Network: Computation in Neural Systems **7** (1996) 87–107. Printed in the UK

Analytical estimates of limited sampling biases in different information measures

Stefano Panzeri†‡§ and Alessandro Treves†

† Biophysics, SISSA, via Beirut 2–4, 34013 Trieste, Italy
‡ Mathematical Physics, SISSA, via Beirut 2–4, 34013 Trieste, Italy

Received 27 March 1995, in final form 27 October 1995

Abstract. Measuring the information carried by neuronal activity is made difficult, particularly when recording from mammalian cells, by the limited amount of data usually available, which results in a systematic error. While empirical *ad hoc* procedures have been used to correct for such error, we have recently proposed a direct procedure consisting of the analytical calculation of the average error, its estimation (up to subleading terms) from the data, and its subtraction from raw information measures to yield unbiased measures. We calculate here the leading correction terms for both the average transmitted information and the conditional information and, since usually one must first regularize the data, we specify the expressions appropriate to different regularizations. Computer simulations indicate a broad range of validity of the analytical results, suggest the effectiveness of regularizing by simple binning and illustrate the advantage of this over the previously used 'bootstrap' procedure.

1. Introduction

The performance of networks of neurons as information processing devices can only be correctly gauged by using the appropriate information measures as performance quantifiers. This is a straightforward notion as far as the analysis of abstract models is concerned, but when it comes to real neurons, whose activity is recorded *in vivo*, extracting information measures is so ridden with subtleties, especially in mammals, that in practice few measures have been produced until now (e.g. Eckhorn and Pöpel 1975, Optican and Richmond 1987, McClurkin *et al* 1991, Gawne and Richmond 1993, Tovee *et al* 1993). Apart from the psychological difficulty of accepting the validity of quantities which are always *relative* to the procedures used to measure them, the biggest source of problems has been the limited size of data samples, which results in measures distorted by a systematic error, occasionally as large as the target quantity itself. As a consequence, important questions such as the type of neuronal coding used by different systems in the mammalian brain, the speed of information processing and its efficiency at the neuronal level, have been most easily approached qualitatively from a theoretical viewpoint, rather than quantitatively from experimental observations‖.

Empirical procedures to correct information measures for limited sampling have been refined (Optican *et al* 1991, Chee-Orts and Optican 1993, Hertz *et al* 1992), but they

§ E-mail: stefano@limbo.sissa.it

‖ Although for some systems, as in the elegant analysis of early vision by Atick and collaborators (Atick and Redlich 1990, Dong and Atick 1995), it is possible to bypass the measurement of information quantities from the data.

0954-898X/96/010087+21$19.50

are not yet satisfactory, for reasons that will be made clear in the following. We have proposed, instead, an approach based on a direct evaluation and subtraction, of the limited sampling bias (Treves and Panzeri 1995). The idea, which had been conceived as early as 40 years ago (Miller 1955), needed to be developed to be applicable to neuronal data, in particular to be adapted to the regularization procedures used with neuronal data; this development is the content of the present report. Of crucial practical importance is the trade off between the information loss due to the regularization and the limited sampling error and our results, which give unbiased estimators of regularized quantities and shift the balance towards choosing milder regularizations.

2. Information measures from limited samples

To be concrete, we consider a 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 set $\mathcal{S}$ of S elements. We wish to measure both the (average) conditional information transmitted when s is presented,

$$I(s) = \int \mathrm{d}r\, P(r|s) \log_2 \frac{p(s|r)}{p(s)} = \int \mathrm{d}r\, P(r|s) \log_2 \frac{P(r|s)}{P(r)} \tag{1}$$

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

$$I = \sum_{s\in\mathcal{S}} p(s) \int \mathrm{d}r\, P(r|s) \log_2 \frac{P(r|s)}{P(r)}. \tag{2}$$

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 by 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 procedure suggested by Optican *et al* (1991) to correct for the error, and successively improved by Chee-Orts and Optican (1993), 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 in some cases 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 and not to measures of conditional information, since the random 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 information measures to different degrees, as will again be clear below, sometimes resulting in an *underestimation* of the bias.

More sophisticated is the procedure suggested by Hertz *et al* (1992), based on a 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 yield unbiased estimates, in the sense that after the regularization one obtains shuffled information very close to zero (Kjaer *et al* 1994). The last flaw in the shuffling procedure also applies here, although, in the sense that negligible shuffled information does not guarantee that the raw estimate is unbiased. Moreover, while *any* regularization results in information loss and in information values relative to *that* regularization, the regularization produced by the artificial neural network is particularly complex and data dependent and it is difficult to assess the relation between the original target information and the regularized measure one obtains. This is made evident in the apparently paradoxical results 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 conceptually appealing approach in cases, such as ours of stimulus–response pairs, when no reasonable assumption on the prior is self-evident.

We have therefore developed an alternative approach, based on the use of the replica trick to compute the *average error* directly as an asymptotic expansion in inverse powers of the sample size. In the first application of the approach (Treves and Panzeri 1995), we have required that the response space be discrete (or, when the original response space is continuous, that it be discretized with a straightforward binning procedure, that simply allots raw responses to the interval in which they happen to lie) and we have implicitly assumed that the conditional response probabilities (for each stimulus) are different from zero in every bin. With these assumptions we have found that the leading contribution to the bias, dependent solely upon the number of stimuli and response bins and already calculated many years ago by Miller (1955), yields most of the error and can thus be subtracted to correct raw estimates. Successive terms in the expansion are of little use: either they are negligible in comparison with the first term, or when N becomes very small, they explode quickly, signalling that data are so scarce that the expansion is meaningless beyond the first term.

Regularization methods different from pure discretizations, such as convolutions with a Gaussian kernel, are often used in practice to manipulate the raw data generated by recording the real responses of the cell(s), which are usually continuous (possibly multidimensional) variables. In this paper, we carry out a similar calculation of the average systematic error when different regularizations are used, to understand how such manipulations interact with the finite-size problem and to find the corresponding correction terms. Moreover, we consider not only mutual information but also conditional information (i.e. relative to a given stimulus) and find the appropriate correction terms, which again turn out to have different forms.

Our analysis also leads us to clarify the role of the previously used shuffled information in the correction procedure and to find simple criteria to establish whether the size of the data is large enough to obtain (given the regularization) a reliable and (after the correction) nearly unbiased measure of information.

3. 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 the next section.

Let us initially consider the more general case in which the (raw) neuronal response is a real (possibly multidimensional) variable†. It is clear from the formula for the mutual information (2), that if one is measuring a continuous output variable, in order to obtain an estimate of the transmitted information from a finite set of N data it is necessary to regularize the raw data in some way; otherwise, the finite number of responses will almost certainly all be 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 a measure of the entropy of the stimulus set only and not of the transmitted 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 data manipulations different from pure discretizations of the response space. These regularizations can be of a simple form, such as a convolution with a Gaussian followed by discretization, or much more complicated, like the neural network used by Hertz *et al* (1992).

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; convolution with a continuous distribution without discretization of the response space.

We shall present in the appendix A the explicit calculations leading to our expression for the bias in the second case only; but, for the sake of generality, we shall briefly discuss how to retrieve the results presented when the other data manipulations are applied.

3.1. Pure discretization of the response space

Let us consider in this section the case in which the real responses have been binned into R different intervals‡ $[m_{j-1}, m_j]$, $j = 1, \cdots, R$, by simply assigning each response to the interval in which it falls. In this case, the binning procedure satisfies an *independence condition*, in the sense that the number of times a given bin r is occupied depends only on the underlying occupancy probability of this bin, and not on the occupancy of other bins

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

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

(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:

$$I_N^{\mathrm{D}}(s) = \sum_{i=1}^{R} p_N(i|s) \log_2 \frac{p_N(i|s)}{p_N(i)} \qquad I_N^{\mathrm{D}} = \sum_{s \in \mathcal{S}} p_N(s) I_N^{\mathrm{D}}(s). \tag{3}$$

In (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 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 the response bin. Similarly, as N increases, the estimate of the transmitted information tends to the information carried by the discretized probabilities:

$$I^{\mathrm{D}}(s) = \sum_{i=1}^{R} p(i|s) \log_2 \frac{p(i|s)}{p(i)} \qquad I^{\mathrm{D}} = \sum_{s \in \mathcal{S}} p(s) I^{\mathrm{D}}(s). \tag{4}$$

By temporarily restricting ourselves to the total transmitted information, it is important to note that the value of the information I^{D} obtained *after* quantization is less than the value of information carried by the continuous responses and, in general, information measures are dependent upon the binning procedure adopted and, most importantly, upon the number of bins R. There is no way of estimating the difference between I and I^{D} from first principles, but a good strategy for controlling 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 eventually require R small.

Of course, the difference between I_N^{D} and I^{D} fluctuates depending on the particular outcomes of the N trials performed. One can, however, estimate the average of the difference, that is the bias, by averaging ($\langle \ldots \rangle$) 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:

$$\langle I_N^{\mathrm{D}} \rangle - I^{\mathrm{D}} = \sum_{m=1}^{\infty} C_m \tag{5}$$

where C_m represents successive contributions to the asymptotic expansion of the bias (the term C_m is proportional to N^{-m}). Here we report just the leading term, whose expression is derived in appendix A:

$$C_1^{\mathrm{D}} = \frac{1}{2N \log 2} \left\{ \sum_s \tilde{R}_s - \tilde{R} - (S-1) \right\} \tag{6}$$

where $\tilde{R}_s$ denotes the number of 'relevant' response bins for the trials with stimulus s, i.e. the response bins in which the occupancy probability $p(i|s)$ (at given s) is non-zero. In the same way, $\tilde{R}$ denotes the number of response bins where $p(i)$ is non-zero. In the case in which each response bin i has a non-zero probability of being occupied for every stimulus s, we recover the result reported in Treves and Panzeri (1995):

$$C_1^{\mathrm{D}} = \frac{1}{2N \log 2}\{(S-1)(R-1)\}. \tag{7}$$

At the end, to correct for the finite-size problem we have to evaluate the correction term in (6), which depends upon the underlying probabilities solely through the $\tilde{R}_s$ parameters, and thus in a much weaker way than the mutual information, which depends upon the full distributions. Therefore, even though the parameters $\tilde{R}_s, \tilde{R}$ have, in turn, to be 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 (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 also differs from the number of bins actually occupied, $\tilde{R}_s$, 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 $\tilde{R}_s$ for calculating C_1 leads, when few trials are available, to an underestimate of the systematic error (the underestimation becoming negligible for $R/N_s \ll 1$ 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 $\tilde{R}_s, \tilde{R}$. As mentioned above, Wolpert and Wolf (1995) show how to calculate any function of the probabilities *given* the experimental frequencies, using the Bayes rule. This requires some knowledge, or some assumption, on the *a priori* probability distributions of the probabilities (see appendix B). 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, which quantity depends on the full details of the probability tables. Nevertheless, we show (in appendix B) how *a correction* to the mutual information depending only upon a few parameters, such as $\tilde{R}_s, \tilde{R}$, can also be well estimated with a crude hypothesis about the prior probability functions. The idea is to use 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 estimate of the probabilities) matches the experimentally observed value.

This estimation, although based on a very simple ansatz (appendix B) on the prior distributions, is sufficient to give good results even for relatively small values of N, as shown in figure 1(*a*). The reason for this good estimation, in our opinion, lies in the fact that only the parameters $\tilde{R}_s, \tilde{R}$ have to be estimated based on the arbitrary ansatz, and the information I depends upon these only in the correction terms.

The observation that the leading bias term (6) can also, in general, be probability dependent leads to a better understanding of the effectiveness with which the shuffled information can correct for limited samples. If the underlying probability is such that each bin has non-zero probability, then the bias should be of the same order for the shuffled and the true probability table and we can correct the measured information by simply subtracting the value of the shuffled information, as previously stated by Treves and Panzeri (1995). If, instead, 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 our C_1^{D} term, (6), should continue to give a good estimate of the bias. Therefore, even when restricting to mutual information and discrete or discretized 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). The shuffling procedure can only be used when all three conditions are met: (i) responses are discrete or simply discretized; (ii) the target is the mutual information; *and* (iii) $p(s, i)$ is deemed to have all non-zero elements.

Our analysis, instead, can also be carried out for the conditional information. Again, we can give an asymptotic expansion for the bias:

$$\langle I_N^{\mathrm{D}}(s)\rangle - I^{\mathrm{D}}(s) = \sum_{m=1}^{\infty} C_m^{\mathrm{D}}(s) \tag{8}$$

and the leading correction term is now:

$$\begin{aligned} C_1^{\mathrm{D}}(s) = \frac{1}{2N\log 2}\widehat{\sum_i}\left\langle\frac{1}{p_N(s)}\right\rangle[1 - p(i|s)] \\ + \frac{1}{2N\log 2}\widehat{\sum_i}\left\{\frac{-p(i|s) + 2p^2(i|s)}{p(i)} - p(i|s)\right\} \end{aligned} \tag{9}$$

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 $\langle p^{-1}(s)\rangle$ the instances with $p_N(s) = 0$ must be excluded. Estimating this expression (9) 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 $C_1^{\mathrm{D}}(s)$ on the probabilities is not as simple as for C_1^{D}, and therefore a Bayesian estimate of $C_1^{\mathrm{D}}(s)$ is more complicated and, without some knowledge on the prior, is not expected to work as well. This handicap may be to some extent circumvented by choosing, when such freedom exists, response bins appropriate to the stimulus s being considered, i.e. collating all bins for which no response to s occurs into a single bin.

All the analytical results presented here for the discrete case are well confirmed by computer simulations presented in Treves and Panzeri (1995) and Panzeri and Treves (1995). New simulations with more realistic probability distributions are presented in the next subsection and confronted with the results obtained with different regularizations.

3.2. 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:

$$\tilde{I}_N^{\mathrm{D}}(s) = \sum_{i=1}^{R} \tilde{p}_N(i|s)\log_2\frac{\tilde{p}_N(i|s)}{\tilde{p}_N(i)} \qquad \tilde{I}_N^{\mathrm{D}} = \sum_{s\in\mathcal{S}} p_N(s)\tilde{I}_N^{\mathrm{D}}(s) \tag{10}$$

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:

$$\tilde{p}_N(i|s) \equiv \frac{1}{N_s}\sum_{j=1}^{N_s} E_i(r_j;\sigma) \qquad \tilde{p}_N(i) \equiv \sum_{s\in\mathcal{S}} p_N(s)\tilde{p}_N(i|s) \tag{11}$$

where $E_i(r_j;\sigma)$ is the integral (over the ith interval) of the kernel function centred in r_j:

$$E_i(r_j;\sigma) = \int_{m_{i-1}}^{m_i} \mathrm{d}r\, K(r, r_j, \sigma). \tag{12}$$

The sum over j in (11) 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† itself: $\sigma = \sigma(s, r_j)$. For large N the raw response distributions approach the underlying ones and thus we can write:

$$\tilde{p}(i|s) = \int \mathrm{d}r\, E_i(r;\sigma) P(r|s) \qquad \tilde{p}(i) = \sum_s p(s)\tilde{p}(i|s). \tag{13}$$

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

$$\tilde{I}(s) = \sum_{i=1}^{R} \tilde{p}(i|s) \log_2 \frac{\tilde{p}(i|s)}{\tilde{p}(i)} \qquad \tilde{I} = \sum_{s\in\mathcal{S}} p(s)\tilde{I}(s). \tag{14}$$

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 top of the loss due to the 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

$$\langle \tilde{I}_N^{\mathrm{D}} \rangle - \tilde{I}^{\mathrm{D}} \simeq \tilde{C}_1^{\mathrm{D}} \qquad \langle \tilde{I}_N^{\mathrm{D}}(s) \rangle - \tilde{I}^{\mathrm{D}}(s) \simeq \tilde{C}_1^{\mathrm{D}}(s) \tag{15}$$

we now find the expressions

$$\tilde{C}_1^{\mathrm{D}} = \frac{1}{2N\log 2}\left\{ \sum_{s\in\mathcal{S}} \widehat{\sum\nolimits_i} \frac{\tilde{q}(i|s)}{\tilde{p}(i|s)} - \widehat{\sum\nolimits_i} \frac{\tilde{q}(i)}{\tilde{p}(i)} - (S-1) \right\} \tag{16}$$

$$\begin{aligned} \tilde{C}_1^{\mathrm{D}}(s) = {} & \frac{1}{N\log 2} \widehat{\sum\nolimits_i} \left\{ \langle p_N^{-1}(s) \rangle \frac{\tilde{q}(i|s) - \tilde{p}^2(i|s)}{2\tilde{p}(i|s)} + \frac{\tilde{p}^2(i|s) - \tilde{q}(i|s)}{\tilde{p}(i)} \right\} \\ & + \frac{1}{2N\log 2} \widehat{\sum\nolimits_i} \left\{ \frac{\tilde{q}(i)\tilde{p}(i|s) - \tilde{p}(i|s)\tilde{p}^2(i)}{\tilde{p}^2(i)} \right\} \end{aligned} \tag{17}$$

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

$$\tilde{q}(i|s) \equiv \int \mathrm{d}r\, P(r|s) E_i^2(r;\sigma) \qquad \tilde{q}(i) \equiv \sum_s p(s)\tilde{q}(i|s). \tag{18}$$

The correction terms (16) and (17) 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 (16) and (17) from the data and, in particular, how to avoid undercounting the bins with non-zero probability over which to take the sums in (16) and (17). 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

† 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 $P_N(\cdot)$.

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‡:

$$\begin{aligned}\tilde{C}_1^{\mathrm{D}} - \langle(\tilde{C}_1^{\mathrm{D}})_N\rangle &\equiv \Delta(\tilde{C}_1^{\mathrm{D}}) \\ &= \frac{1}{2N\log 2}\left\{\sum_s \widehat{\sum}_i \left[1-\tilde{p}(i-1|s)-\tilde{p}(i|s)-\tilde{p}(i+1|s)\right]^{N_s}\right\} \\ &\quad -\frac{1}{2N\log 2}\left\{\widehat{\sum}_i \left[1-\tilde{p}(i-1)-\tilde{p}(i)-\tilde{p}(i+1)\right]^{N}\right\}.\end{aligned} \tag{19}$$

This approximate form for the underestimation of $\tilde{C}_1^{\mathrm{D}}$ 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, $\Delta(\tilde{C}_1^{\mathrm{D}})$ can be added to $\tilde{C}_1^{\mathrm{D}}$ to marginally improve the estimation of the bias.

As for the validity of the bootstrap procedure, the fact that the correction terms (16) and (17) 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 and the shuffled responses), the shuffled information might easily *underestimate* the bias, because it might reflect an effectively stronger regularization. Thus, in these 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.

To support these analytical results, we performed explicit numerical simulations. We chose as 'test' underlying probabilities Poisson distributions, which are fair simple models of the spiking activity of neurons under certain conditions (Abeles *et al* 1990, Levine and Troy 1986, Scobey and Gabor 1989). We generated 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

$$\bar{r} = -\log\left(1-\frac{x}{2a}\right) \quad \text{if } x < 2a \qquad r = 0 \quad \text{if } x > 2a. \tag{20}$$

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 ($t = 500$ ms in the present simulations) followed the Poisson distribution

$$P(n|s) = \frac{[\bar{r}(s)t]^n \exp[-\bar{r}(s)t]}{n!}. \tag{21}$$

To measure, from N trials, the information carried by the firing rates generated in this way, we used a regularization procedure similar to that used in Rolls *et al* (1995a). The range of

‡ 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.

responses was 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). A smoothing procedure was 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).

Figure 1 shows, 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 (figure 1(a)), the asymptotic value of the discretized information is only a few per cent 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 the Bayesian estimates for $\tilde{R}$, $\tilde{R}_s$ 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

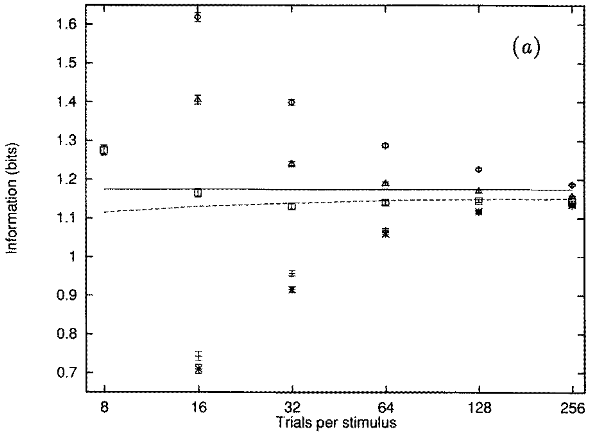

Figure 1. Mutual information values for the distribution of stimuli and Poisson responses described in the text (the sparseness of the mean firing rates is $a = 0.4$), with $S = 16$ and $R = 16$ and different values of N_s. The three panels correspond to (a) $\gamma = 0.0$ (pure discretization); (b) $\gamma = 0.5$; (c) $\gamma = 1.0$. 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 (this latter value varies with N because the regularization smoothing width is data dependent). In the first panel, compared to these reference values are, for each N, the raw estimates (◇), the estimates corrected by subtracting the C_1 term calculated by estimating the relevant bins by counting the number of actually occupied ones (△), estimating the effective bins with the Bayesian procedure described in the text (□), taking all bins to be relevant ($\tilde{R}_s = R = 16$) (+) and the estimate corrected by the bootstrap method (∗). In the second and third panels, we plot only the raw estimates (◇), those corrected by subtracting $\tilde{C}_1^{\mathrm{D}} + \Delta(\tilde{C}_1^{\mathrm{D}})$ (□) and by the bootstrap method (∗). 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.

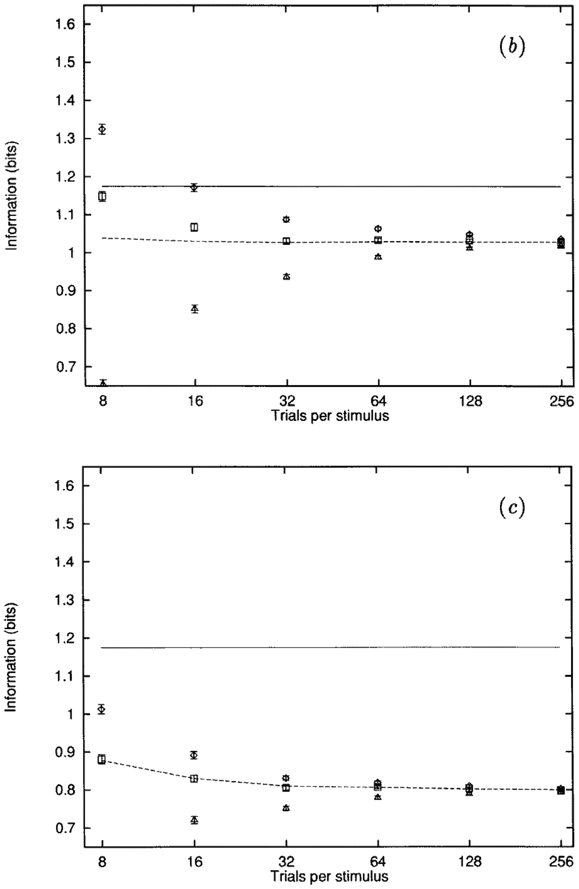

Figure 1. Continuation.

actually occupied response bins, or on the total number of bins, is not much more effective than the bootstrap or even raw estimates (note that, unlike in Treves and Panzeri (1995), we use here more realistic response distributions with many empty bins, which accounts for the poor performance of subtracting either the shuffled information or C_1 calculated naively). When a weaker (figure 1(*b*)) or stronger (figure 1(*c*)) 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 asymptotic (but strongly downward biased) values is faster (particularly in figure 1(*c*)). The conclusion appears to be that smoothing with a Gaussian does more damage than good, although we do note that (i) there may be other reasons for smoothing (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 response, smoothing may induce much smaller loss than in our examples, with possibly faster convergence with sample size.

Figure 2 shows the worth of subtracting the $\tilde{C}_1^{\mathrm{D}}(s)$ 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

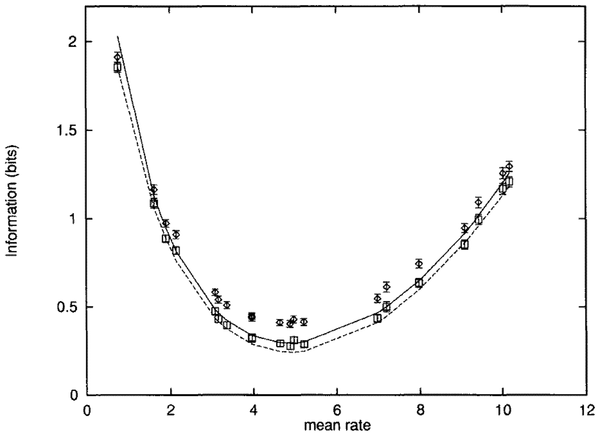

Figure 2. Values for the information conditional to which of $S = 20$ (simulated) stimuli was presented, plotted against the mean rate $\bar{r}(s)$ to each stimulus (on an arbitrary scale). The firing rates are distributed with sparseness $a = 0.7$. Again, the full curve indicates the real and the dashed one the regularized information values; and the symbols indicate raw and subtracted measures, each with standard deviation of the mean over 100 measures. Here $R = 10$, $N = 300$, $\gamma = 0.33$.

values within the narrow range spanned by the difference between real and regularized information values. The C-shape of the information versus rate plot and whether or not (as in this simulation) it touches the rate axis are interesting facts, discussed by Rolls *et al* (1995b).

3.3. Neural network regularization

When a regularization similar to that introduced by Hertz *et al* (1992) is used, the bias can be evaluated in a similar fashion. The idea of Hertz and his co-workers is to use a feed-forward neural network to fit, from the real responses r_j to a given stimulus s, the conditional probability that a stimulus is recognized as the ith:

$$\tilde{p}_N(i|s) \equiv \frac{1}{N_s}\sum_{j=1}^{N_s} E_i(r_j;\omega) \qquad \tilde{p}_N(i) \equiv \sum_{s\in\mathcal{S}} p_N(s)\tilde{p}_N(i|s) \tag{22}$$

where

$$E_i(r;\omega) = \frac{\exp\left[\sum_l (W_{il}H_l + B_i)\right]}{\sum_{j=1}^{S}\exp\left[\sum_l (W_{jl}H_l + B_j)\right]} \qquad H_l(r) = \tanh\left[\sum_{m=1}^{Q}\omega_{lm}r_m + b_l\right]. \tag{23}$$

In (23), 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 Kjaer *et al* (1994) for details). 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 $\tilde{p}(i|s)$ (22) can be interpreted as the conditional probability with which a response elicited by stimulus s may be attributed to stimulus i.

It is to be noted that, in this procedure (Hertz *et al* 1992, Kjaer *et al* 1994), to avoid overfitting, the (fitting) parameters W, ω, b, B entering in the neural network are adjusted on a set of 'training' data and the information is calculated on a set of 'test' experimental data. Thus, in the context of evaluating the finite size bias, the parameter N is the number of 'test' stimulus–response pairs. Without going further into the details of the procedure†, it is sufficient, for our purposes, to remark that the form of the regularized probability distributions (22) is the same as in (11), except that $E_i(r)$ is no longer evaluated simply by integrating a continuous kernel over the ith bin, but with the more complicated rule (23). This does not affect the results for the bias, which are therefore the same as in subsection 3.2, with the only difference that $E_i(r)$ must be computed from (23) instead of (12).

In practice, the information estimate produced by the network is unlikely to require any finite size subtraction as, if anything, it suffers more from the loss due to regularization. A comparison of the binning-and-correcting procedure and the neural network procedure on a large set of realistic simulated data is the object of another study (Golomb *et al* 1996).

3.4. 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

$$\tilde{I}_N(s) = \int \mathrm{d}r\, \tilde{P}_N(r|s) \log_2 \frac{\tilde{P}_N(r|s)}{\tilde{P}_N(r)} \qquad \tilde{I}_N = \sum_{s\in\mathcal{S}} p_N(s)\tilde{I}_N(s) \tag{24}$$

where the $\tilde{P}_N$ are the experimental distributions, obtained by convolving the experimental responses r_i with a kernel distribution $K(r, r_i, \sigma)$:

$$\tilde{P}_N(r|s) = \frac{1}{N_s}\sum_{j=1}^{N_s} K(r, r_i, \sigma) \qquad \tilde{P}_N(r) = \sum_{s\in\mathcal{S}} p_N(s)\tilde{P}_N(r|s) \tag{25}$$

where the sum over j is performed over all the N_s experimental responses to the stimulus s. As N increases, the raw response distributions approach the underlying ones:

$$\tilde{P}(r|s) = \int \mathrm{d}r_1\, K(r, r_1, \sigma) P(r_1|s) \qquad \tilde{P}(r) = \sum_{s} p(s)\tilde{P}(r|s) \tag{26}$$

and the raw estimates of information tend to:

$$\tilde{I}(s) = \int \mathrm{d}r\, \tilde{P}(r|s) \log_2 \frac{\tilde{P}(r|s)}{\tilde{P}(r)} \qquad \tilde{I} = \sum_{s\in\mathcal{S}} p(s)\tilde{I}(s). \tag{27}$$

The expressions we find in this case for the bias are

$$\tilde{C}_1 = \frac{1}{2N\log 2}\left\{\int \mathrm{d}r \left[\sum_{s\in\mathcal{S}}\left(\frac{\tilde{Q}(r|s)}{\tilde{P}(r|s)}\right) - \frac{\tilde{Q}(r)}{\tilde{P}(r)}\right] - (S-1)\right\} \tag{28}$$

$$\begin{aligned}\tilde{C}_1(s) = {}& \frac{1}{N\log 2}\int \mathrm{d}r \left[\langle p_N^{-1}(s)\rangle \frac{\tilde{Q}(r|s) - \tilde{P}^2(r|s)}{2\tilde{P}(r|s)} + \frac{-\tilde{Q}(r|s) + \tilde{P}^2(r|s)}{\tilde{P}(r)}\right] \\ & + \frac{1}{2N\log 2}\int \mathrm{d}r\, \frac{-\tilde{P}^2(r)\tilde{P}(r|s) + \tilde{Q}(r)\tilde{P}(r|s)}{\tilde{P}^2(r)}\end{aligned} \tag{29}$$

† It should be noted that the mutual information defined in Kjaer *et al* (1994) is not fully equivalent to the mutual information carried by the regularized probabilities (14).

where

$$\tilde{Q}(r|s) = \frac{1}{N_s}\sum_{j=1}^{N_s} K^2(r, r_i, \sigma) \qquad \tilde{Q}(r) = \sum_{s\in\mathcal{S}} p_N(s)\tilde{Q}_N(r|s). \tag{30}$$

In the continuous case, the problem of underestimation of the correction terms (28) and (29) when calculated from data, is absent, 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 continuum limit, and the simple expressions above hold, whenever the discretization is sufficiently fine with respect to the width of the kernel.

4. The bias with fixed number of trials per stimulus

In the previous section we studied the finite size distortions when the stimuli are drawn at random from a discrete set. Here we present the results 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 (A4)–(A6), 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

$$\tilde{C}_1^{\mathrm{D}} = \frac{1}{2N\log 2}\left\{\widehat{\sum}_i\left[\sum_{s\in\mathcal{S}}\left(\frac{\tilde{q}(i|s)}{\tilde{p}(i|s)} + \frac{p_N(s)\tilde{p}^2(i|s)}{\tilde{p}(i)}\right) - \frac{\tilde{q}(i)}{\tilde{p}(i)}\right] - S\right\} \tag{31}$$

$$\begin{aligned}\tilde{C}_1^{\mathrm{D}}(s) = {} & \frac{1}{N\log 2}\widehat{\sum}_i\left\{\langle p_N^{-1}(s)\rangle\frac{\tilde{q}(i|s) - \tilde{p}^2(i|s)}{2\tilde{p}(i|s)} + \frac{\tilde{p}^2(i|s) - \tilde{q}(i|s)}{\tilde{p}(i)}\right\} \\ & + \frac{1}{2N\log 2}\widehat{\sum}_i\left\{\frac{\tilde{p}(i|s)\tilde{q}(i)}{\tilde{p}^2(i)} - \sum_{s'}\frac{p(s')\tilde{p}^2(i|s')\tilde{p}(i|s)}{\tilde{p}^2(i)}\right\}\end{aligned} \tag{32}$$

where the notation is the same as in section 3.2. The results corresponding to the other regularizations considered in the previous section can be derived by taking the appropriate limits, as explained in the appendix.

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 validity of our correction, which in the discrete case works well even down to $N_s \sim R$. Given the effectiveness of the binning procedure, we recommend limiting the regularization to simple binning, unless motivated by other considerations (as mentioned in section 3.2). The important question of the 'optimal choice' of the number of bins for an experiment with S stimuli and N_s trials per stimulus remains. A reasonable answer 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. This choice should effectively minimize the combined error due to regularization and finite sampling. In figure 3 information estimates obtained by choosing $R = N_s$ are compared, for different values of N_s, with the full, unregularized, value of the information carried by the Poisson distribution of responses. Results appear to be reasonable in the whole N_s range explored. However, the correction procedure based on binning indicates the minimum appropriate number of trials per stimulus in real experiments. The correction functions reasonably down to $N_s \simeq R$, and the minimum number of response bins which may not throw away information, if the appropriate code is used, is just the same as the number of stimuli, $R = S$. Therefore a minimum number of $N_s = S$ trials per stimulus is a fair demand to be made on the design of experiments from which information estimates are to be derived.

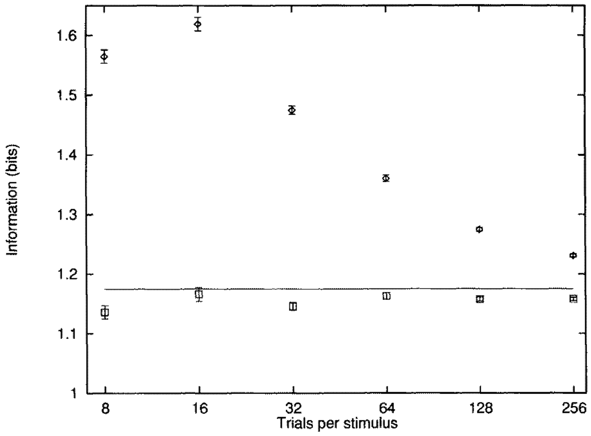

Figure 3. Mutual information values for the distribution of stimuli and Poisson responses described in the text. Here $a = 0.4$, $S = 16$ and the response space is purely discretized. The number of bins R is fixed equal to N_s. The symbols have the same meaning as in figure 1(*a*). Note that for $N_s = 16$, also $R = 16$ and the result is the same as shown in figure 1(*a*). For higher values of N_s, results approach the unregularized value of the information, whereas in figure 1(*a*) they approached the value regularized with $R = 16$ bins.

6. Conclusions

The work reported here has no deep theoretical significance; conceptually, it is on a par with calculating the correct expression for the moments of a Gaussian distribution, for example, when these have to be estimated from the data. It is, however, of practical importance, especially, although by no means solely, for the analysis of neuronal activity recorded in the mammalian nervous system *in vivo*. Measuring the information carried by neuronal activity has been avoided by many neurophysiologists because of the seemingly huge amount of data required to obtain reasonable statistics, and the outcomes of such measurements have been widely accepted only in a few instances, e.g. when performed in insects (Bialek *et al* 1991), in which data collection is not a constraint and the results appeared *hard* (just as the nervous systems examined appear to be hard-wired).

Our procedure for evaluating the bias and correcting information estimates will not, as is evident from the figures, be of any help when data are so scarce as to make the expansion meaningless; nor, obviously, when they are abundant enough to make any correction superfluous. The procedure is only useful for a range of sample sizes, which range is, however, roughly of the order of magnitude of that in which typical neurophysiological experiments lie. The data collected in such experiments are, then, available for information measurements, at the very limited cost of adding a very quick routine to standard data analysis packages (Rolls *et al* 1995a–c)†. The diffusion of the practice of measuring (accurately) the information content of neuronal activity is likely to greatly enhance our quantitative understanding of the processing of information in the nervous system.

Appendix A.

We give here the derivation of the results presented in previous sections. 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. The method used here is different from that employed in Treves and Panzeri (1995) and closer to that of Carlton (1969); results are, in any case, fully equivalent.

We start by calculating the average of the total amount of information (10), which can be expressed as follows:

$$\langle \tilde{I}_N^{\mathrm{D}} \rangle = \sum_{s\in\mathcal{S}} \widehat{\sum}_i \langle p_N(s)\tilde{p}_N(i|s)\log_2 \tilde{p}_N(i|s)\rangle - \widehat{\sum}_i \langle \tilde{p}_N(i)\log_2 \tilde{p}_N(i)\rangle \qquad \text{(A1)}$$

where $\tilde{p}(\cdot)$ is defined in (11) and the hat on the sum over response bins i in (A1) 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 disappear from the average). Now we can used the following series expansion for the logarithm:

$$-\log_2(\tilde{p}_N(\cdot)) = \frac{1}{\log 2}\sum_{j=1}^{\infty}\frac{(1-\tilde{p}_N(\cdot))^j}{j}. \qquad \text{(A2)}$$

This expansion (A2) is convergent for all values of $\tilde{p}_N(\cdot)$, since $0 < \tilde{p}_N(\cdot) \leqslant 1$ (note that in our calculation the configuration $\tilde{p}_N(\cdot) = 0$ can be excluded since it gives a vanishing contribution to the average). Taking term by term expectations in (A1) we find:

$$\begin{aligned}\langle \tilde{I}_N^{\mathrm{D}} \rangle = \frac{-1}{\log 2}\sum_s \widehat{\sum}_i \sum_{j=1}^{\infty}\left\langle p_N(s)\tilde{p}_N(i|s)\frac{(1-\tilde{p}_N(i|s))^j}{j}\right\rangle \\ +\frac{1}{\log 2}\widehat{\sum}_i \sum_{j=1}^{\infty}\left\langle \tilde{p}_N(i)\frac{(1-\tilde{p}_N(i))^j}{j}\right\rangle \\ = \frac{-1}{\log 2}\sum_s \widehat{\sum}_i \sum_{j=1}^{\infty}\sum_{k=0}^{j}\frac{(-1)^k}{j}\binom{j}{k}\langle p_N(s)\tilde{p}_N^{k+1}(i|s)\rangle \\ +\frac{1}{\log 2}\widehat{\sum}_i \sum_{j=1}^{\infty}\sum_{k=0}^{j}\frac{(-1)^k}{j}\binom{j}{k}\langle \tilde{p}_N^{k+1}(i)\rangle \qquad \text{(A3)}\end{aligned}$$

where in the last step we used the binomial decomposition for $(1-\tilde{p}_N(\cdot))^j$. 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 $N_s \equiv Np_N(s)$) simply by assuming that the probability of obtaining a raw response r (given the stimulus s) is given

† We are happy to make the required routine available via the Internet.

by $P(r|s)\,\mathrm{d}r$ 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 $\langle(\tilde{p}_N(i))^k\rangle$, since the parameters specifying the kernel can be stimuli dependent, we must decompose $\tilde{p}_N(i)$ as $\tilde{p}_N(i) = \sum_s p_N(s)\tilde{p}_N(i|s)$, average first over the responses (at fixed stimulus) and finally over $p_N(s)$ with the multinomial distribution. In general, we obtain the following expressions:

$$\langle\tilde{p}_N^k(i|s)\rangle = \tilde{p}^k(i|s) + \frac{1}{N_s}\binom{k}{2}\tilde{p}^{k-2}(i|s)\left[\tilde{q}(i|s) - \tilde{p}^2(i|s)\right] + \mathrm{o}\left(\frac{1}{N_s\tilde{p}(i|s)}\right) \tag{A4}$$

$$\langle p_N(s)\tilde{p}_N^k(i|s)\rangle = p(s)\tilde{p}^k(i|s) + \frac{1}{N}\binom{k}{2}\tilde{p}^{k-2}(i|s)\left[\tilde{q}(i|s) - \tilde{p}^2(i|s)\right] + \mathrm{o}\left(\frac{1}{Np(s)\tilde{p}(i|s)}\right) \tag{A5}$$

$$\langle\tilde{p}_N^k(i)\rangle = \tilde{p}^k(i) + \frac{1}{N}\binom{k}{2}\tilde{p}^{k-2}(i)\left[\tilde{q}(i) - \tilde{p}^2(i)\right] + \mathrm{o}\left(\frac{1}{N\tilde{p}(i)}\right) \tag{A6}$$

where $\tilde{q}(\cdot)$ is defined in (18). Ignoring the third term in each of (A4)–(A6) and then substituting (A4)–(A6) into (A3), we find an expression for the bias which is exact up to $\mathrm{O}(1/N^2)$ terms and is a good approximation to the bias if in each bin $N_s\tilde{p}(i|s) \gg 1$:

$$\begin{aligned}
\langle\tilde{I}_N^{\mathrm{D}}\rangle \simeq{}& \frac{-1}{\log 2}\sum_s \widehat{\sum}_i \sum_{j=1}^{\infty}\frac{1}{j}\left[1 - \tilde{p}(i|s)\right]^{j-2}\Bigg\{\tilde{p}(i|s)\left[1 - \tilde{p}(i|s)\right]^2 \\
&+\frac{1}{2N}\left[\tilde{q}(i|s) - \tilde{p}^2(i|s)\right]\left[j(j-1)\tilde{p}(i|s) - 2j(1 - \tilde{p}(i|s))\right]\Bigg\} \\
&+\frac{1}{\log 2}\widehat{\sum}_i \sum_{j=1}^{\infty}\frac{1}{j}\left[1 - \tilde{p}(i)\right]^{j-2}\Bigg\{\tilde{p}(i)\left[1 - \tilde{p}(i)\right]^2 \\
&+\frac{1}{2N}\left[\tilde{q}(i) - \tilde{p}^2(i)\right]\left[j(j-1)\tilde{p}(i) - 2j(1 - \tilde{p}(i))\right]\Bigg\} \\
={}& \tilde{I}^{\mathrm{D}} + \tilde{C}_1^{\mathrm{D}}
\end{aligned} \tag{A7}$$

where $\tilde{I}^{\mathrm{D}}$ is given in (14) and $\tilde{C}_1^{\mathrm{D}}$ is the leading contribution to the bias:

$$\tilde{C}_1^{\mathrm{D}} = \frac{1}{2N\log 2}\left\{\widehat{\sum}_i\left[\sum_s\left(\frac{\tilde{q}(i|s)}{\tilde{p}(i|s)}\right) - \frac{\tilde{q}(i)}{\tilde{p}(i)}\right] - (S-1)\right\}. \tag{A8}$$

By going further in the $1/N$ expansion when considering the averages (A4)–(A6), 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:

$$\begin{aligned}
\tilde{C}_2^{\mathrm{D}} ={}& \frac{1}{12N^2\log 2}\left\{\sum_{s\in\mathcal{S}}\langle p_N^{-1}(s)\rangle\left[\widehat{\sum}_i\frac{-2\tilde{p}(i|s)\tilde{t}(i|s) + 3\tilde{q}^2(i|s)}{\tilde{p}^3(i|s)} - 1\right]\right\} \\
&-\frac{1}{12N^2\log 2}\left\{\widehat{\sum}_i\frac{-2\tilde{p}(i)\tilde{t}(i) + 3\tilde{q}^2(i)}{\tilde{p}^3(i)} + 1\right\}
\end{aligned} \tag{A9}$$

where

$$\tilde{t}(i|s) \equiv \int \mathrm{d}r\; P(r|s)E_i^3(r|s) \qquad \tilde{t}(i) \equiv \sum_s p(s)\tilde{t}(i|s). \tag{A10}$$

Higher-order terms (valid in the discrete case) are reported in Treves and Panzeri (1995). In fact, (A8) is derived (as the leading term in the bias) under the condition that $N_s \tilde{p}(i|s) \gg 1$ 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 $\tilde{C}_1^{\rm D}$) under the less stringent condition that $\tilde{C}_1^{\rm D} \ll 1$. Therefore, the higher-order contributions are, in any case, close to negligible whenever $\tilde{C}_1^{\rm D}$ is a good approximation for the bias. When this is not the case, because the condition $N_s \tilde{p}(i|s) \gg 1$ is severely violated, computer simulations indicate that taking higher-order terms into account (which is itself not easy), does not help; on the contrary, in such a low-N regime in which $\tilde{C}_1^{\rm D}$ is often already too large, the next terms become huge and signal the breakdown of the expansion procedure.

If one is interested in measuring, instead of the average 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 $\langle \tilde{I}(s)^{\rm D} \rangle$ from (10),

$$\langle \tilde{I}_N^{\rm D}(s) \rangle = \frac{1}{\log 2} \sum_{i=1}^{R} \langle \tilde{p}_N(i|s) \log \tilde{p}_N(i|s) \rangle - \frac{1}{\log 2} \sum_{i=1}^{R} \langle \tilde{p}_N(i|s) \log \tilde{p}_N(i) \rangle \tag{A11}$$

after using the convergent expansion (A2) for the logarithm, one has to calculate the average of $\langle \tilde{p}_N(i|s) \tilde{p}_N^k(i) \rangle$ up to the next-to-leading order

$$\begin{aligned} \langle \tilde{p}_N(i|s) \tilde{p}_N^k(i) \rangle = \tilde{p}(i|s) \tilde{p}^k(i) + \frac{1}{N} \binom{k}{2} \tilde{p}(i|s) \tilde{p}^{k-1}(i) \left[1 - \tilde{p}(i)\right] \\ + \frac{k}{N} \left[1 - \tilde{p}(i|s)\right] \tilde{p}(i|s) \tilde{p}^{k-1}(i) + \mathrm{o}\left(\frac{1}{N p(s) \tilde{p}(i|s)} \right). \end{aligned} \tag{A12}$$

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

$$\langle \tilde{I}_N^{\rm D}(s) \rangle - \tilde{I}^{\rm D}(s) \simeq \tilde{C}_1^{\rm D}(s) \tag{A13}$$

with

$$\begin{aligned} C_1^{\rm D}(s) = \frac{1}{N \log 2} \widehat{\sum}_i \left\{ \langle p_N^{-1}(s) \rangle \frac{\tilde{q}(i|s) - \tilde{p}^2(i|s)}{2 \tilde{p}(i|s)} + \frac{\tilde{p}^2(i|s) - \tilde{q}(i|s)}{\tilde{p}(i)} \right\} \\ + \frac{1}{2N \log 2} \widehat{\sum}_i \left\{ \frac{\tilde{q}(i) \tilde{p}(i|s) - \tilde{p}(i|s) \tilde{p}^2(i)}{\tilde{p}^2(i)} \right\}. \end{aligned} \tag{A14}$$

The *discrete case* (for which the results are fully discussed in section 3.1) 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 (A8) that the leading bias term takes the form:

$$\begin{aligned} C_1^{\rm D} &= \frac{1}{2N \log 2} \left\{ \sum_{s \in \mathcal{S}} \widehat{\sum}_i \left[1 - p(i|s)\right] - \widehat{\sum}_i \left[1 - p(i)\right] \right\} \\ &= \frac{1}{2N \log 2} \left\{ \sum_s \tilde{R}_s - \tilde{R} - (S - 1) \right\}. \end{aligned} \tag{A15}$$

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

$$\langle I_N^{\rm D} \rangle - I^{\rm D} \simeq -\widehat{\sum}_i \left\{ \log_2 \left(1 + \frac{1 - p(i)}{N p(i)} \right) - \frac{1}{2N \log 2} \frac{p(i) \left[1 - p(i)(N - 1)\right]}{(N p(i) + 1 - p(i))^2} \right\}$$

$$+\sum_{s\in\mathcal{S}}\widehat{\sum}_i\left\{\log_2\left(1+\frac{1-p(i|s)}{N_s p(i|s)}\right)\right.$$

$$\left.-\frac{1}{2N_s\log 2}\,\frac{p(i|s)\left[1-p(i|s)(N_s-1)\right]}{(N_s p(i|s)+1-p(i|s))^2}\right\}. \tag{A16}$$

The expression (A16) for the bias agrees with our expression (6), up to the $1/N$ order, but is very different when going to higher orders. The procedure employed by Carlton to derive the result (A16) is similar to that presented here, in the sense that he uses the expansion (A2) 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 (A4)–(A6), 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 (A4)–(A6)) a partial re-summation (to all orders in $1/N$) of the complete expression for the bias. 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 (A16) in the simulation reported in figure 1(*a*), we 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 3.4, 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 $R_{i;n}$ ($i = 0, \pm 1, \pm 2, \ldots$ labels each interval)

$$R_{i;n}\equiv\left\{r;\ \frac{i}{2^n}\varrho\leqslant r<\frac{i+1}{2^n}\varrho\right\}. \tag{A17}$$

The discrete probabilities (13) have the form

$$\tilde{p}_n(i|s)\equiv\int_{R_{i;n}}\mathrm{d}r\ \tilde{P}(r). \tag{A18}$$

By introducing the function

$$\Gamma_n(r)\equiv\frac{2^n}{\varrho}\tilde{p}_n(i)\ \text{for}\ r\in R_{i;n} \tag{A19}$$

we have the identity

$$\tilde{p}_n(i)\log_2\left(\frac{2^n}{\varrho}\tilde{p}_n(i)\right)=\int_{R_{i;n}}\Gamma_n(r)\log_2\Gamma_n(r)\,\mathrm{d}r \tag{A20}$$

from which we can derive

$$\sum_{i,s}\tilde{p}_n(s,i)\log_2\frac{\tilde{p}_n(s,i)}{p(s)\tilde{p}_n(i)}=\sum_s\int\mathrm{d}r\ \Gamma_n(s,r)\log_2\frac{\Gamma_n(s,r)}{p(s)\Gamma(r)}. \tag{A21}$$

Now, with the hypothesis that $\tilde{P}(r)$, $\tilde{P}(r|s)$ are bounded and continuous almost everywhere (Ihara 1993), we have that in the $n\to\infty$ limit $\Gamma(r|s)\to\tilde{p}(r|s)$ and in the same limit the (infinitely) discretized information (A21) tends to the continuous one (24), whereas the infinitely discretized term (A8) tends to that derived in the continuous case (28).

Appendix B.

In this appendix we briefly discuss the Bayesian-like method we use to extract the values of $\tilde{R}_s$, $\tilde{R}$ from the data. Let us first recall some terminology from Bayes theory (see, for example, Wolpert and Wolf (1995) and the recent review of MacKay (1995) on Bayes theory and data modelling). The meaning of the various parameters is explained in section 3.1.

If we wish to measure a function $G\left(\{P(r|s)\}\right)$ of the set of probabilities $\{P(r|s)\}$ and we know the prior probability distribution of the probabilities $\mathcal{P}\left(\{P(r|s)\}\right)$, then the Bayesian estimate of the function $G\left(\{P(r|s)\}\right)$ has the following expression as a function of the set of experimental data $\{n(r|s)\}$:

$$\hat{G}\left(\{n(r|s)\}\right) = \int \left(\prod_r \mathrm{d}P\,(r|s)\right) \mathcal{P}\left(\{P(r|s)\} \,|\, \{n(r|s)\}\right) G\left(\{P(r|s)\}\right) \quad \text{(B1)}$$

where $\mathcal{P}\left(\{P(r|s)\} \,|\, \{n(r|s)\}\right)$ is the 'posterior' conditional probability of the underlying probabilities *given* the experimental outcome which is calculated with Bayes's theorem:

$$\mathcal{P}\left(\{P(r|s)\} \,|\, \{n(r|s)\}\right) = \frac{\mathcal{P}\left(\{n(r|s)\} \,|\, \{P(r|s)\}\right) \mathcal{P}\left(\{P(r|s)\}\right)}{\mathcal{P}\left(\{n(r|s)\}\right)} \quad \text{(B2)}$$

where

$$\mathcal{P}\left(\{n(r|s)\}\right) = \int \left(\prod_r \mathrm{d}P\,(r|s)\right) \mathcal{P}\left(\{n(r|s)\} \,|\, \{P(r|s)\}\right) \mathcal{P}\left(\{P(r|s)\}\right) \quad \text{(B3)}$$

and the 'likelihood' probability distribution is binomially distributed:

$$\mathcal{P}\left(\{n(r|s)\} \,|\, \{P(r|s)\}\right) = N_s! \prod_r \frac{P(r|s)^{n(r|s)}}{n(r|s)!}. \quad \text{(B4)}$$

The procedure we use here to evaluate $\tilde{R}_s$ is the following:

- We first pick for $\tilde{R}_s$ one of the allowed values, $R_s \leqslant \tilde{R}_s \leqslant R$.
- We construct, by using (B1), the Bayes estimate $\hat{P}(r|s)$ 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 $\tilde{R}_s - R_s$ empty bins is a different constant, fixed by requiring that the probability of that bin being empty is h times larger than the probability of being occupied, where $h_s = N_s/R_s$. This last requirement simply reflects the fact that when the responses are concentrated into a few bins (i.e. high 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 $\tilde{R}_s$ 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 the Bayes approach is expected to give better results.
- We pick other values for $\tilde{R}_s$ and we finally choose as an estimate for $\tilde{R}_s$ the value of $\tilde{R}_s$, which gives the expectation value of the number of occupied bins that is closest to the experimental value of R_s. It is easy to show that this expectation value has the following expression:

$$\langle R_s \rangle = \sum_r \left[1 - \left(1 - \hat{P}(r|s)\right)^{N_s}\right]. \quad \text{(B5)}$$

- The procedure is the same for the evaluation of $\tilde{R}$, the only difference being that the Bayesian estimate for $\hat{P}(r)$ should be calculated now from N and not N_s trials.

As shown in figure 1(*a*) this procedure, although based on a very crude ansatz on the prior, is sufficient to give reliable results even up to relatively small values of N.

Acknowledgments

Our work was stimulated by that of Edmund Rolls, Barry Richmond, John Hertz and their co-workers. Paolo De Los Rios read the manuscript carefully and Marco Billó helped with the figures. Partial support was from CNR, INFM, INFN and the EC's HCM Programme.

References

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

Atick J J and Redlich A N 1990 Towards a theory of early visual processing *Neural Comput*. **2** 308–20

Bialek W, Rieke R, de Ruyter van Steveninck R R and Warland D 1991 Reading a neural code *Science* **252** 1854–7

Carlton A G 1969 On the bias of information estimates *Psych. Bull*. **71** 108–9

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

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

Dong D W and Atick J J 1995 Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus *Network: Comput. Neural Syst.* **6** 159–78

Gawne T J and Richmond B J 1993 How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci*. **13** 2758–71

Golomb, Hertz J A, Panzeri S, Richmond B J and Treves A 1996 in preparation

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 Sys*. **3** suppl 91–103

Ihara S 1993 *Information Theory for Continuous Systems* (Singapore: World Scientific)

Kjaer T W, Hertz J A and Richmond B J 1994 Decoding cortical neuronal signals: network models, information estimation and spatial tuning *J. Comput. Neurosci*. **1** 109–39

Levine M W and Troy J B 1986 The variability of maintained discharge of cat dorsal-lateral geniculate cells *J. Physiol*. **375** 219–46

MacKay D J C 1995 Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks *Network: Comput. Neural Syst.* **6** 469–505

McClurkin J W, Gawne T J, Optican L M and Richmond B J 1991 Lateral geniculate neurons in behaving primates. II. Encoding of visual information in the temporal shape of the response *J. Neurophysiol*. **66** 794–808

Miller G A 1955 Note on the bias on information estimates *Information Theory in Psychology; Problems and Methods II-B* pp 95-100

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** 162–78

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** 305–10

Panzeri S and Treves A 1995 Correcting measures of information for limited data samples *Proc. 3rd Workshop: Neural Networks: from biology to high energy physics (Elba) Int. J. Neural Sys*. **7** (Suppl) 133–7

Rolls E T, Critchley H D and Treves A 1995a The representation of olfactory information in the primate orbitofrontal cortex *J. Neurophysiol*. in press

Rolls E T, Tovee M J and Treves A 1995b Information in the neuronal representation of individual stimuli in the primate temporal visual cortex, submitted

—— 1995c The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex, submitted

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

Tovee M J, Rolls E T, Treves A and Bellis R P 1993 Information encoding and the responses of single neurons in the primate temporal visual cortex *J. Neurophysiol*. **70** 640–54

Treves A 1990 Graded-response neurons and information encodings in autoassociative memories *Phys. Rev.* A **42** 2418–30

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

Wolpert D H and Wolf D R 1995 Estimating functions of probability distributions from a finite set of samples *Phys. Rev.* E **52** 6841–54