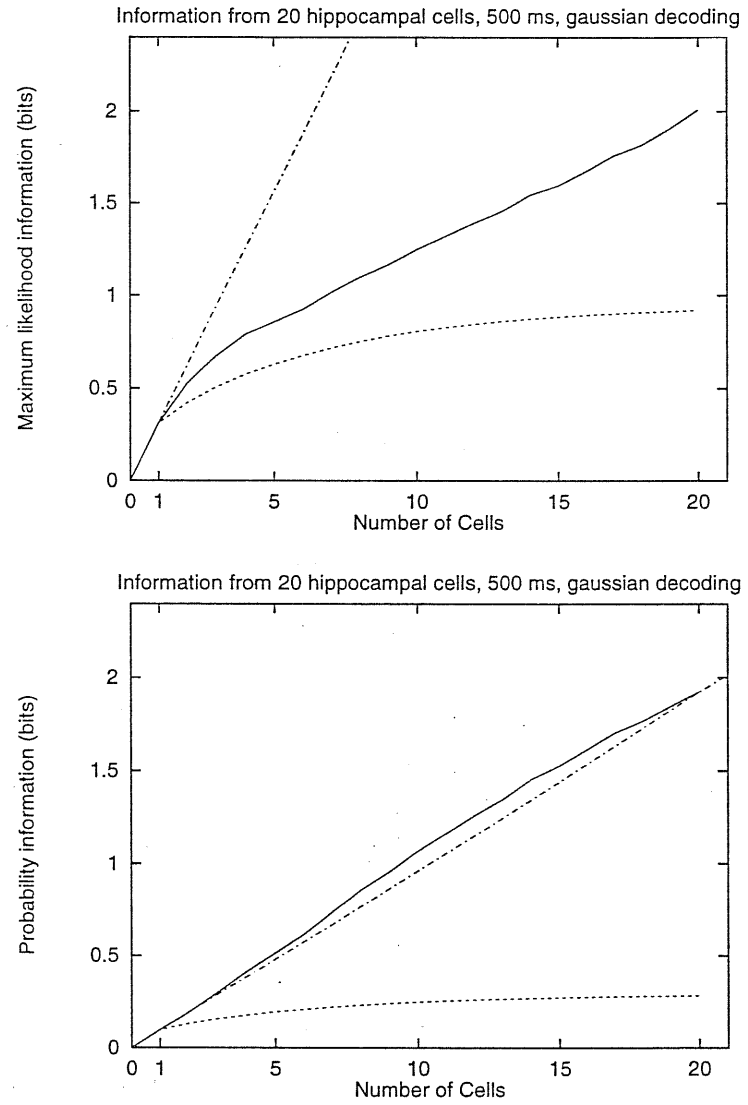


Figure 3.8:

(Top): Average *maximum likelihood* information about 16 spatial views extracted from subsets of hippocampal cells from a sample of 20 cells (——). The other lines represent the information that would be extracted in the case of full redundancy (– – –) and full independence (— · —). The size of the time bin used for counting spikes is 500 ms. (Bottom) The same as the top figure, but for the *probability* information.

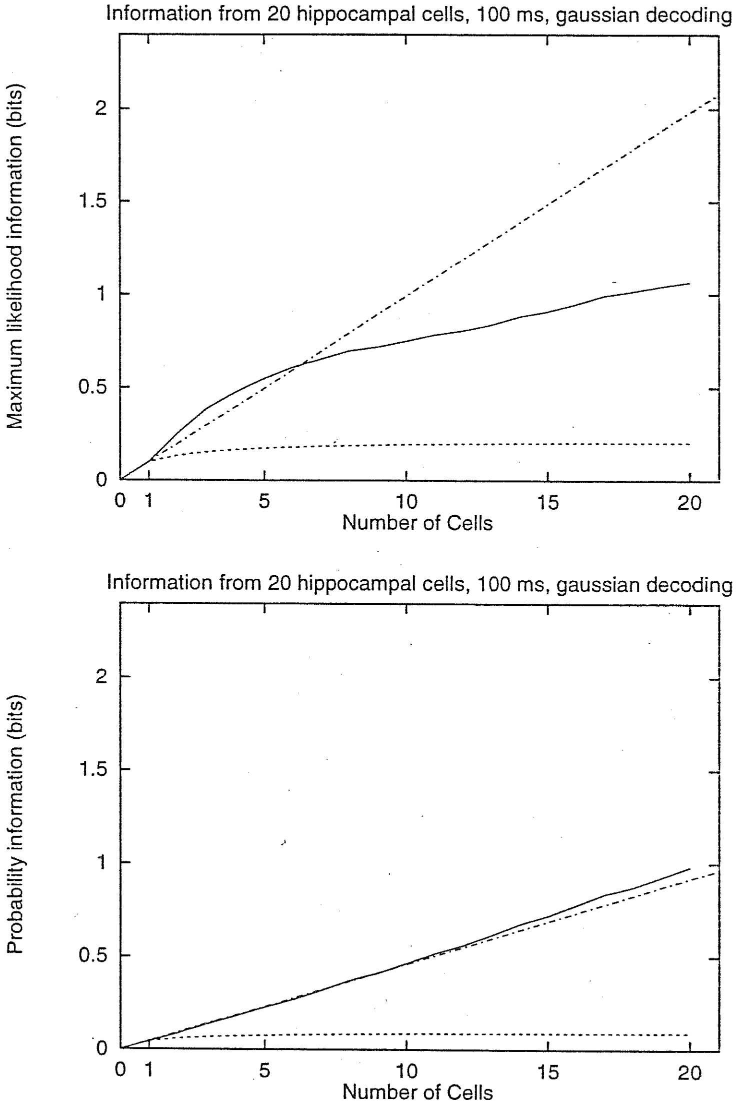
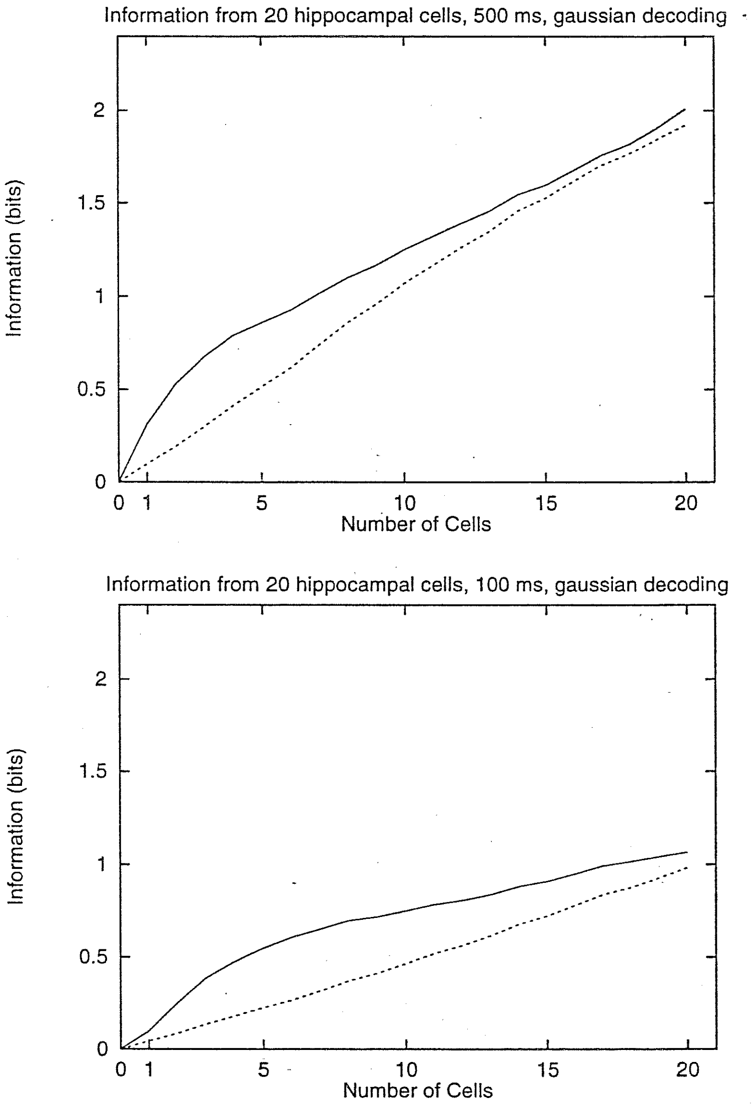


Figure 3.9:

The same as in fig. 3.8, but the size of the time bin used for counting spikes is now 100 ms.


Information from 20 hippocampal cells, 500 ms, gaussian decoding

Information from 20 hippocampal cells, 100 ms, gaussian decoding

Figure 3.10:

Average *maximum likelihood* information (——) and *probability* information (– – –) about 16 spatial views extracted from subsets of hippocampal cells from a sample of 20 cells. Responses are sampled over 500 ms (Top) and 100 ms (Bottom).

This principle states that the purpose of the first stages of sensory processing is to minimize the redundancy present into the signals coming from the external environment. According to this principle, cells in higher stages of the sensory pathway should represent feautures in a compact form, in other words each cell should carry nearly independent information. The 'minimal redundancy' strategy provides evolutionary advantages, such as a saving of neuronal dynamical ranges, and cognitive advantages related to learning of associations (Atick 1992; Barlow 1989).

This optimization principle has been successfully used in the study of the early visual system. In the case of the visual system, one can in fact observe that the typical incoming signal (*i.e.*, 'natural' images) is highly redundant (nearby pixels are highly correlated), and that the structure of correlations is reproducible among different natural images. Starting from the properties of natural images, several investigators have argued that the retinal and LGN systems are designed to reduce redundancy and to code for natural scenes in a compact form. The observed properties of the receptive fields of retinal ganglion cells and LGN cells turn out to be compatible with the principle of efficent coding of natural visual scenes (Atick and Redlich 1990; Dong and Atick 1996).

In this section. we report the results of an attempt to understand if redundancy reduction is working also in the coding of painful stimulations by the somatosensory system of the rat. This issue was analyzed recording simultaneously from small sets of cells in two successive stages of the somatosensory pathway, VPL thalamic nucleus and SI cortex, when a set of noxious stimulations was applied to the periphery. Then the redundancy of the representation of those stimuli in the thalamus and in the SI cortex was quantified by means of the information theoretical quantities discussed previously. The main concern in exploring this issue is the lack of a good characterization of a 'natural' statistics of this particular kind of stimulations. This fact can render any analysis of responses from a limited set of stimuli hardly generalizable to the case of a larger, more ecological, set of stimulations. The situation is further complicated by the fact that adaptation to the noxious stimuli strongly constraints the number of different stimulations that one can apply to the animal. Nevertheless, one can argue that, if a sensible reduction of redundancy of neuronal representation in succesive somatosensory stages is regularly observed when different sets of noxious stimulations are applied, then a redundancy reduction mechanism is effectively at work in this particular modality.

Methods

We have analyzed responses of simultaneously recorded cells in different regions of the somatosensory pathway. Data was kindly provided by Gabriele Biella and coworkers. In a series of experiments, they performed extensive and concurrent recordings from the VPL thalamic nucleus and from the granular zone of the SI cortical region, in anesthetized and curarized rats. A group of up to five electrodes in a ring configuration (max 500 μm of the array diameter, *ad*) was placed in the VPL at 25 degrees in the parasagittal plane ($\sim 5.5 - 6$ mm in transverse distance depth), while a ring of up to seven electrodes (max *ad* 600 μm) was placed in the cortex

(0.5-1 mm from bregma). A set of electrical stimulations of a given intensity (from 1.5 to 8 mA), or a set of more "natural" stimulations, like brushing, pinching and thermal stimulations, were delivered to the periphery (palmar region of the hindpaw). More details on the recording procedure and on the data analysis performed are reported in Biella *et al.* (1995) and Panzeri *et al.* (1995a).

Results

The goal of this study is to understand which is the typical encoding strategy used by these brain regions to represent this simple class of painful stimulations, and to understand whether there are significant differences between the neuronal representation in the two regions.

For this purpose, we start by describing the results of the recording of a relatively large set of cortiacl cells, whose distribution of messages among different neurons reproduces well what has been typically found in the various recording sessions. In this experiment, the activity of 7 cells in SI was simultaneously recorded, during the presentation of a set of electrical stimulations, 1 sec. long, consisting in a series of very short (less than 1 msec) electrical pulses with the same intensity (3 ma) and with different frequencies (10,20,40 Hz). Different stimuli were presented in a pseudorandom order. To reduce adaptation, stimulations were separated by a time interval of 3-10 sec. The spontaneous activity of the cells (which will be called in the following the "0 Hz stimulation") was also recorded. Then the mutual information between stimuli (stimulations of 10,20,40, and 0 Hz) and responses was calculated, on the basis of 10-15 trials per stimulus.

We report in Tab. 3.6 the average responses to different stimuli of each cell in the sample, and in Tab. 3.7 the value of the stimulus specific and the average information carried by firing rate distributions of each cortical cell in the sample. The information is in this case evaluated from the (discretized) responses. It is important to note that all the cells show a high degree of selectivity to the painful stimuli, and high information values. The "preferred stimulus", *i.e.,* the stimulus that carries the highest stimulus specific information, and is better predictable from responses, is the same for *all* the cell in the sample. In this case the cells discriminate pretty well between noxious and non noxious stimulations, but are less selective for particular noxious stimulations. Moreover, among the different noxious stimulations, the most discriminable one is generally the same for the cells considered.

From the last observation it is straightforward to conclude that the recorded cells convey essentially the same information, perhaps only reflecting the intensity of the responses of pain receptors with the intensity of their single-cell response. This point is well illustrated by the fact that (fig. 3.10) the information extracted from the population[8] is very close to the information available in the case of *full redundancy.*

[8] The information extracted from the population is quantified with the maximum likelihood information, which, in this case of strong selectivity, gives good estimations of the single cell information extracted from the discretized responses (Tab. 3.7)

We have analyzed not only this particular set of cortical cells responding to this set of stimulations, but also data from different samples of cortical or thalamic neurons, responding to 'more natural' stimulations, or to electrical stimulations of different intensities, or to electrical stimulations of different frequencies, or to groups of stimuli of these different classes mixed together. But the behaviour that we have found was always the same, *i.e.*, the neurons in the sample which showed selectivity with respect to the stimuli presented had very correlated response profiles, perhaps reflecting the intensity of the responses of pain receptors with the intensity of their single-cell response.

No marked differences has been found between the encoding of thalamus and cortex. As an example, we report in Fig. 3.12 the information extracted from the firing rates of a set of three cells in the cortex and five cells in the thalamus, concurrently recorded. In this case, the set of stimulations consited of brushing, mechanical pinching, thermal stimulation with a fixed temperature, and spontaneous activity. Firing rates were recorded over 500 ms and from the response vector of the population the maximum likelihood information (3.24) was extracted. Again, it is clear that the information about the stimuli carried by the two different populations is almost fully redundant, with no significant differences between thalamus and cortex. It is interesting to note that, in the case of the cortex (Fig. 3.12, bottom), the linear increase of information with the size of the population is only an artifact of the average over the possible subsamples with the same size, and reflects the fact that one of the three cell carried a lot of information (nearly 1 bit), whereas the other carried small amounts of information.

Discussion

We have found that the coding of simple painful stimulations by two regions of the somatosensory system of the rat is very redundant, and there is no evidence of a redundancy reduction in successive stages. The result of our analysis could be related to the particularly simple set of painful stimulations used, but anyway it suggests that the nature of sensory information processing may be very different when considering different modalities and different conditions. Further, it emphasizes the necessity, in order to perform an analysis of redundancy which can give nontrivial results, of having a system which discriminates among a large variety of stimuli, and of testing it with an appropriate and representative subset of those stimuli.

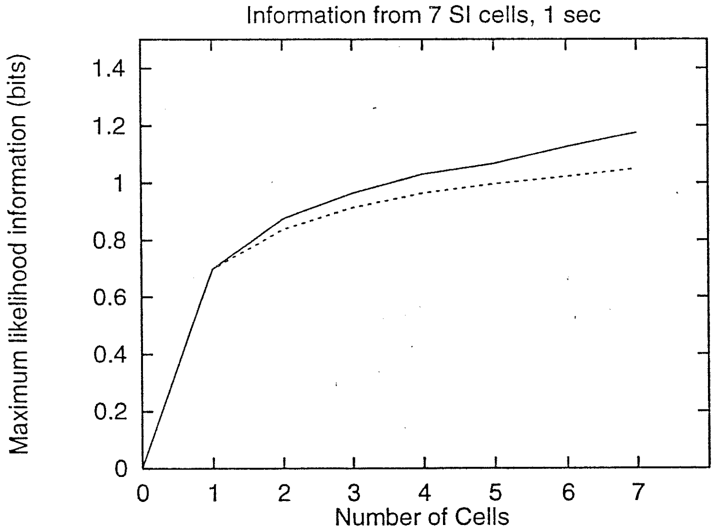


Figure 3.11:
Average maximum likelihood information about 4 electrical stimulations of different intensities extracted from subsets of cells (from a sample of 7) in the rat SI cortex (——). The broken line (– – –) represent the information that would be extracted in the case of *full redundancy*. The number of spikes emitted by each cell are counted within a time window of 1 sec (during the presentation of the stimulus).

Cell	$I(\mathcal{S},\mathcal{R})$	Preferred view	I_{max}
0	0.195	2	0.458
1	0.210	7	1.852
2	0.481	2	2.881
3	0.222	1	0.887
4	0.171	2	0.668
5	0.396	2	0.898
6	0.364	10	0.839
7	0.330	16	0.668
8	0.176	12	0.564
9	0.185	14	0.943
10	0.228	4	0.752
11	0.151	2	1.294
12	0.224	5	0.499
13	0.367	6	1.561
14	0.046	10	0.235
15	0.227	1	1.129
16	1.035	16	3.596
17	0.789	13	2.620
18	0.952	4	3.709
19	0.263	13	0.918

Table 3.4:
Values of the mutual information contained in the firing rates ($I(\mathcal{S},\mathcal{R})$), the preferred view, and the maximum value of the stimulus specific information (I_{max}) for each of the 20 cells in the sample. The responses are recorded over 500 ms.

	100 ms	500 ms
$I(\mathcal{S},\mathcal{R})$	0.135	0.351
I_{ml}	0.099	0.312
I_p	0.044	0.096

Table 3.5:
Average over the population of 20 hippocampal cells of mutual information extracted form single units by means of evaluation from discretized responses ($I(\mathcal{S},\mathcal{R})$), *maximun likelihood* information (I_{ml}), and *probability* information (I_p).

Cell	10 Hz	20 Hz	40 Hz	0 Hz
0	63.0	92.0	113.8	43.4
1	57.9	83.2	75.0	35.8
2	48.1	132.2	123.6	31.4
3	41.8	52.2	42.8	19.8
4	61.6	71.6	65.7	42.8
5	61.3	93.3	91.0	41.8
6	45.6	81.3	78.4	28.6

Table 3.6:
Average firing rate responses to each stimulus for the 7 different cells in the rat SI cortex. Firing rates are measured over a time period of 1 sec during the stimulus presentation.

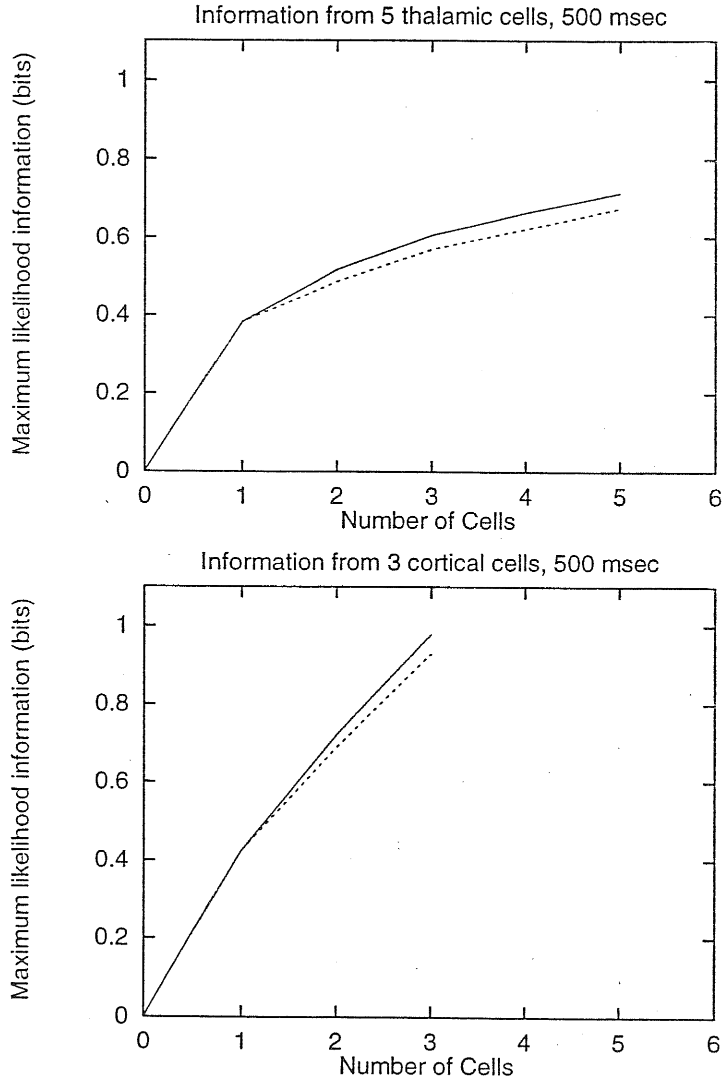


Figure 3.12:
Average maximum likelihood information (——) about 4 'natural' stimulations (brushing, pinching, thermal stimulation and spontaneous activity) extracted from a sample of 5 cells in the VPL thalamic nucleus (Top), and 3 cells in the SI cortex of a rat. The broken line (– – –) represent the information that would be extracted in the case of *full redundancy*. The number of spikes emitted by each cell are counted within a time window of 500 msec (during the presentation of the stimulus).

Cell	I	$I(s = 10Hz)$	$I(s = 20Hz)$	$I(s = 40Hz)$	$I(s = 0Hz)$
0	0.490	0.155	0.416	0.416	0.888
1	0.766	0.383	0.748	0.348	1.474
2	1.085	0.981	1.113	0.923	1.186
3	0.569	0.044	0.505	0.128	1.513
4	0.343	0.000	0.169	0.281	0.907
5	0.855	0.551	0.701	0.647	1.404
6	0.956	0.812	0.838	0.578	1.469

Table 3.7:
Mutual information and stimulus specific information carried by the firing rate distributions of the 7 different cells in the rat SI cortex.

Chapter 4

A quantitative model of information processing within the hippocampus

In the previous chapters, we have discussed how to quantify, by using information theory, the way in which external correlates are coded in the spike train of single (or multiple) units. The full power of the quantitative approach based on information theory can nevertheless be achieved only if one is able to relate the behaviour of real brain structures, as observed by recording the activity of its units, to models that can establish (or predict) quantitive relations betwen structures and functions (in the sense of information processing).

A brain region which could be fruitfully studied within this approach is the hippocampal formation. In fact, in the last few years a number of papers (Treves and Rolls 1992, 1994; Rolls 1995; Treves 1995; Treves *et al.* 1996)) has related, within the theory that describes the hippocampus as a device for the on-line storage of complex memories, several anatomical and physiological aspects of the hippocampal organization to the requirement of optimizing the functions it performs. The approach reported in these papers is based on the idea that, from an abstract, information-theoretical level, the function of the hippocampus as a memory device should be optimized to store and retrieve efficiently information; in this way, one can study which parts of the hippocampus seem to be designed to perform a task pertaining to such a memory device. By using this approach, Treves and Rolls were able to show, among other results, that the crucial autoassociative operations are ascribed mainly to the recurrent CA3 network. As shown by Treves (1995), the Schaffer collateral connections from CA3 to CA1 may still be important, both in completing information retrieval and in re-expanding, with minimal information loss, the highly compressed representation retrieved in CA3.

In this chapter we try to extend those results by taking explicitly into account the contribution to the retrieval given by the other input system to CA1, the direct perforant path projection from entorhinal cortex. These connections may allow CA1 to integrate the complete but compressed representation retrieved form CA3, with the partial (but information richer) representation, available in entorhinal cortex, of those elements of memory that served as a cue. To quantify these effects, we have introduced a suitably realistic formal model of the relevant circuitry, taking into account explicitly both the input systems to CA1, and evaluated its performance in the sense of information theory. Then we have provided, by means of the techniques of statistical mechanics, an analytical expression for the amount of information present in the model CA1 output, about what has been presented in the entorhinal cortex. This analytical solution of the problem is given as a function of saddle point parameters, which are related, through the saddle point equations, to the parameters characterizing the network. The saddle point equations turn out to be very difficult to solve, and we have not studied, so far, a region of the parameter space big enough to give a full computational account of the functions attibuted to the perforanth path connections to CA1. Nevertheless, we think that it is worthwile to report the analytical solution of the model, both because the calculation is non trivial and because, in the spirit of this dissertation, it is important to develop methods which can give accurate descriptions of the informational performances of biologically plausible neural networks.

This chapter is organized as follows: in the first section we briefly review the basic facts of the hippocampal organization relevant to our analysis, together with the basic computational and functional hypotheses contained in the Treves-Rolls model (Treves and Rolls 1992, 1994) which are the starting point of our analysis. The second section is devoted to the definition of our model, and to the discussion of the underlying hypothesis. In the third section we report the analytical evaluation of the information present in the CA1 output about a pattern presented in the entorhinal cortex, and we discuss how to find the numerical solutions of the saddle point equations. Finally, in the last section we conclude by discussing the results, and the possibility of using data available from simultaneous or sequential recording of hippocampal cells (Wilson and Mc Naugthon 1993, Skaggs *et al.* 1993, Treves *et al.* 1996, Rolls *et al.* 1995a) to validate some of the building hypothesis of the model, and to check its quantitative predictions.

4.1 The Hippocampal System

The very brief review of the salient features of hippocampal organization, and the discussion of computational hypotheses for the CA3 and CA1 region that we discuss here, are not presented to give a new, original view of hippocampal structure, but to emphasize which hypotheses we are going to test with our quantitative model.

4.1.1 Hippocampal architecture and plasticity

The hippocampus receives, via the adjacent parahippocampal gyrus and entorhinal cortex, inputs from virtually all association areas in the neocortex, including those in the parietal, temporal, and frontal lobes (Squire et al., 1989). Therefore the hippocampus has available highly elaborated multimodal information, which has already been processed extensively along different sensory pathways. An extensively divergent

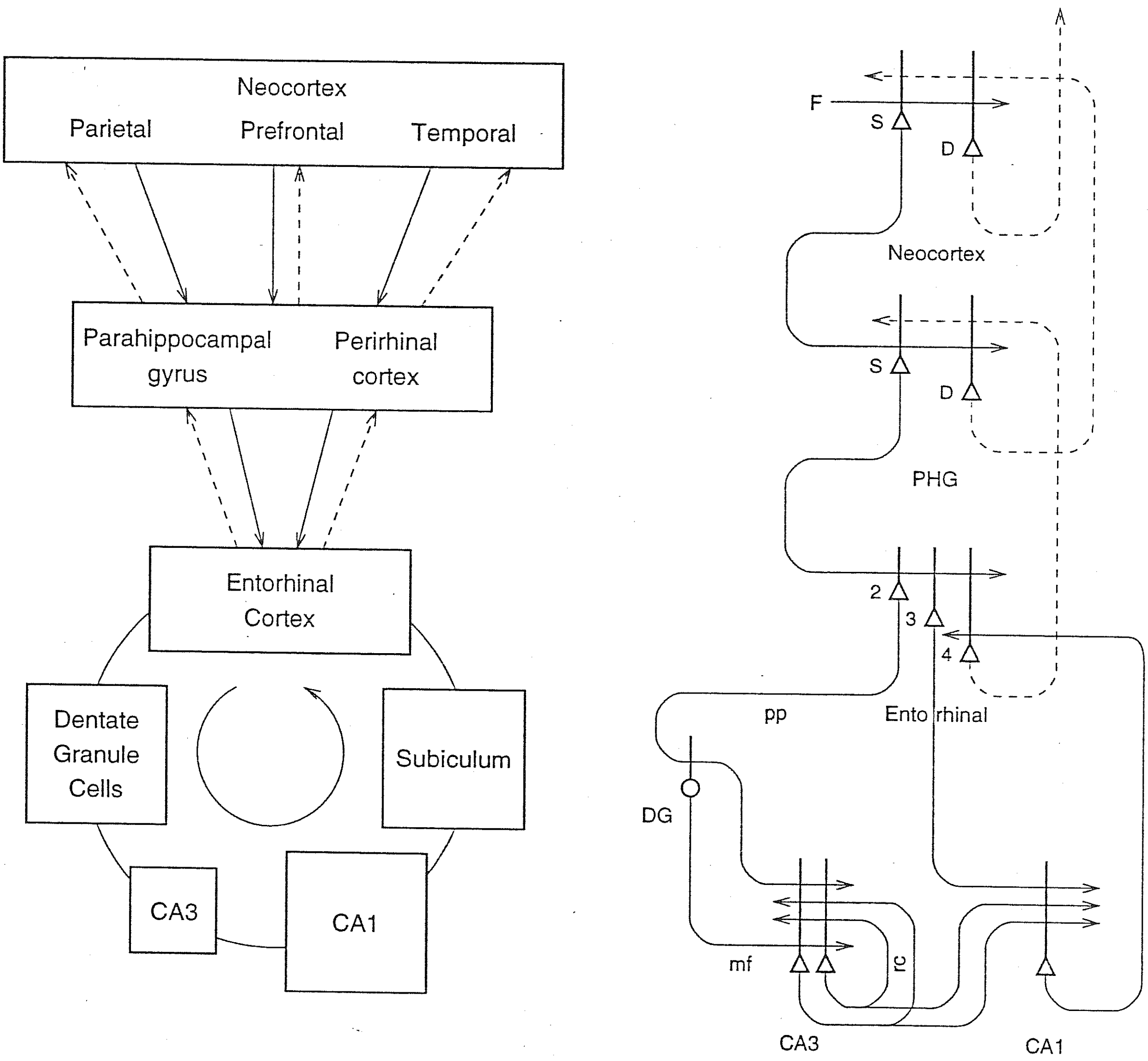


Figure 4.1:
Forward connections (solid lines) from areas of cerebral association neocortex via the parahippocampal gyrus and perirhinal cortex, and entorhinal cortex, to the hippocampus; and backprojections (dashed lines) via the hippocampal CA1 pyramidal cells, subiculum, and parahippocampal gyrus to the neocortex. There is great convergence in the forward connections down to the single network implemented in the CA3 pyramidal cells; and great divergence again in the backprojections. Left: block diagram. Right: more detailed representation of some of the principal excitatory neurons in the pathways. Abbreviations - D: Deep pyramidal cells. DG: Dentate Granule cells. F: Forward inputs to areas of the association cortex from preceding cortical areas in the hierarchy. mf: mossy fibres. PHG: parahippocampal gyrus and perirhinal cortex. pp: perforant path. rc: recurrent collateral of the CA3 hippocampal pyramidal cells. S: Superficial pyramidal cells. 2: pyramidal cells in layer 2 of the entorhinal cortex. 3: pyramidal cells in layer 3 of the entorhinal cortex. The thick lines above the cell bodies represent the dendrites. (From Rolls (1995)).

system of output projections enables the hippocampus to feed back into most of the areas from which it receives inputs.

Information is processed within the hippocampus along a distinctly unidirectional path, consisting of three major stages, as shown in Fig. 4.1 (Amaral and Witter, 1989; Amaral, 1993). Axonal projections mainly from layer 2 of entorhinal cortex reach the granule cells in the dentate gyrus via the perforant path (pp), and also proceed to make synapses on the apical dendrites of pyramidal cells in the next stage, CA3. A different set of fibres projects from entorhinal cortex (mainly layer 3) directly onto the third processing stage, CA1.

There are about 10^6 dentate granule cells in the rat, and more than 10 times as many in man (more detailed anatomical studies are available for the rat) (Amaral et al., 1990; West and Gundersen, 1990). They project to CA3 cells via the mossy fibres (mf), which form a relatively dilute (low probability of connection) but possibly powerful synaptic matrix; each fibre makes, in the rat, about 15 synapses onto the proximal dendrites of CA3 pyramidal cells. As there are some 3×10^5 CA3 pyramidal cells in the rat (SD strain; 2.3×10^6 in man, Seress, 1988), each of them receives no more than around 50 mossy synapses. (The connectivity is thus 0.005%.) By contrast, there are many more - possibly weaker - direct perforant path inputs onto each CA3 cell, in the rat of the order of 4×10^3. The largest number of synapses (about 1.2×10^4 in the rat) on the dendrites of CA3 pyramidal cells is, however, provided by the (recurrent) axon collaterals of CA3 cells themselves (rc). The CA3 system thus provides a single network, with a connectivity of approximately 4% between the different CA3 neurons. The implication of this widespread recurrent collateral connectivity is that each CA3 cell can transmit information to every other CA3 cell within 2-3 synaptic steps. The CA3 system therefore is, far more than either DG or CA1, a system in which intrinsic, recurrent excitatory connections are, at least numerically, dominant with respect to excitatory afferents.

In addition, there are also intrinsic connections with a variety of numerically limited and mainly inhibitory populations of interneurons. Here we assume that such connections perform only the function of keeping the activity of hippocampal neurons within well-defined bounds, and not of providing signals specific to the information being processed in the system.

Extrinsic axonal projections from CA3, the Schaffer collaterals, provide the major input to CA1 pyramidal cells, of which there are about 4×10^5 in the (SD) rat. The CA1 pyramidal cells are characteristically smaller than the CA3 ones and, across different species, come in larger numbers. In terms of cell numbers, therefore, information appears to be funnelled from DG through the CA3 bottleneck, and then spread out again into CA1. The output of CA1 returns (directly and via the subiculum) to the entorhinal cortex, from which it is redistributed to neocortical areas.

Neurophysiological evidence also indicates that many of the synapses within the hippocampus are modified as a result of experience, in a way explicitly related to the types of learning for which the hippocampus is necessary, as shown in studies (Morris, 1989) in which the blocking of such modifiability with drugs results in specific learning impairments.

Studies on long-term potentiation (LTP) have shown that some synaptic systems (in DG and CA1, but probably also pp and rc synapses in CA3) display a "Hebbian", or associative, form of plasticity, whereby presynaptic activity concurrent with strong postsynaptic depolarization can result in a strengthening of the synaptic efficacy (Brown et al., 1990; Miles, 1988). Such strengthening appears to be associated with the activation of NMDA (N-Methyl-D-Aspartate) receptors (Collingridge and Singer, 1990), and it is possible that the same synapses display also (associative) long-term depression (Levy and Desmond, 1985; Levy et al, 1990). Also MF synapses are known to display long-term activity-dependent synaptic enhancement, but this form of enhancement appears not to be associative (Brown et al., 1990).

4.1.2 Computational hypotheses arising from functional constraints: a quick review of the Treves-Rolls model

Here we briefly review the functional role suggested for various hippocampal subfields in the work by Treves and Rolls, by arguments based on a computational analysis. The aim of the section is not to summarize this interesting model of hippocampal function, or to present it as an established 'truth', but to discuss and emphasize only those hypothesis that we have taken as starting point of our analysis.

Rolls (1987, 1989) has suggested that the reason why the hippocampus is used for episodic memory, is that the hippocampus contains one stage, the CA3 stage, which acts, by means of its extensive network of recurrent collaterals (RC), as an autoassociation memory. This hypothesis implies that any new event to be memorised is given a unitary representation as a firing pattern of CA3 pyramidal cells, that the pattern is stored in associatively modifiable synapses from the recurrent collateral axons, and that subsequently the extensive recurrent collateral connectivity allows for the retrieval of a whole representation to be initiated by the activation of some small part of the same representation (the cue). Computational constraints suggested specific roles for the two input systems to the CA3 network (Treves and Rolls 1992). In particular, the mossy fibers, and with them the whole dentate network, might be regarded as a device to force very efficient information storage into CA3, by virtue of their strong (and sparse) influence on the CA3 cell firing rates. The perforant path connection to CA3, instead, should help in the process of retrieval from a partial cue, in that this large system of weak associatively modifiable synapses can relay a signal specific enough to initiate retrieval.

The information retrieved within the CA3 autoassociator has to be sent back to neocortical areas, with minimal waste. In this view, the CA1 network can be considered (Treves and Rolls 1994) to be both the first step in the relay from CA3 back to neocortex, and the last stage of the hippocampal associative memory system. One crucial advantage of having CA1 after the CA3 stage is that the very compressed representation provided by CA3 pyramidal cells can be reexpanded onto the larger number of CA1 pyramidal cells, resulting in the same information being coded in a much more robust manner. Obviously the recoding is effective only if at least it *preserves* the overall information content of the representation. Further, CA1 can also contribute to associative retrieval itself by *increasing* this information content over and beyond that of the representation retrieved from CA3. A quantitative assessment of such information (Treves, 1995) shows that both preservation and increase can occur, depending on the balance of the various noise parameters in the different subfields, if the CA3-to-CA1 connections, the Schaffer collaterals, are endowed with associative Hebbian modifiability. In particular, there is an optimal range of the plasticity parameter which is the one that matches the plasticity of CA3 recurrent connections. This analysis then suggest that the two synaptic systems (within CA3 and from CA3 to CA1) may be optimally organized if they share the same molecular and biophysical mechanism based on NMDA-receptor-dependent potentation.

Another interesting feature of the CA1 network is its double set of afferents, with each cell receiving a definite proportion of extrinsic inputs not only from Schaffer collaterals but also from direct perforant path projections from (mainly) layer 3 of entorhinal cortex. The existence of this second set of inputs indicates that, in some conditions, there is need, after the CA3 stage, of some information closely related to the original input given to the hippocampus. It has been suggested (Treves and Rolls 1994) that the perforant path projection may serve, during retrieval, to integrate the description of the full event recalled from CA3, with the information rich description of only those elements of the respresentation used as a cue provided by the entorhinal/perforant path signal. Nevertheless, from this semiqualitative analysis is not clear at all in which region of the noise parameters the effect of this input could be necessary for the recoding in CA1 with minimal waste, and when this input could be irrelevant. The aim of the next section is to extend the analysis, introduced by Treves (1995) to quantify the information relayed by Schaffer collaterals, to include the direct perforant path projections to CA1. It is hoped that this work will clarify the contribution of direct entorhinal inputs to the information content in the hippocampal output, providing indications as of the reason for the relative abundance of perforant path and Schaffer collateral synapses onto CA1 cells, and making predictions about the synaptic systems under considerations (within CA3, from CA3 to CA1, and from entorhinal cortex to CA1).

4.2 The Model

The information content of a 'CA1' firing pattern has been evaluated analytically using a formal model. The model describes, in simplified form, the Schaffer collateral connections from the N pyramidal cells of CA3 to the M pyramidal cells of CA1, and the direct perforant path connections from the L pyramidal cells of the entorhinal cortex (EC) to CA1. It considers the connections from EC to CA3 in a simplified way, under the hypothesis, discussed in the previous chapter, that the mossy fibers and the dentate network are able to force a very efficient information storage in CA3. Morevoer, the model does not include directly the effect of inhibitory interneurons, but it is asssumed that they exert a general regulation of pyramidal cells activity, setting up effective thresholds for the pyramidal cells, such as to produce activity distributions with a given sparseness. The weak (Amaral and Witter, 1989) CA1 recurrent collateral system is neglected.

The system works as follow: a given pattern of activity in EC is presented to the hippocampus via the projection systems. Two distinct modes of operation of the hippocampus are envisioned: *storage* and *retrieval*. During storage the synaptic efficacies on both the Schaffer collaterals and perforant path (PP) connections are modified in a Hebbian way reflecting the conjunction of pre- and post-synaptic activity – but the modification is not immediate and thus does not influence the current CA1 output. During retrieval, the Schaffer collaterals relay a pattern of activity retrieved from CA3, and the PP relay the pattern[1] represented in EC. The synaptic efficacies of these connections,

[1] In fact, only a part of the EC pattern (a cue).

while not being presently modified, reflect all previous storage events. In this way, five different patterns of neuronal firing are considered (Table 4.1)

	CA3	CA1	EC
storage	$\{\eta_i\}$	$\{\zeta_j\}$	$\{\nu_k\}$
retrieval	$\{V_i\}$	$\{U_j\}$	$\{\nu_k\}$

Table 4.1: The five different firing pattern appearing in the analysis. Each symbol denotes the firing rate of the cell indexed by the subscript.

- $\{\nu_k\}$ are the firing rates of each cell k of Entorhinal Cortex, which together code for the information to be stored, and later retrieved, from the hippocampus. The firing pattern $\{\nu_k\}$ represent the information received by EC from neocortical association areas. Statistically, the probability density of finding a given firing pattern is taken to be a product, for each cells, of a certain "typical" firing rate distribution:

$$P(\{\nu_k\}) = \prod_k P_\nu(\nu_k) d\nu_k \tag{4.1}$$

  This assumption means that each cell in EC is taken to code for independent information, an highly idealized version of the idea that by this stage most of the redundancy present in earlier representations has been removed. The assumption of independent input patterns, necessary for our solution of the model, appears to be reliable for the hippocampal cells, where there is some evidence that each cell code for nearly independent information [2], but has not been verified for the entorhinal cortex, where quite possibly the cells might be further from this assumption. Work is in progress to provide analytical solution of the model even in presence of symple types of correlation among EC cells.

- $\{\eta_i\}$ are the firing rates produced in each cell i of CA3, *during the storage* of the EC representation; they are determined by the matrix multiplication of the pattern $\{\nu_k\}$ with the synaptic weights I_{ik} (of zero mean, as explained below, and variance σ_I^2) followed by Gaussian distortion, (inhibition-dependent) thresholding and stepwise rectification:

$$\eta_i = \Xi\left(\eta_0 + \sum_{k=1}^{L} d_{ik} I_{ik} \nu_k + \delta_i^S\right)$$
$$< (I_{ik})^2 >= \sigma_I^2; \quad < (\delta_i^S)^2 >= \sigma_{\delta^S}^2 \tag{4.2}$$

[2] For the rat hippocampus see Treves *et al.* 1996, whereas for primates see the discussion in chapter 3 of this thesis

The connections between EC and CA3 are effectively implemented by the Gaussian connections I_{ik}. As discussed previously, this model does not describe explicitly the dentate network, but the sparse and information rich representation set up on the CA3 cells by the dentate network is achieved by tuning the parameters σ_I^2, $\sigma_{\delta S}$ in such a way that nearly all the information present in EC is stored into the sparse firing activity of the CA3 cells. The function Ξ has been defined as a step linearwise function (and not a threshold linear one, like for the other transfer functions) in order to replace one of the integrals in function Λ (defined in the Results section, eq. .(4.20)) with a sum over steps, thus decreasing the complexity of the numerical solution of the saddle point equations. The function $\Xi(x)$ has the following form:

1. When $-\infty < x < 0$, then $\Xi(x) = 0$.
2. When $m_l \leq x < m_{l+1}$ ($m_l \equiv l\ \mathrm{m_a}$), and $l = 0, \cdots, l_{\max} - 2$, then the function Ξ has the value $\Xi(x) \equiv \xi_l \equiv (l + \frac{1}{2})\mathrm{m_a}$.
3. When $m_{l_{\max}-1} \leq x < \infty$, then $\Xi(x) \equiv \xi_{l_{\max}-1} \equiv (l_{\max} + \frac{1}{2})\mathrm{m_a}$

By increasing the parameter $\mathrm{m_a}$, the number of steps $l_{\max}$ necessary to span a reasonable range of x decreases. We note that, for the sake of compactness of notations in the result section, we define also $m_{-1} = -\infty$ and $\xi_{-1} = 0$. The synaptic matrix is sparse as each CA3 cell receives inputs from D cells in EC.

$$d_{ik} \in \{0, 1\}, \qquad < d_{ik} > L = D\ . \tag{4.3}$$

- $\{\zeta_j\}$ are the firing rates produced in each cell j of CA1, *during the storage* of the CA3 representation; they are determined by the matrix multiplication of the pattern $\{\eta_i\}$ and $\{\nu_k\}$ respectively with the synaptic weights J_{ij}^S and K_{jk}^S followed by Gaussian distortion, (inhibition-dependent) thresholding and rectification:

$$\zeta_j = \left[\zeta_0 + \sum_{i=1}^{N} c_{ij} J_{ij}^S \eta_i + \sum_k b_{jk} K_{jk}^S \nu_k + \epsilon_j^S\right]^+$$

$$< (J_{ij}^S)^2 >= \sigma_J^2; \quad < (K_{jk}^S)^2 >= \sigma_K^2; \quad < (\epsilon_j^S) >= \sigma_{\epsilon S}^2 \tag{4.4}$$

(the rectifying function $[x]^+ = x$ for $x > 0$, and 0 otherwise, ensures that a firing rate is a positive quantity). The sparse synaptic matrix is fixed in this way:

$$< c_{ij} > N = C; \qquad < b_{jk} > L = B_j \tag{4.5}$$

The average number B_j of PP inputs to the CA1 cell j is taken to vary across different CA1 cells. (The average of B_j across cells is denoted as B). Anatomical studies (Witter 1993) reveal that restricted populations in the EC activate specific regions of CA1. This is consistent with the hypothesis (Treves and Rolls 1994) that the PP carries detailed information only on a *limited* number of elements of the episode.

- $\{V_i\}$ are the firing rates in the pattern retrieved from CA3, and they are taken to reproduce the $\{\eta_i\}$ with some Gaussian distortion (noise), followed by rectification:

$$\begin{aligned} V_i &= [\eta_i + \delta_i]^+ \\ &< (\delta_i)^2 >= \sigma_\delta^2 \end{aligned} \tag{4.6}$$

  σ_δ can be related e.g. to interference effects due to the loading of other memory patterns in CA3 (see below and Treves and Rolls, 1991). This and all the other noise terms are all taken to have zero means.

- $\{U_j\}$ are the firing rates produced in CA1 during retrieval:

$$\begin{aligned} U_j &= \left[U_0 + \sum_i c_{ij} J^R_{ij} V_i + \sum_k b_{jk} K^R_{jk} \nu_k + \epsilon^R_j\right]^+ , \\ < (\epsilon^R_j)^2 > &= \sigma^2_{\epsilon^R} . \end{aligned} \tag{4.7}$$

  It is important to note that, in the present version of the model, the PP connections contribute to the firing rates in CA1 during retrieval, eq (4.7), with the whole episode presented in the EC during storage. The possibility that only a limited number of units of the EC pattern are active in the retrieval phase, can be effectively and easily taken into account by switching off, with a quenched probability, a given fraction (*e.g.*, 3/4) of the b_{jk} connections in eq. (4.7) that were already switched on during storage. This latter modification of the model leads only to a sligth as easily computable modification of the analytical results (4.18). However, we decided to begin the numerical study of the solutions of the model from the simplest form (4.7).

The weights of the SC synaptic matrix during retrieval of a *specific* pattern,

$$J^R_{ij} = \cos(\theta_\mu) J^S_{ij} + \gamma_s^{1/2}(\theta_\mu) H(\eta_i, \zeta_j) + \sin(\theta_\mu) J^N_{ij} \tag{4.8}$$

consist of

1. the original weight during storage, J^S_{ij}, damped by a factor $cos(\theta_\mu)$, where $0 < \theta_\mu < \pi/2$ parametrizes the time elapsed between the storage and retrieval of pattern μ (μ is a shorthand for the pattern pentaplet $\{\nu_k, \eta_i, V_i, \zeta_j, U_j\}$).

2. the modification due to the storage of μ itself, represented by a Hebbian term $H(\eta_i, \zeta_j)$ – reflecting the association of patterns $\{\eta_i\}$ and $\{\zeta_j\}$ – also normalized so that

$$< (H(\eta, \zeta))^2 >= \sigma_J^2; \tag{4.9}$$

   γ_s measures the degree of *plasticity*, i.e. the mean square contribution of the modification of the Schaffer collaterals induced by one pattern, over the overall variance, across time, of the synaptic weight.

3. the superimposed modifications J^N reflecting the successive storage of new intervening patterns, again normalized such that

$$< (J^N_{ij})^2 >= \sigma_J^2. \tag{4.10}$$

In a symilar way, the weights of the PP synaptic matrix during retrieval of a specific pattern

$$K^R_{jk} = \cos(\theta_\mu)K^S_{jk} + \gamma_p^{1/2}(\theta_\mu)\widetilde{H}(\nu_k,\zeta_j) + \sin(\theta_\mu)K^N_{jk} \ , \tag{4.11}$$

have a similar structure, with the corresponding Hebbian term and the superimposed modifications again normalized as:

$$\begin{aligned} < \widetilde{H}(\nu_k,\zeta_j) > &= \sigma_K^2 & (4.12)\\ < (K^N_{jk})^2 > &= \sigma_K^2 & (4.13) \end{aligned}$$

in such a way that γ_p has the meaning of plasticity of the perforant path fibers.

The mean value of each synaptic weight has been collapsed with the threshold term (an approximation valid when the mean firing levels are strictly regulated by inhibition) and the gain of the threshold-linear transfer function (see Treves, 1990) has been set to one by rescaling the weights. It is convenient also to set for the Hebbian terms the specific form

$$\begin{aligned} H(\eta_i,\zeta_j) &= g(\zeta_j-\zeta_0)(\eta_i-\eta_0),\\ \widetilde{H}(\nu_k,\zeta_j) &= h(\zeta_j-\zeta_0)(\nu_k-\nu_0),\\ \nu_0 &= <\nu>_\nu \ . \end{aligned} \tag{4.14}$$

where the parameter $g = \frac{g'}{\sqrt{C}}$ and $h = \frac{h'}{\sqrt{B}}$ ensure the normalization given in Eqs. (4.9,4.12). Note that the presynaptic ans postsynaptic cell are taken to vary independently across memeory patterns, because of the extensive convergence of both SC and PP fibers; this implies that the variance in the synaptic weights is just a sum of terms from each memory pattern, and justify the interpretation of $1/\gamma_s$ and $1/\gamma_p$ as the effective number of patterns in storage in the Schaffer collaterals and perforant path to CA1 respectively.

The aim is to calculate how much, on average, of the information present in a given original pattern $\{\nu_k^\lambda\}$ is still present in the effective output of the system at the time λ is retrieved, i.e. in the pattern $\{U_j^\lambda\}$, that is to average the mutual information $i(\{\nu_k^\lambda\},\{U_j^\lambda\})$ over the *quenched*[3] variables $d_{ik}, b_{jk}, c_{ij}, l_{ik}, J^S_{ij}, J^N_{ij}, K^S_{ij}, K^N_{ij}$. The average over the quenched variables is necessary, because no meaning could possibly be assigned to a result specific to certain values of each of the quenched connections. Extensive quantities (*i.e.,* scaling proportionally to the number of units of the system), like the mutual information, are expected anyway to coincide with their average (Mezard *et al.* 1987).

[3] The term quenched in this study means independent of the specific distribution realized in pattern λ.

This amount of information has to be compared, for different values of the network parameters, to other quantities, like the entropy of the pattern presented in EC, and the information about that pattern that has been stored (or retrieved) in CA3. For the sake of brevity, we report only the results for the calculation of $i(\{\nu_k^\lambda\}, \{U_j^\lambda\})$, which is both the more difficult and more interesting calculation, but it will be clear from the discussion how to compute also the information stored or retrieved at intermediate stages of the network, like CA3.

4.3 Results of the analytical evaluation

In this section we report the results of the evaluation of the average mutual information $i(\{\nu_k^\lambda\}, \{U_j^\lambda\})$ between the given original EC pattern $\{\nu_k^\lambda\}$ and the CA1 output at the time λ is retrieved, i.e. the pattern $\{U_j^\lambda\}$. The details of the calculations are only sketched, pointing out the relevant facts and the techniques used in each step.

The first step of the calculation is the evaluation of the joint probability $P(\{\nu_k\}, \{U_j\})$. The latter quantity can be written (simplifying the notation) as

$$\begin{aligned} P(\nu, U) &= P(U \mid \nu)P(\nu) = \int_V \int_\zeta \int_\eta dV d\zeta d\eta P(U \mid V, \zeta, \eta, \nu)P(V \mid \eta, \nu) \\ &\times \; P(\zeta \mid \eta, \nu)P(\eta \mid \nu)P(\nu) \end{aligned} \tag{4.15}$$

where the different probability densities in (4.15) implement the model defined above.

The (average) amount of information is evaluated using the replica trick (Nadal and Parga, 1993; Treves 1995). This trick, often used in the context of the statistical physics approach to spin glass theory and neural networks (Mezard *et al.* 1987; Amit 1989), is based on writing the logarithm as a limit of a power:

$$\log(x) = \lim_{n \to 0} \frac{x^n - 1}{n} \tag{4.16}$$

The average of the logarithm of the probabilities is then calculated by introducing n replicas, performing the average over the quenched variables and taking at the end the $n \to 0$ limit. In this way one ends up with the expression:

$$\begin{aligned} < i(\nu, U) >_{d,b,c,I,J^S,J^N,K^S,K^N} &= \lim_{n \to 0} \frac{1}{n} < \int d\nu dU P(\nu, U) \left\{ \left[\frac{P(\nu, U)}{P(\nu)} \right]^n \right. \\ &- \left. \Big[P(U) \Big]^n \right\} >_{d,b,c,I,J^S,J^N,K^S,K^N} . \end{aligned} \tag{4.17}$$

where one needs to introduce $n+1$ replicas of the variables $\delta_i, \delta_i^S, \epsilon_j^S, \epsilon_j^R, V_i, \zeta_j, \eta_i$, and, for the second term in curly brackets only, ν_k. Then, the integrals in (4.17) are evaluated, in the limit of infinite number of neurons[4], by the saddle point method. In order to find

[4] The numbers of neurons L, N, M contained respectively in EC, CA3, CA1 are supposed to scale to infinity with a fixed ratio.

the saddle points, we then use the 'replica symmetry ansatz', which consists in imposing that both single and double-replica saddle point parameters do not depend on the replica index. Only at the end, after the large N limit, the $n \rightarrow 0$ limit is performed [5].

This procedure leads to the following expression:

$$\begin{aligned}
< i(U,\nu) > &= \mathrm{extr}_{y_A,\tilde{y}_A,w_A,\tilde{w}_A,z_A,\tilde{z}_A}\Big\{\sum_j \Gamma(y_A,w_A,z_A,q^0,B_j,\gamma_s,\gamma_p) \\
&- \frac{N}{2}\Big(y_A\tilde{y}_A + 2w_A\tilde{w}_A + z_A\tilde{z}_A\Big) + N\ \Lambda(\tilde{y}_A,\tilde{w}_A,\tilde{z}_A,q^0)\Big\} \\
&- \mathrm{extr}_{q_B,\tilde{q}_B,y_B,\tilde{y}_B,w_B,\tilde{w}_B,z_B,\tilde{z}_B}\Big\{\sum_j \Gamma(y_B,w_B,z_B,q_B,B_j,\gamma_s,\gamma_p) \\
&- \frac{N}{2}(y_B\tilde{y}_B + 2w_B\tilde{w}_B + z_B\tilde{z}_B) - \frac{L}{2}q_B\tilde{q}_B \\
&+ L\int_{-\infty}^{\infty}\Delta s < e^{-\frac{\tilde{q}_B\nu^2}{2} - s\sqrt{\tilde{q}_B}\nu} >_\nu \ln < e^{-\frac{\tilde{q}_B\nu^2}{2} - s\sqrt{\tilde{q}_B}\nu} >_\nu \\
&+ N\ \Lambda(\tilde{y}_B,\tilde{w}_B,\tilde{z}_B,q_B)\Big\} \qquad (4.18)
\end{aligned}$$

where

$$< (\cdot) >_\nu = \int d\nu P_\nu(\nu)(\cdot) \ , \qquad (4.19)$$

and taking the extremum means evaluating each of the two terms, separately, at a saddle-point over the variables indicated (and dividing by ln 2 to yield a result in bits). Let us now explain what are the symbols appearing in our result (4.18). The functionΛ is given by:

$$\begin{aligned}
\Lambda(\tilde{y},\tilde{w},\tilde{z},q) &= \int \frac{ds dr d\tau}{(2\pi)^{3/2}\sqrt{(\tilde{y}\tilde{z} - \tilde{w}^2)\sigma_I^2 Dq}} \exp -\frac{(\tau - \eta_0)^2}{2\sigma_I^2 Dq} \\
&\times\ F\left[\ln F + \frac{1}{2}\begin{pmatrix} s & r \end{pmatrix}\begin{pmatrix} y & w \\ w & z \end{pmatrix}^{-1}\begin{pmatrix} s \\ r \end{pmatrix}\right] \qquad (4.20)
\end{aligned}$$

and F has the following expression

$$\begin{aligned}
F(r,s,\tau,\tilde{y},\tilde{w},\tilde{z},q) &= \sum_{l=-1}^{l_{\max}-1}\left[\phi\left(\frac{m_{l+1} - \tau}{\sqrt{\sigma_{\delta_S}^2 + \sigma_I^2 D(q^0 - q)}}\right) - \phi\left(\frac{m_l - \tau}{\sqrt{\sigma_{\delta_S}^2 + \sigma_I^2 D(q^0 - q)}}\right)\right] \\
&\times \left\{\phi\left[\frac{\xi_l - \sigma_\delta^2(s + \tilde{w}\xi_l)}{\sigma_\delta\sqrt{1 + \sigma_\delta^2\tilde{y}}}\right]\frac{1}{\sqrt{1 + \sigma_\delta^2\tilde{y}}}\exp -\frac{[s + \xi_l(\tilde{w} + \tilde{y})]^2}{2\tilde{y}(1 + \sigma_\delta^2\tilde{y})}\right.
\end{aligned}$$

[5] Problems arising with the replica symmetry ansatz and with the inversion of the two limits $n \rightarrow 0$ and $N \rightarrow \infty$ are discussed *e.g.*, in Mezard *et al.* 1995; For the range of validity of the replica symmetry ansatz in networks of threshold linear units see Treves (1991)

$$
\begin{aligned}
&+ \quad \phi\left[\frac{-\xi_l}{\sigma_\delta}\right] \exp - \frac{[\tilde{w}\xi_l + s]^2}{2\tilde{y}} \Bigg\} \\
&\times \quad \exp - \frac{[\xi_l(\tilde{y}\tilde{z} - \tilde{w}^2) - (\tilde{w}s - \tilde{y}r)]^2}{2\tilde{y}(\tilde{y}\tilde{z} - \tilde{w}^2)}
\end{aligned}
\tag{4.21}
$$

and the following notations for the gaussian integration measure and the error function are introduced:

$$
\Delta s \equiv (ds/\sqrt{2\pi}) \exp -s^2/2 \qquad \phi(x) \equiv \int_{-\infty}^{x} \Delta s. \tag{4.22}
$$

Γ is effectively an entropy term for the CA1 activity distribution, given by

$$
\begin{aligned}
\Gamma(y, w, z, q, B_j, \gamma_s, \gamma_p) &= \int \frac{ds_1 ds_2}{2\pi\sqrt{\det \mathbf{T}'}} \exp - \begin{pmatrix} s_1 & s_2 \end{pmatrix} \frac{(\mathbf{T}')^{-1}}{2} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \\
&\times \left[\int_{-\infty}^{0} dU G(U) \ln \int_{-\infty}^{0} dU' G(U') \right. \\
&+ \left. \int_{0}^{\infty} dU G(U) \ln G(U) \right],
\end{aligned}
\tag{4.23}
$$

where

$$
\begin{aligned}
G(U) &= G(U; s_1, s_2, y, w, z, q, B_j, \gamma_s, \gamma_p) \\
&= \phi\left[\frac{(\zeta_0 - s_2)(T_y + 2\gamma_j T_w + \gamma_j^2 T_z) + (U - U_0 + s_1 + \gamma_j s_2)(T_w + \gamma_j T_z)}{\sqrt{(T_y T_z - T_w^2)(T_y + 2\gamma_j T_w + \gamma_j^2 T_z)}}\right] \\
&\quad \times \frac{1}{\sqrt{2\pi(T_y + 2\gamma_j T_w + \gamma_j^2 T_z)}} \exp - \frac{(U - U_0 + s_1 + \gamma_j s_2)^2}{2(T_y + 2\gamma_j T_w + \gamma_j^2 T_z)} \qquad (4.24) \\
&+ \phi\left[\frac{-(\zeta_0 - s_2)T_y - (U - U_0 + s_1 + \gamma_j \zeta_0)T_w}{\sqrt{(T_y T_z - T_w^2)T_y}}\right] \\
&\quad \times \frac{1}{\sqrt{2\pi T_y}} \exp - \frac{(U - U_0 + s_1 + \gamma_j \zeta_0)^2}{2T_y}, \qquad (4.25)
\end{aligned}
$$

and

$$
\begin{aligned}
T_y &= \sigma_{\epsilon R}^2 + \sigma_J^2 C(y^0 - y) + \sigma_K^2 B_j (q^0 - q) \\
T_w &= \cos(\theta)[\sigma_J^2 C(w^0 - w) + \sigma_K^2 B_j (q^0 - q)] \\
T_z &= \sigma_{\epsilon S}^2 + \sigma_J^2 C(z^0 - z) + \sigma_K^2 B_j (q^0 - q) \qquad (4.26) \\
\mathbf{T}_j' &= \begin{pmatrix} \tau_y & \tau_w \\ \tau_w & \tau_z \end{pmatrix} \\
\tau_y &= \sigma_J^2 C y + \sigma_K^2 B_j q \\
\tau_w &= \cos(\theta)[\sigma_J^2 C w + \sigma_K^2 B_j q] \\
\tau_z &= \sigma_J^2 C z + \sigma_K^2 B_j q
\end{aligned}
$$

are effective noise terms.

$$\gamma_j = C\gamma_s^{1/2}gx^0 + B_j\gamma_p^{1/2}ht^0 \tag{4.27}$$

y, w, z, q are saddle-point parameters (conjugated to $\tilde{y}, \tilde{w}$, $\tilde{z}$ and $\tilde{q}$), and $x^0, y^0, w^0, z^0, t^0, q^0$ are corresponding single-replica parameters fixed, essentially by the normalization of the probabilities, as

$$\begin{aligned}
x^0 &= \frac{1}{N}\sum_i < (\eta_i - \eta_0)V_i >= \sum_{l=-1}^{l_{\max}-1}\left[\phi\left(\frac{m_{l+1}-\eta_0}{\sqrt{\sigma_{\delta_S}^2+\sigma_I^2 Dq^0}}\right) - \phi\left(\frac{m_l-\eta_0}{\sqrt{\sigma_{\delta_S}^2+\sigma_I^2 Dq^0}}\right)\right] \\
&\times \quad (\xi_l - \eta_0)\left[\xi_l\phi\left(\frac{\xi_l}{\sigma_\delta}\right) + \frac{\sigma_\delta}{\sqrt{2\pi}}\exp -\frac{1}{2}\left(\frac{\xi_l}{\sigma_\delta}\right)^2\right] \\
y^0 &= \frac{1}{N}\sum_i < V_i^2 >= \sum_{l=-1}^{l_{\max}-1}\left[\phi\left(\frac{m_{l+1}-\eta_0}{\sqrt{\sigma_{\delta_S}^2+\sigma_I^2 Dq^0}}\right) - \phi\left(\frac{m_l-\eta_0}{\sqrt{\sigma_{\delta_S}^2+\sigma_I^2 Dq^0}}\right)\right] \\
&\times \quad \left\{\left[\sigma_\delta^2 + \xi_l^2\right]\phi\left(\frac{\xi_l}{\sigma_\delta}\right) + \frac{\xi_l\sigma_\delta}{\sqrt{2\pi}}\exp -\frac{1}{2}\left(\frac{\xi_l}{\sigma_\delta}\right)^2\right\} \\
w^0 &= \frac{1}{N}\sum_i < \eta_i V_i >= \sum_{l=-1}^{l_{\max}-1}\left[\phi\left(\frac{m_{l+1}-\eta_0}{\sqrt{\sigma_{\delta_S}^2+\sigma_I^2 Dq^0}}\right) - \phi\left(\frac{m_l-\eta_0}{\sqrt{\sigma_{\delta_S}^2+\sigma_I^2 Dq^0}}\right)\right] \\
&\times \quad \xi_l\left[\xi_l\phi\left(\frac{\xi_l}{\sigma_\delta}\right) + \frac{\sigma_\delta}{\sqrt{2\pi}}\exp -\frac{1}{2}\left(\frac{\xi_l}{\sigma_\delta}\right)^2\right] \\
z^0 &= \frac{1}{N}\sum_i \eta_i^2 = \sum_{l=-1}^{l_{\max}-1}\left[\phi\left(\frac{m_{l+1}-\eta_0}{\sqrt{\sigma_{\delta_S}^2+\sigma_I^2 Dq^0}}\right) - \phi\left(\frac{m_l-\eta_0}{\sqrt{\sigma_{\delta_S}^2+\sigma_I^2 Dq^0}}\right)\right]\xi_l^2 \\
t^0 &= < (\nu - \nu_0)\nu >_\nu \\
q^0 &= < \nu^2 >_\nu
\end{aligned} \tag{4.28}$$

4.3.1 Model parameters and experimental quantities

Since the model we are studying is supposed to reproduce the behaviour of a real neural network, its parameters should be in some way related to experimental quantitities. This subsection is devoted to a brief discussion of how the parameters can be measured or fixed in a reasonable way.

Very important parameters included in the model are γ_s and γ_p, which represent, as stated above, the degree of plasticity of the Schaffer and perforant path connections respectively, expressed as the ratio between the mean square change in synaptic strength due to the storage of one memory pattern, and the overall variance in synaptic strength. If such variance is entirely due to memory storage, one can say, inversely, that the memory system holds of the order of γ_s^{-1} (or γ_p^{-1})patterns at any one time. This is only a rough measure, as in fact patterns are not necessarily ever completely effaced in the model, but rather their traces may be only gradually overwritten by other intervening patterns, depending on the chosen dependence of θ on real time, as parametrized by the factor $\cos(\theta)$. One interesting model of the time-dependence of the plasticity parameters could be the gradual decay of memory traces used in Treves (1995).

The evaluation has been carried out for an arbitrary distribution P_ν, the only hypothesis being that of independence among different EC cells. For the numerical evaluation of the saddle point equations, one can choose, for simplicity, a binary form, in which a cell is firing at a rate ν^* with probability a, and silent otherwise

$$P_\nu(\nu) = (1-a)\delta(\nu) + a\delta(\nu - \nu^*) \tag{4.29}$$

where the firing rate of all EC active cells has been set to ν^*, and a is a *sparse coding* parameter (3.13). In alternative, more complex form, *e.g.*, that fit real distributions observed in EC cells, can be chosen.

Let us now discuss the noise parameters. σ_δ is measured on the η^* scale (i.e., as a frequency), and is chosen to account both for actual noise in retrieval from CA3 (which corresponds to the standard deviation in the rates recorded during successive trials of an already learned task, as e.g. in Rolls *et al*, 1989; *fast* noise in thermodynamics jargon), and for interference (the so-called *quenched* noise) caused by memory loading (which would be observable by comparing responses during and after one-shot learning, and which most theoretical models would predict to grow with the square root of the load; Treves, 1990). Moreover, CA1 rates are sensitive to fast noise of s.d. $\sigma_{\epsilon^S}, \sigma_{\epsilon^R}$, again measured in Hz. Data on trial to trial variability in firing rates in CA3 and CA1 are presently too scarce to allow a systematic analysis; but these model parameters can be set to reproduce experimental data as it becomes available. In contrast, the noise of s.d. σ_{δ^S}, present in CA3 during storage, is not a parameter with an experimental correlate, and, as discussed in the previous section, should be chosen to be very small, in order to store in CA3 as much as possible of the information presented in CA3.

The threshold terms ζ_0, U_0, also measured as frequencies (given the unit gain), are given negative values such as to produce activity distributions in CA1 with a given sparseness. These parameters, which summarize a variety of effects and have no immediate correlate, can be given appropriate values inferred indirectly from the measured sparseness (see *e.g.*, Barnes *et al.* 1990). The same applies to the parameters characterizing the η distribution (mainly the function $\Xi(\cdot)$).

For the purpose of the numerical study of the saddle point equations, the three variances of the synaptic efficacies $\sigma_I^2, \sigma_J^2, \sigma_K^2$ can be set to the same value. The relative abundance (or estimated strength) of the SC and PP connections can be tuned by means of the mean connectivity parameters B, C, with the constraint

$$\sigma_J^2(B+C) = 1 \ . \tag{4.30}$$

The latter requirement is just set to be consistent with the gain of the CA1 transfer function being set to 1. The value for the connectivity parameter D for the "fake" connections EC-to-CA3 plays a minor role, as far as the ratio between the values of $\sigma_I^2 D$ and $\sigma_{\delta S}$ is such as to achieve efficient information storage in CA3.

Finally, the ratios M/L and N/L between cells in different regions can be selected looking at anatomical data. The number L of EC cells is itself irrelevant if one considers only the amount of information per EC cell; moreover L is considered strictly infinite in the analytical evaluation.

4.3.2 Numerical solutions of the saddle point equations.

It is self-evident that the numerical study of the saddle point equations introduced in the last sections is quite hard, and we were not able to study, so far, a region of the parameter space big enough to get useful insights on the functions of the perforant path to CA1. Nevertheless, let us briefly describe the numerical technique that we (*i.e.*, mainly Carlo Fulvi Mari) are using.

One of the problems in evaluating (4.18) is the large dimensionality of the space in which one has to find the saddle point (6 dimensions for the first term, and 8 dimensions for the second saddle point). However, the dimensionality of the saddle point parameter space can be effectively reduced, for the purpose of the numerical evaluation, by using a trick that we describe focusing on the first term in (4.18) (that with the A subscripts on the saddle point variables). We start with an initial guess for the values of (y_A, w_A, z_A) (let us denote the values by (y_A^n, w_A^n, z_A^n)). Then we calculate the corresponding values for the 'tilde' parameters $(\tilde{y}_A^n, \tilde{w}_A^n, \tilde{z}_A^n)$ by evaluating the derivatives of the function Γ in (y_A^n, w_A^n, z_A^n). Then we use $(\tilde{y}_A^n, \tilde{w}_A^n, \tilde{z}_A^n)$ to compute (deriving this time Λ) the new, more refined guess for the (y_A, w_A, z_A). The procedure is iterated until we reach the fixed point for (y_A, w_A, z_A). In this way the effective dimensionality of the saddle point parameters is reduced by a factor of 2.

Currently, the numerical investigation of saddle point equations is devoted to the study of the dependence of information on certain parameters, like the relative abundance of Schaffer and perforant path connections. In particular, we are studying the way that information varies with B, that, having fixed the condition (4.30) and certain values for the noise parameters, parametrizes the relative strength of perforant path versus Schaffer connections. The comparison between the information relayed by the CA1 cells when B is very low, and when B is of order of what is found in real hippocampus, can shed light on the specific contribution of the perforant path projections to the CA1 processing.

4.4 Discussion

We conclude by discussing both how the model could be improved and which assumptions and results could be tested by experiments, leading to an improvement of our understanding of the hippocampus from the viewpoint discussed here.

4.4.1 How good is the model ?

The biological plausibility of the model has been discussed before. Here we point at the improvements of the model that can be carried out, and are in fact under investigation:

- At the moment we are numerically studying the simplest case in which the average number of perforant path connections per CA1 cell (B_j) is constant along the CA1 field. It would be interesting to take instead into account the real topographic organization of the perforant path to CA1 (Witter 1993).
- A similar problem is the study of the effects of taking into account the real structure of the Schaffer collateral connectivity, and in particular of the fact that the convergence from CA3 to CA1 is not constant along the length of CA1 field. This issue is under investigation by Simon Schultz (Schultz 1996).
- A problem which requires an improvement of the analytical techniques, but is interesting, is the introduction of spatially correlated input patterns in EC, both to fit what is found with real data and to understand how the redundancy of messages varies at different stages of hippocampal information processing.
- The fact that the dentate gyrus is not modelled realistically is in our opinion a minor point, as the understanding of its function is not the goal of the present study, and the hypothesis on its role can be studied as a separate issue (Treves and Rolls 1992).

4.4.2 Relating analytical models to experimental measures

We conclude by discussing how the underlying hypotheses and the predictions of the model can be tested and related to the behaviour of hippocampus *in vivo*, as observed by recording the activity of its units. From our viewpoint, the combination of a quantitative analysis of real data and a realistic model can greatly improve our understanding of the hippocampus. Most important directions to take:

- The comparison between data recorded from different populations (*e.g.,* Barnes *et al.* 1990; Treves *et al.* 1996) has already been useful, but more systematic and quantitative analyses of parallel recording are needed to quantify the informational properties of each stage of the system.
- It would be also insightful to observe if there are systematic differences in the information content of hippocampal cells in different experimental conditions. If the variations in external conditions can be related in some way to variations in the noise parameters of a model like ours [6], one can perform severe tests of the model.

[6] *e.g.,* in the experiment about view cells in primate hippocampus, chapter 3, one can imagine varying the noise of retrieval by using curtains in order to present only a partial visual cue and check to which extent the response is invariant in different parts of hippocampus. This procedure, devised by Rolls, Robertson and Georges-Francois 1996, is now under experimental study.

- The above model is fully based on the assumption that information is represented basically in the *firing rate* of the cells. Although there is some evidence that in the rat hippocampus the temporal modulation can play some role (*e.g.,* O'Keefe and Recce 1993), this may be not true for the primate hippocampus. The study reported in chapter 3 is in our opinion an interesting starting point for understanding this important issue, and is fully consistent with the hypothesis that firing rate describe well the neuronal representation of the external correlates.

Appendix A

Explicit evaluation of the bias

In this appendix we give the derivation of the results presented in chapter 2. In the calculation we consider the case in which the data are treated by convolving responses with a kernel distribution and then by discretizing the response space into R intervals. Finally, however, we show how to recover the results appropriate to the other data manipulations, namely pure discretization and pure convolution with continuous distributions. Moreover, we explain why the bias evaluation, given by Carlton (1969), and often quoted in the literature, is wrong.

We start by calculating the average of the total amount of information (2.15), which can be expressed as follows:

$$< \tilde{I}_N^D >= \sum_{s\in\mathcal{S}} \widehat{\sum_i} < p_N(s)\tilde{p}_N(i|s)\log_2 \tilde{p}_N(i|s) > - < \widehat{\sum_i} \tilde{p}_N(i)\log_2 \tilde{p}_N(i) > \qquad \text{(A.1)}$$

where $\tilde{p}(\cdot)$ is defined in (2.16) and the hat on the sum over response bins in (A.1) denotes that we must exclude from that sum, for each term of the sum over stimuli, the bins in which $\tilde{p}(i|s) = 0$ (in fact, in those bins, the only permitted outcome is $\tilde{p}_N(i|s) = 0$ and they trivially disapper from the average)). Now we can use the following series expansion for the logarithm:

$$-\log_2(\tilde{p}_N(\cdot)) = \frac{1}{\log 2}\sum_{j=1}^{\infty}\frac{(1-\tilde{p}_N(\cdot))^j}{j}\,. \qquad \text{(A.2)}$$

This expansion (A.2) is convergent for all values of $\tilde{p}_N(\cdot)$, since $0 < \tilde{p}_N(\cdot) \le 1$ (note that in our calculation the configuration $\tilde{p}_N(\cdot) = 0$ can be excluded). Taking term by term expectations in (A.1) we find:

$$\begin{aligned} < \tilde{I}_N^D > &= \frac{-1}{\log 2}\sum_{s\in\mathcal{S}}\widehat{\sum_i}\sum_{j=1}^{\infty} < p_N(s)\tilde{p}_N(i|s)\frac{(1-\tilde{p}_N(i|s))^j}{j} > \\ &+ \frac{1}{\log 2}\widehat{\sum_i}\sum_{j=1}^{\infty} < \tilde{p}_N(i)\frac{(1-\tilde{p}_N(i))^j}{j} > \\ &= \frac{-1}{\log 2}\sum_{s\in\mathcal{S}}\widehat{\sum_i}\sum_{j=1}^{\infty}\sum_{k=0}^{j}\frac{(-1)^k}{j}\binom{j}{k} < p_N(s)\tilde{p}_N^{k+1}(i|s) > \\ &+ \frac{1}{\log 2}\widehat{\sum_i}\sum_{j=1}^{\infty}\sum_{k=0}^{j}\frac{(-1)^k}{j}\binom{j}{k} < \tilde{p}_N^{k+1}(i) > \end{aligned} \qquad \text{(A.3)}$$

where in the last step we used the binomial decomposition for $(1-\widetilde{p}(\cdot))^j$. We can now calculate the average by the following procedure. First we average over responses (at fixed stimulus s and number of presentations per stimulus $N_s = Np_N(s)$) simply by assuming that the probability of obtaining a raw response r (given the stimulus s) is given by $P(r|s)dr$, and by substituting the sum over outcomes with the corresponding (correctly normalized) integral in the response space. We are then left with an average over $p_N(s)$, with a multinomial distribution. Note that, in averaging terms of the form $<(\widetilde{p}_N(i))^k>$, since the parameters specifying the kernel can be stimulus dependent, we must decompose $\widetilde{p}_N(i)$ as $\widetilde{p}_N(i) = \sum_s p_N(s)\widetilde{p}_N(i|s)$, average first over the responses (at fixed stimulus) and finally over $p_N(s)$ with the multinomial distribution. In this way we obtain the following expressions:

$$\begin{aligned} <\widetilde{p}_N^k(i|s)> &= \widetilde{p}^k(i|s) + \frac{1}{N_s}\binom{k}{2}\widetilde{p}^{k-2}(i|s)\left[\widetilde{q}(i|s) - \widetilde{p}^2(i|s)\right] + o\left(\frac{1}{N_s\widetilde{p}(i|s)}\right) && \text{(A.4)} \\ <p_N(s)\widetilde{p}_N^k(i|s)> &= p(s)\widetilde{p}^k(i|s) + \frac{1}{N}\binom{k}{2}\widetilde{p}^{k-2}(i|s)\left[\widetilde{q}(i|s) - \widetilde{p}^2(i|s)\right] + o\left(\frac{1}{Np(s)\widetilde{p}(i|s)}\right) && \text{(A.5)} \\ <\widetilde{p}_N^k(i)> &= \widetilde{p}^k(i) + \frac{1}{N}\binom{k}{2}\widetilde{p}^{k-2}(i)\left[\widetilde{q}(i) - \widetilde{p}^2(i)\right] + o\left(\frac{1}{N\widetilde{p}(i)}\right) && \text{(A.6)} \end{aligned}$$

where $\widetilde{q}(\cdot)$ is defined in (2.23). Ignoring the third term in each of the (A.4)-(A.6) and then substituting (A.4)-(A.6) into (A.3), we find an exact expression for the bias which is exact up to $O(1/N^2)$ terms and is a good approximation to the bias if in each bin $N_s\widetilde{p}(i|s) \ll 1$:

$$\begin{aligned} <\widetilde{I}_N^D> \simeq\ & \frac{-1}{\log 2}\sum_{s\in\mathcal{S}}\widehat{\sum}_i\sum_{j=1}^{\infty}\frac{1}{j}\left[1-\widetilde{p}(i|s)\right]^{j-2}\Big\{\widetilde{p}(i|s)\left[1-\widetilde{p}(i|s)\right]^2 \\ +\ & \frac{1}{2N}\left[\widetilde{q}(i|s) - \widetilde{p}^2(i|s)\right]\left[j(j-1)\widetilde{p}(i|s) - 2j(1-\widetilde{p}(i|s))\right]\Big\} \\ +\ & \frac{1}{\log 2}\widehat{\sum}_i\sum_{j=1}^{\infty}\frac{1}{j}\left[1-\widetilde{p}(i)\right]^{j-2}\Big\{\widetilde{p}(i)\left[1-\widetilde{p}(i)\right]^2 \\ -\ & \frac{1}{2N}\left[\widetilde{q}(i) - \widetilde{p}^2(i)\right]\left[j(j-1)\widetilde{p}(i) - 2j(1-\widetilde{p}(i))\right]\Big\} \\ \equiv\ & \widetilde{I}^D + \widetilde{C}_1^D \end{aligned} \tag{A.7}$$

where $\widetilde{I}^D$ is given in (2.19) and $\widetilde{C}_1^D$ is the leading contribution to the bias:

$$\widetilde{C}_1^D = \frac{1}{2N\log 2}\left\{\widehat{\sum}_i\left[\left(\sum_{s\in\mathcal{S}}\frac{\widetilde{q}(i|s)}{\widetilde{p}(i|s)}\right) - \frac{\widetilde{q}(i)}{\widetilde{p}(i)}\right] - (S-1)\right\} \tag{A.8}$$

By going further in the $1/N$ expansion when considering the averages (A.4)-(A.6), one can also obtain the next terms in the $1/N$ expansion of the bias by the same procedure. Here we report only the results for the second term:

$$\begin{aligned} \widetilde{C}_D^2 =\ & \frac{1}{12N^2\log 2}\left\{\sum_{s\in\mathcal{S}}<p_N^{-1}(s)>\left[\widehat{\sum}_i\frac{-2\widetilde{p}(i|s)\widetilde{t}(i|s) + 3\widetilde{q}^2(i|s)}{\widetilde{p}^3(i|s)} - 1\right]\right\} \\ -\ & \frac{1}{12N^2\log 2}\left\{\left[\widehat{\sum}_i\frac{-2\widetilde{p}(i)\widetilde{t}(i) + 3\widetilde{q}^2(i)}{\widetilde{p}^3(i)} + 1\right]\right\} \end{aligned} \tag{A.9}$$

where

$$\widetilde{t}(i|s) \equiv \int drP(r|s)E_i^3(i|s) \qquad \widetilde{t}(i) \equiv \sum_{s\in\mathcal{S}}p(s)\widetilde{t}(i|s)\ . \tag{A.10}$$

Higher order terms are reported, for the discrete case, in Treves and Panzeri (1995). In fact (A.8) is derived (as in the leading term in the bias) under the condition that $N_s\widetilde{p}(i|s) \gg 1$ in each interval; whereas by inspecting the higher order expansion terms, one can, as mentioned in Treves and Panzeri (1995), expect them to be successively smaller (and negligible with respect to $\widetilde{C}_1^D$), under the less stringent condition $\widetilde{C}_1^D \ll 1$. Therefore, the higher order corretions are, in any case, close to negligible whenever $\widetilde{C}_1^D$ is a good approximation for the bias. When this is not the case, because the condition $N_s\widetilde{p}(i|s) \gg 1$ is severely violated, computer simulations indicate that taking higher order corrections into account (which is itself not easy), does not help; on the contrary, in such a low-N regime in which $\widetilde{C}_1^D$ is often already too large, the next terms become huge and signal the breakdown of the expansion procedure (Treves and Panzeri 1995; Strong *et al.* 1996).

If one is interested in measuring, instead of the averaged transmitted information, the conditional transmitted information, relative to a given stimulus s, a similar calculation can be performed to obtain the bias of this quantity. The main technical step which is different is that when calculating $< \widetilde{I}(s)^D >$ from (2.15),

$$< \widetilde{I}_N^D(s) >= \widehat{\sum}_i < \widetilde{p}(i|s) \log_2 \widetilde{p}_N(i|s) > -\widehat{\sum}_i < \widetilde{p}_N(i|s) \log_2 \widetilde{p}_N(i) > \tag{A.11}$$

after using the expansion (A.2) for the logarithm, one has to calculate the average of $< \widetilde{p}_N(i|s)\widetilde{p}_N^k(i) >$ up to the next-to-leading order:

$$\begin{aligned} < \widetilde{p}_N(i|s)\widetilde{p}_N^k(i) > &= p_N(i|s)\widetilde{p}^k(i) + \frac{1}{N}\binom{k}{2}\widetilde{p}^{k-1}(i)\widetilde{p}(i|s)\,[1-\widetilde{p}(i)] \\ &+ \frac{k}{N}\,[1-\widetilde{p}(i|s)]\,\widetilde{p}(i|s)\widetilde{p}^{k-1}(i) + o\left(\frac{1}{Np(s)\widetilde{p}(i|s)}\right) \end{aligned} \tag{A.12}$$

Our result, again valid when $N_s\widetilde{p}(i|s) \gg 1$ in each interval, is now expressed as:

$$< \widetilde{I}_N^D(s) > -\widetilde{I}^D(s) \simeq \widetilde{C}_1^D(s) \tag{A.13}$$

with

$$\begin{aligned} \widetilde{C}_1^D(s) &= \frac{1}{N\log 2}\widehat{\sum}_i \left\{ < p_N^{-1}(s) > \frac{\widetilde{q}(i|s)-\widetilde{p}^2(i|s)}{2\widetilde{p}(i|s)} + \frac{\widetilde{p}^2(i|s)-\widetilde{q}(i|s)}{\widetilde{p}(i)} \right\} \\ &+ \frac{1}{2N\log 2}\widehat{\sum}_i \left\{ \frac{\widetilde{q}(i)\widetilde{p}(i|s)-\widetilde{p}(i|s)\widetilde{p}^2(i)}{\widetilde{p}^2(i)} \right\} \end{aligned} \tag{A.14}$$

The *discrete case* (for which the results are fully reported in section 2.1.1) can be easily derived by choosing a Gaussian as kernel function and then taking the limit of zero convolution width. In this case, it is easy to show from (A.8) that the leading bias term takes the form:

$$\begin{aligned} C_1^D &= \frac{1}{2N\log 2}\left\{ \sum_{s\in\mathcal{S}}\widehat{\sum}_i [1-p(i|s)] - \widehat{\sum}_i [1-p(i)] \right\} \\ &= \frac{1}{2N\log 2}\left\{ \sum_{s\in\mathcal{S}}\widetilde{R}_s - \widetilde{R} - (S-1) \right\}\,. \end{aligned} \tag{A.15}$$

It should be noted that in the discrete case the following evaluation of the bias of the nutual information was derived by Carlton (1969):

$$\begin{aligned} < I_N^D > -I^D &\simeq -\widehat{\sum}_i \left\{ \log_2\left(1+\frac{1-p(i)}{Np(i)}\right) - \frac{1}{2N\log 2}\frac{p(i)[1-p(i)(N-1)]}{(Np(i)+1-p(i))^2} \right\} \\ &+ \sum_{s\in\mathcal{S}}\widehat{\sum}_i \left\{ \log_2\left(1+\frac{1-p(i|s)}{N_s p(i|s)}\right) \right. \\ &- \left. \frac{1}{2N_s\log 2}\frac{p(i|s)[1-p(i|s)(N_s-1)]}{(N_s p(i|s)+1-p(i|s))^2} \right\} \end{aligned} \tag{A.16}$$

The expansion in $1/N$ of (A.16), agrees with our expression (2.6) up the $1/N$ order, but is very different form the real bias expansion when going to higher orders. This is because the procedure employed by Carlton to derive the result (A.16) uses the expansion (A.2) for the logarithm and takes term by term expectations by truncating averages of powers of $p(\cdot)$ to the next-to-leading order, as in (A.4-(A.6), but with a trick (valid only in the discrete case) used to obtain (without going further in $1/N$ in the evaluation of the averages (A.4-(A.6)) a *partial* re-summation (to all orders in $1/N$) of the complete bias expression. This partial re-summation, however, is of dubious value from the conceptual point of view and gives utterly nonsensical results when checked numerically. In fact, for example, by using the correction term (A.16) in the simulation reported in figure 2.3 , we have obtained an estimate of the bias much larger than the raw information in the N_s range 8-128.

The *continuum limit*, results for which are presented in section 2.1.4, can be reached when $R \to \infty$, as follows. Let us denote some typical size of the response by ρ (taken here to be uni-dimensional) and let us introduce the following succession of infinite discretizations, indexed by n, into intervals $R_{i;n}$ ($i = 0, \pm 1. \pm 2, \cdots$ labels each interval):

$$R_{i;n} \equiv \left\{ r; \frac{i}{2^n}\rho \le r < \frac{i+1}{2^n}\rho \right\} . \tag{A.17}$$

The discrete probabilities (2.18) have the form:

$$\widetilde{p}_n(i|s) \equiv \int_{R_{i;n}} dr \widetilde{P}(r) . \tag{A.18}$$

By introducing the function

$$\Gamma_n(r) \equiv \frac{2^n}{\rho}\widetilde{p}_n(i) \qquad \text{for } r \in R_{i;n} \tag{A.19}$$

we have the identity:

$$\widetilde{p}_n(i) \log_2 \left(\frac{2^n}{\rho}\widetilde{p}_n(i) \right) = \int_{R_{i;n}} \Gamma_n(r) \log_2 \Gamma_n(r) dr \tag{A.20}$$

from which we can derive

$$\sum_{i,s} \widetilde{p}_n(i|s) \log_2 \frac{\widetilde{p}_n(s,i)}{p(s)\widetilde{p}_n(i)} = \sum_{s \in S} \int dr \Gamma_n(s,r) \log_2 \frac{\Gamma_n(s,r)}{p(s)\Gamma_n(r)} . \tag{A.21}$$

Now, with the hypotheses that $\widetilde{P}(r), \widetilde{P}(r|s)$ are bounded and continuous almost everywhere (Ihara 1993), we have that in the $n \to \infty$ limit $\Gamma(r|s) \to \widetilde{p}(r|s)$ and in the same limit the (infinitely) discretized information (A.21) tends to the continuous one (2.29), whereas the infinitely discretized term (A.8) tends to that derived in the continuous case (2.33).

Acknowledgments

The results presented in the thesis are the outcome of a large collaboration; thus I feel that I must start thanking the people who directly contributed to this project.

First of all, Alessandro Treves. I was lucky to have such an excellent thesis advisor, who provided a constant guidance over all my three years in SISSA, and introduced me gradually into the world of neuroscience. Then Edmund Rolls, who motivated, with his experimental and theoretical work, most of the research presented here.

Moreover, Alessandro Treves and Edmund Rolls developed the decoding method discussed in chapter 3. Bill Skaggs derived the formula for the average information per spike. David Golomb created the useful database of simulated responses of the LGN cells. Barry Richmond and John Hertz tested their neural network procedure for information analysis by using David's database. Robert Robertson and Pierre Georges-Francois recorded, in the lab of Edmund Rolls, the activity of hippocampal view cells. Gabriele Biella, Lilette Riva and Pasquale Gurzi recorded response of cells in the rat SI cortex. Carlo Fulvi Mari is studying the numerical solutions of the saddle point equations derived and described in chapter 4.

I owe also an enormous debt to Alessandro D'Adda and Michele Caselle, outstanding men and outstanding scientists. Although they did not contribute directly to the work presented here, their role was anyway fundamental, because they inspired my love for science and math.

I also thank all the friends that I have met during my time in Torino University and in SISSA, and especially Marco Billó, Gabriele Gionti, Paolo De Lo Rios and Marco Scalerandi, with whom I have shared many important events of my life.

This thesis has been completed at the Department of Experimental Psychology of the University of Oxford. I thank for the many interesting discussions all the people that I have met in this departement, and in particular Roland Baddeley, Martin Elliffe, Tim Milward, Nestor Parga and Simon Schultz.

Finally, my deepest thanks go to my parents and to my brother, who always supported and encouraged me in all of my interests.

This thesis is dedicated to Loredana.

Bibliography

Abbott, L.F., Rolls, E.T., and Tovee, M.J. 1996. Representational capacity of face coding in monkeys. *Cerebral Cortex*, **6**, 498-505.

Abeles, M., Vaadia, E., and Bergman, H. 1990. Firing patterns of single units in the prefrontal cortex and neural network models. *Network* **1**, 13-25.

Amaral, D.G. 1993. Emerging principles of intrinsic hippocampal organization. *Current Opinion in Neurobiology* **3**, 225-229.

Amaral, D.G., Ishizuka, N., Claiborne, B. 1990. Neurons, numbers and the hippocampal network. *Prog in Brain Res* **83**, 1-11.

Amaral, D.G., Witter, M.P. 1989. The three-dimensional organization of the hippocampal formation: a review of anatomical data. *Neurosci.* **31**, 571-591.

Amit, D.J. 1989. Modelling Brain Function, Cambridge Univ Press, New York.

Atick, J.J. 1992. Could information theory provide an ecological theory of sensory processing? *Network* **3**, 213-251.

Atick, J.J. and Redlich, A.N. 1990. Towards a theory of early visual processing. *Neural Comp.* **2**, 308-320.

Barlow, H.B. 1961. Possible principles underlying the transformation of sensory messages. *Sensory Communication*, W.A. Rosenblith Ed., MIT Press, Cambridge, MA.

Barlow, H.B. 1989. Unsupervised learning. *Neural Comp.* **1**, 295-311.

Barnes, C.A., McNaughton, B.L., Mizumori, S.J., Lim, L.H. 1990. Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Prog in Brain Res* **83**, 287-300.

Bialek, W. 1991. Optimal signal processing in the nervous system. In: *Princeton lectures on biophysics*, W. Bialek Ed., World Scientific, London, UK.

Bialek, W., Callan, C.G., Strong, S.P 1996. Field theories for learning probability distributions. *Los Alamos archives cond-mat 9607180, submitted*

Bialek, W., DeWeese M., Rieke, F., and Warland, D. 1993. Bits and brain: information flow in the nervous system. *Physica* **A200**, 581-593.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R.R., and Warland, D. 1991. Reading a neural code. *Science* **252**, 1854-1857.

Biella, G., Riva, L., Sotgiu, M.L. 1996. Spino-thalamo-cortical correlation changes and oscillatory responses during noxious and non-noxious stimulations in rats. *Soc. Neurosc. Abs.* **21**, 114.

Brown, T.H., Kairiss, E.W., Keenan, C.L. 1990. Hebbian synapses: biophysical mechanisms and algorithms. *Ann Rev of Neurosci* **13**, 475-511.

Carlton, A.G. 1969. On the bias of information estimate. *Psych. Bull.* **71**, 108-109.

Chee-Orts, M.N., and Optican, L.M. 1993. Cluster method for analysis of transmitted information in multivariate neuronal data. *Biol. Cybernet.* **69**, 29-35.

Collingridge, G.L., Singer, W. 1990. Excitatory amino acid receptors and synaptic plasticity. *Trends Pharm. Sci.* **11**, 290-296.

Cover, T.M., and Thomas, J.A. 1991. *Elements of information theory.* John Wiley, New-York.

de Ruyter van Steveninck, R.R., and Laughlin, S.B. 1996. The rates of information transfer at graded-potential synapses. *Nature* **379**, 642-645.

Dong, D.W., and Atick, J.J. 1995. Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network* **6**, 159-178.

Eckhorn, R., Grüsser, O-J., Kröller,J., Pellnitz, K., and Pöpel, B. 1976. Efficiency of different neural codes: information transfer calculations for three different neuronal systems. *Biol. Cybernet.* **22**, 49-60.

Eckhorn, R., and Pöpel, B. 1975. Rigorous and extended application of information theory to the afferent visual system of the cat. II. Experimental results. *Kybernetik* **17**, 7-17.

Efron, B. 1982 The Jackknife, the bootstrap and other resampling plans. *Philadelphia, PA: Society for industrial and applied mathematics.*

Gawne T.J., and Richmond. B.J. 1993. How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.* **13**, 2758-2771.

Georgopoulos, A.P., Schwartz, A., Kettner, R.E. 1986. Neural population coding of movement direction. *Science* **233**, 1416-1419.

Georgopoulos, A.P., Taira, M., Lukashin, A. 1993. Cognitive neurophysiology of the motor cortex. *Science* **260**, 47-52.

Gochin, P.M., Colombo, M., Dorfmam, G.A., Gerstein, G.L. and Gross, C.G. 1994. Neural ensemble encoding in inferior temporal cortex. *J. Neurophysiol.* **71**, 2325-2337.

Golomb, D., Hertz, J., Panzeri, S., Richmond, B.J., Treves, A. 1996 How well can we estimate the information carried in neuronal responses from limited samples?. *Neural Comp.*, in press.

Golomb, D., Kleinfeld, D., Reid, R.C., Shapley, R.M., and Shraiman, B.I. 1994. On temporal codes and the spatiotemporal response of neurons in the lateral geniculate nucleus. *J. Neurophysiol.* **72**, 2990-3003.

Heller, J., Hertz, J.A., Kjaer, T.W., and Richmond, B.J. 1995. Information flow and temporal coding in primate pattern vision. *J. Comp. Neurosci.* **2**, 175-193.

Hertz, J.A., Kjaer, T.W., Eskander, E.N., and Richmond, B.J. 1992. Measuring natural neural processing with artificial neural networks. *Int. J. Neural Syst.* **3** (suppl), 91-103.

Kjaer, T.W., Hertz, J.A., and Richmond, B.J. 1994. Decoding cortical neuronal signals: networks models, information estimation and spatial tuning. *J. Comp. Neurosci.* **1**, 109-139.

Levine, M.W., and Troy, J.B. 1986. The variability of maintained discharge of cat dorsal-lateral geniculate cells. *J. Physiol.* **375**, 219-246.

Levy, W.B., Colbert, C.M., Desmond, N.L. 1990. Elemental adaptive processes of neurons and synapses: a statistical/computational perspective. In: *Neuroscience and connectionist theory (Gluck M, Rumelhart D, eds)*, Ch 5, pp 187-235. Hillsdale, N.J: Erlbaum

Levy, W.B., Desmond, N.L. 1985. The rules of elemental synaptic plasticity. In: *Synaptic modification, neuron selectivity, and nervous system organization (Levy WB, Anderson JA, Lehmkuhle S, eds)*, Ch 6, pp 105-121. Hillsdale, New Jersey: Erlbaum.

McClurkin, J.W., Optican, L.M., Richmond, B.J. and Gawne, T.J. 1991. Concurrent processing and complexity of temporally encoded neuronal messages in visual perception. *Science* **253**, 675-677.

Miller, G.A. 1955. On the bias of information estimates. *Information theory in psychology; problems and methods*, II-B, 95-100.

Mezard, M., and Parisi, G., and Virasoro, M.A. 1987. Spin glass theory and beyond, World Scientific,Singapore

Miles, R. 1988. Plasticity of recurrent excitatory synapses between CA3 hippocampal pyramidal cells. *Society for Neuroscience Abstracts* **14**, 19.

Morris, R.G.M. 1989 Does synaptic plasticity play a role in information storage in the vertebrate brain? In: *Parallel Distributed Processing: Implications for Psychology and Neurobiology* (Morris RGM, ed), Ch 11, pp 248-285. Oxford: Oxford University Press.

Nadal, J.-P., and Parga, N. 1993. Information processing by a perceptron in an unsupervised learning task. *Network* **4**, 295-312.

Olshausen, B.A., and Field, D.J. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607-609.

O'Keefe, J., and Recce, M.L. 1993. Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* **3**, 317-330.

Optican, L.M., Gawne, T.J., Richmond, B.J., and Joseph, P.J. 1991. Unbiased measures of transmitted information and channel capacity from multivariate neuronal data. *Biol. Cybernet.* **65**, 305-310.

Optican, L.M., and Richmond, B.J. 1987. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex: III. Information theoretic analysis. *J. Neurophysiol.* **57**, 162-178.

Panzeri, S., Biella, G., Rolls, E.T., Skaggs, W.E., Treves, A. 1996a. Speed, noise, information and the graded nature of neuronal responses. *Network* **7**, 365-370.

Panzeri, S., Biella, G., Sotgiu, M.L., Treves, A. 1995a. Information theoretical analysis of thalamocortical ensemble coding during noxious stimulation. *Soc. Neurosci. Abstr.* **21**, 114

Panzeri, S., and Treves, A. 1995b. Correcting measures of information for limited data samples. *Int. J. Neural Syst.* **7 (sup.)**, 133-137.

Panzeri, S., and Treves, A. 1996b. Analytical estimates of limited sampling biases in different information measures. *Network* **7**, 87-107.

Quenouville, M. 1949. Approximate tests of correlation in time series. *J. Roy. Statist. Soc.* **B 11**, 18-84.

Reid, R.C., and Shapley, R.M. 1992. Spatial structure of cone inputs to receptive fields in primate lateral geniculate nucleus. *Nature* **356**, 716-718.

Rieke, F., Warland, D., and Bialek, W. 1993. Coding efficiency and information rates in sensory neurons. *Europhys. Lett.* **22**, 151-156.

Rolls, E.T. 1995. A model of the operation of the hippocampus and entorhinal cortex in memory. *Int. J. Neural Syst.* **7 (sup.)**, 51-70.

Rolls, E.T, Critchley, H.D., and Treves, A. 1996a. Representation of olfactory information in the primate orbitofrontal cortex. *J. Neurophysiol.* **75**, 1982-1996.

Rolls, E.T., Robertson, R.G., and Georges-Francois, P. 1995a. The representation of space in primate hippocampus. *Soc. Neurosc. Abs.* **21**, 1494.

Rolls, E.T., Robertson, R.G., and Georges-Francois, P. 1996b. Spatial view cells in the primate hippocampus. *Submitted*

Rolls, E.T., and Tovée, M.J. 1995b. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex *J. Neurophysiol.* **73**, 713-726.

Rolls, E.T., Treves, A., Tovée, M.J, and Panzeri, S. 1996c. Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *submitted.*

Rolls, E.T., Treves, A., and Tovée, M.J. 1996d. The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Experimental Brain Research*, in press.

Rolls, E.T., Yaxley, S., and Sienkiewicz 1990. Gustatory respons~~es~~of single neurons in the orbitofrontal cortex of the macaque monkey *J. Neurophysiol.* **64**, 1055-1066.

Schultz, S.R., *et al.* 1996. Effects of realistic ocnnectivity and firing rate distributions on the information relayed by the Schaffer collaterals. In preparation.

Scobey, R.P., and Gabor, A.J. 1989. Orientation discrimination sensitivity of of single units in cat primary visual cortex. *Exp. Brain Res.* **77**.

Seress, L. 1988. Interspecies comparison of the hippocampal formation shows increased emphasis on the regio superior in the Ammon's horn of the human brain. *J. Hirnforsc.* **29**, 335-340.

Shannon, C.E., 1948. A mathematical theory of communication. *AT&T Bell Labs. Tech. J.* **27**, 379–423.

Skaggs, W.E., and McNaughton, B.L. 1992. Quantification of what it is that hippocampal cell firing encodes. *Soc. Neurosci. Abs.* **18** 1216;

Skaggs, W.E., McNaughton, B.L., Gothard, K., Markus, E. 1993. An information theoretic approach to deciphering the hippocampal code, in *Advances in Neural Information*

Processing Systems **5**, eds S. J. Hanson, J. D. Cowan, C. L. Giles, Morgan Kaufmann, San Mateo, pp 1030-1037.

Smith, D.V., and Travers, J.B. 1979. A metric for the breadth of tuning of gustatory neurons. *Chem Senses Flavour* **4**, 215-229.

Squire, L.R., Shimamura, A.P., Amaral, D.G. 1989. *Memory and the hippocampus.* In: Neural models of plasticity: Theoretical and empirical approaches (Byrne J, Berry WO, eds). Ch 12, pp 208- 239. New York: Academic Press.

Strong, S.P., Koberle, R., de Ruyter van Steveninck, R.R, and Bialek, W. 1996. Entropy and information in neural spike trains. *Los Alamos archives cond-mat 9603127, submitted.*

Theunissen, F.J., Roddey, J.C., Stufflebeam, S., Clague, H., and Miller J.P. 1996. *J. Neurophysiol.* **75**, 1345-1364.

Tovée, M.J., Rolls, E.T., Treves, A., Bellis, R.P. 1993. Information encoding and the response of single neurons in the primate temporal visual cortex. *J. Neurophysiol.* **70**, 640-654.

Treves, A., 1990. Graded-response neurons and information encodings in autoassociative memories. *Phys. Rev.* **A42**, 2418-2430.

Treves, A. 1991. Are spin-glass effects relevant to understanding realistic autoassociative networks?. *J. Phys.* **A 24**. 2645-2654.

Treves, A. 1995. Quantitative estimate of the information relayed by the schaffer collaterals. *J. Comp. Neurosci.* **2**, 259-272.

Treves, A. 1996. On the perceptual structure of face space, *Biosystems* **15**, in press.

Treves, A., and Panzeri, S. 1995. The upward bias in measures of information derived from limited data samples. *Neural Comp.* **7**, 399-407.

Treves, A., and Rolls, E.T. 1991. What determines the capacity of autoassociative memories in the brain? *Network* **2**, 371-397

Treves, A., and Rolls, E.T. 1992. Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus* **2**, 189-199.

Treves,A., and Rolls,E.T. 1994, A computational analysis of the role of the hippocampus in memory. *Hippocampus* **4**, 374-391.

Treves, A., Skaggs, W.E., Barnes, C.A. 1996. How much of the hippocampus can be explained by functional constraints? *Hippocampus* **6**, in press.

Tsodyks, M.V., and Feigel'man, M.V. 1988. The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett.* **6**, 101-105

West, M.J., Gundersen, H.G.J. 1990. Unbiased stereological estimation of the numbers of neurons in the human hippocampus. *J Comp Neurol* **296**, 1-22.

Wilson, M., and McNaughton, B.L. 1993. Dynamics of the hippocampal ensemble code for space. *Science* **261**, 1055-1058.

Witter, M.P. 1993. Organization of the entorhinal-hippocampal system: a review of current anatomical data. *Hippocampus* **3**, 33-44.

Wolpert, D.H., and Wolf, D.R. 1995. Estimating functions of probability distributions from from a finite set of samples. *Phys. Rev.* **E52**, 6841-6854.