LETTER Communicated by Emery Brown

On Decoding the Responses of a Population of Neurons from Short Time Windows

Stefano Panzeri
Neural Systems Group, Department of Psychology, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, U.K.

Alessandro Treves
SISSA—Programme in Neuroscience, 34013 Trieste, Italy

Simon Schultz
Edmund T. Rolls
University of Oxford, Department of Experimental Psychology, Oxford OX1 3UD, U.K.

The effectiveness of various stimulus identification (decoding) procedures for extracting the information carried by the responses of a population of neurons to a set of repeatedly presented stimuli is studied analytically, in the limit of short time windows. It is shown that in this limit, the entire information content of the responses can sometimes be decoded, and when this is not the case, the lost information is quantified. In particular, the mutual information extracted by taking into account only the most likely stimulus in each trial turns out to be, if not equal, much closer to the true value than that calculated from all the probabilities that each of the possible stimuli in the set was the actual one. The relation between the mutual information extracted by decoding and the percentage of correct stimulus decodings is also derived analytically in the same limit, showing that the metric content index can be estimated reliably from a few cells recorded from brief periods. Computer simulations as well as the activity of real neurons recorded in the primate hippocampus serve to confirm these results and illustrate the utility and limitations of the approach.

1 Introduction

Understanding the way in which stimuli are represented by neuronal responses operationally amounts to being able to reconstruct (that is, identify or decode) the external correlates from the responses. Thus, decoding is useful in providing both insight into how the brain itself might use the information encoded in the neuronal responses and a tool to quantify the accuracy

with which the variables characterizing the stimuli can be estimated from the observation of the activity of populations of neurons (Georgopoulos, Schwartz, & Kettner, 1986; Seung & Sompolinsky, 1993; Abbot, 1994; Snippe, 1996; Rolls, Treves, & Tovée, 1997; Zhang, Ginzburg, McNaughton, & Sejnowski, 1998). Moreover, when used in particular information measures, decoding is often an essential part of the procedure for their estimation, needed in order to reduce the dimensionality of the response space (Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1996; Rolls & Treves, 1998).

Using the limit of short time windows can facilitate analysis of the representation of information by neurons. First, there is substantial evidence that in many cases information is transmitted by neuronal activity in very short times, suggesting that it may also be decoded in short times. At the level of single cortical cells in primates, much of the information that can be extracted from their responses (even to static stimuli) is found to be present already in rather short periods of 20–50 ms (Oram & Perrett, 1992; Tovée, Rolls, Treves, & Bellis, 1993; Heller, Hertz, Kjaer, & Richmond, 1995). At the level of populations, information is transmitted much faster, at least to the extent that the different cells in the population carry independent information (Rolls, Treves, & Tovée, 1997). Event-related potential studies of the human visual system provide further evidence that the processing of information in a multiple-stage neural system can be extremely rapid (Thorpe, Fize, & Marlot, 1996). Second, over time windows much shorter than the mean interspike interval, the response of each individual cell can be taken to be binary: it either emits a spike or does not. This simplifies the estimation of accuracy variables derived from the response; in particular, with populations of a few cells, it again reduces the dimensionality of their response space to allow the estimation of transmitted information.

In this article we combine the two approaches by studying the accuracy of decoding procedures in reconstructing the information transmitted by the activity of neuronal populations in short timescales. The simplification brought about by the short time limit makes it possible to establish analytical results of practical import—for example, that it is in most cases better to estimate transmitted information directly from the stimuli decoded as most likely rather than from the full distribution of stimulus likelihoods, in that less or no information is lost in the decoding step itself. Analytical results are valid only in the limit of short timescales, and since they derive from the first-order terms of a Taylor expansion in time, to which single cells contribute additively and independently, they cannot provide clues on the effects of correlations.[1] However both computer simulations and the analysis of real data indicate that the range of validity of the main conclusions extends to time windows and population sizes typical of many neuro-

[1] The effects of correlations are studied in a companion paper (Panzeri, Schultz, Treves, & Rolls, 1999) that makes use of second-order terms in the expansion.

physiological recording experiments, thus suggesting appropriate uses of decoding procedures in practical cases.

2 Basic Concepts

2.1 Stimulus-Response Information and Limited Sampling. In this article we consider experiments in which the responses of several cells to repeated presentations of the same stimuli are recorded. Stimuli are taken from a discrete, nonmetric set $\mathcal{S}$ of S elements, each occurring with probability $P(s)$.[2] Responses are described simply by a vector $\mathbf{n}$ of spike counts, to which each of C neurons contributes a component given by the number of spikes n_c emitted in the time window $[t_0, t_0 + t]$. This description does not assume rate coding, but simply derives from the fact that at the level of first-order terms in an expansion in t, more complex descriptions of the response aimed at capturing temporal codes, for example, are not relevant. In the $t \to 0$ limit, in fact, the only possible responses are 0 or 1 spike per neuron. Further, different cells could be recorded sequentially or simultaneously, since this makes no difference at the first order in t. We treat elsewhere the effects of correlations among cells, which obviously can be satisfactorily observed only with simultaneous recording. The probability of events with response $\mathbf{n}$ is denoted as $P(\mathbf{n})$, and the joint probability distribution as $P(s, \mathbf{n})$.[3]

The information that the neuronal responses convey about the set of stimuli can be written as a function of response probabilities and of the time window length t (Shannon, 1948):

$$I(t) = \sum_{s \in \mathcal{S}} \sum_{\mathbf{n}} P(s, \mathbf{n}) \log_2 \frac{P(s, \mathbf{n})}{P(s)P(\mathbf{n})} \,. \tag{2.1}$$

Ideally, one would measure $I(t)$ by directly applying equation 2.1. In practice, however, $P(s, \mathbf{n})$ is not available, and one has to use instead the frequency table computed on the basis of N stimulus-response pairs, $P_N(s, \mathbf{n})$. If $P_N(s, \mathbf{n})$ is simply inserted in equation 2.1 in place of $P(s, \mathbf{n})$, it is known that information is usually grossly overestimated because of the undersampling due to the limited number of trials usually available (Miller, 1955). A number of methods, including some based on bootstrap (Optican, Gawne, Richmond, & Joseph, 1991) or jackknife (Efron, 1982) procedures, have been

[2] We consider nonmetric sets of stimuli for the sake of generality because in many experiments, the set of stimuli is a complex set of objects, like two-dimensional visual patterns or faces, for which a notion of distance between stimuli is not well defined. An extension to continuous stimuli is given in the appendix.

[3] The response probabilities $P(s, \mathbf{n})$ are a function of the time window length t, as made explicit in the short time limit in equations 2.6 and 2.7. The time dependence of the various information quantities introduced in the text arises from the dependence of the response probabilities on time.

developed to correct for the sampling bias. It is possible, for example, to subtract a correction term calculated from the data, which results in equivalent accuracy with samples an order of magnitude smaller in sizes (Treves & Panzeri, 1995). This term, δI, is dependent on any regularization (e.g., binning or smoothing) of the responses, which should be kept minimal because regularization itself causes an information loss. If the responses are discretized into R bins, δI depends solely on the number R_s of bins relevant (i.e., with some probability of being occupied) for each stimulus (Panzeri & Treves, 1996):

$$\delta I = \frac{1}{2N \ln 2} \left[\sum_s R_s - R - (S-1) \right]. \tag{2.2}$$

The correction is reliable, as a rule of thumb, if there are at least as many trials per stimulus as response bins R. This indicates that the number of trials required to control undersampling grows exponentially with population size, because $R = \prod_c n_c^{\max} \simeq (n^{\max})^C$, even when finite sampling corrections (Treves & Panzeri, 1995) are applied. Thus a direct calculation of transmitted information from a large population of cells is in practice impossible with the amount of data that can be obtained from a mammalian cortical recording session. Nevertheless, for very short time windows, such that one or two spikes are emitted at most by any cell, it is possible to calculate this "true information" for ensembles comprising up to a few cells. This will provide a useful comparison for the decoded information values obtained below.

2.2 Taylor Expansion in the Short Time Limit. The instantaneous rate at which information accumulates from time t_0 can be examined by considering directly the time derivatives of information at t_0. To first order, $I(t)$ is approximated by the Taylor expansion

$$I(t) = t\, I_t + O(t^2), \tag{2.3}$$

where I_t is the first time-derivative of $I(t)$ calculated at t_0. We assume that the firing-rate distribution reflects a stationary random process: individual trials to a given stimulus are drawn at random from the same probability $P(\mathbf{n}|s)$ conditional to stimulus s and are therefore statistically indistinguishable. Under this assumption, the mean firing rate $\overline{r}_c(s)$ (i.e., the mean spike count divided by t) is a well-defined quantity. The bar denotes averaging over population responses $\mathbf{n}$ with probability $P(\mathbf{n}|s)$ conditional to stimulus s:

$$\overline{(\cdot)} \equiv \sum_{\mathbf{n}} P(\mathbf{n}|s)(\cdot). \tag{2.4}$$

We also assume that the probability of observing one spike emitted by a cell c in the time window $[t_0, t_0 + t]$ conditional on the emission of a different

spike by any other neuron in the population, when a stimulus s is presented, is proportional to t,

$$P(\text{spike from cell } i \text{ in } [t_0, t_0 + t] \mid \text{spike from cell } j \text{ in } [t_0, t_0 + t]) = \overline{r}_i(s)\, t(1 + \gamma_{ij}(s)). \tag{2.5}$$

$\gamma_{ij}(s)$ is a scaled cross-correlation factor and measures the fraction of coincidences above (or below) that expected from uncorrelated responses, normalized to the number of coincidences expected in the uncorrelated case. If we call conditional firing rate the average rate of a cell c conditional on at least one spike having been emitted by a different neuron in the same window, equation 2.5 just means that all instantaneous conditional firing rates are finite. This is a very natural assumption and is violated only in the rather implausible case of spikes locked to one another with infinite time precision. In any case, the validity of equation 2.5 can be verified for any given data set.

The t expansion of response probabilities is then essentially an expansion in the total number of spikes emitted by the population in response to a stimulus. The only responses with nonzero probabilities up to the order t^k are those with up to k spikes in total from the whole population; the only events with nonzero probability are therefore to first order in t those with no more than one spike emitted in total:

$$p(\mathbf{0}|s) = 1 - t \sum_{c=1}^{C} \overline{r}_c(s) + O(t^2) \tag{2.6}$$

$$p(\mathbf{e}_c|s) = t\, \overline{r}_c(s) + O(t^2) \qquad c = 1, \ldots, C, \tag{2.7}$$

where $\mathbf{0}$ is the response vector with zero spikes emitted by each cell; $\mathbf{e}_c$ is the response vector with one spike in the cth cell component and zero in the other ones. The first-order probabilities do not depend on the correlation coefficients $\gamma_{ij}(s)$; the effects of correlations are relevant only at second order and are studied in (Panzeri et al., 1999).

Substituting the first-order probabilities (see equations 2.6 and 2.7) into the definition of information (see equation 2.1), we obtain the generalization at the population level of the formula derived for the case of single cells by Bialke, Rieke, de Ruyter van Steveninck, & Warland, (1991) and Skaggs, McNaughton, Gothard, and Markus (1993):

$$I_t = \sum_{c=1}^{C} \sum_{s \in \mathcal{S}} P(s) \overline{r}_c(s) \log_2 \frac{\overline{r}_c(s)}{\overline{r}_c}, \tag{2.8}$$

where $\overline{r}_c = \sum_s P(s) \overline{r}_c(s)$, the grand mean rate of cell c to all stimuli. Since only two spiking events (zero or one spike) are relevant to first order, this is

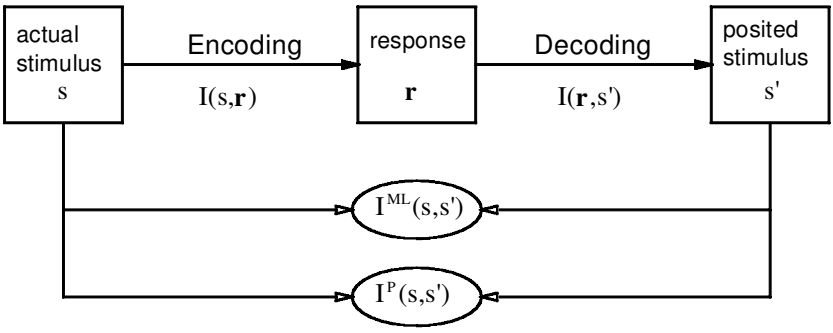

Figure 1: Schematic description of the encoding-decoding relationship.

in fact the first derivative of the information carried by the full spike train, and not only by the mean firing rates.

It is interesting to note from equation 2.8 that as long as conditional rates do not diverge as $t \to 0$, the characteristic timescale for information processing in a population is just C times shorter than the average timescale for single cells. For large enough populations, therefore, most of the information carried by the network can be extracted from time windows so short that the responses of individual cells are all binary (zero or one spike).

2.3 Decoded Information. Other than by focusing on very short windows, transmitted information measures from populations can be obtained also by first replacing neuronal responses with any functions of the responses themselves, chosen such as to have lower entropy (i.e., lower dimensionality), or fewer possible states. Decoding which compresses the original high-dimensional response space into a set that has the same structure as the stimulus set (the set of predicted, or posited, stimuli), is an interesting example of such a transformation. This is in some cases a drastic reduction, but it is appropriate because the minimum number of regularized response classes that do not throw away information about which stimulus has occurred is the number of stimuli. Therefore, if it is accurate, decoding is valuable in itself and also provides a useful tool to estimate the information conveyed by large populations of cells, as schematized in Figure 1.

The original transmitted information can be estimated by considering the mutual information between the stimuli and the most likely stimulus in each trial—what we shall call the maximum likelihood information, I^{ml} (Gochin, Colombo, Dorfman, Gerstein, & Gross, 1994; Victor & Purpura, 1996; Rolls, Treves, & Tovée, 1997; Rolls & Treves, 1998). A slightly

more complex variant (Heller et al., 1995; Gawne, Kjaer, Hertz, & Richmond, 1996; Rolls, Treves, & Tovée, 1997) includes a step that extracts from the responses in each trial not only the single most likely stimulus, but all the probabilities that each of the possible stimuli in the set was the actual one. The joint probabilities of actual and posited stimuli can be averaged across trials, and another information quantity, I^p, can be calculated from such a probability matrix of presenting a stimulus and decoding another one. Neither I^{ml} nor I^p, calculated after decoding, can be higher than the information I contained in the neuronal responses, because the decoding step, if performed correctly, cannot add new information of its own. On the other hand, in order to use I^{ml} or I^p as reasonable approximations to I, the decoding procedure should be efficient; the stimulus should be reconstructed with minimal error so that the difference between "true" information in the neuronal responses and decoded information remains as small as possible.

The distinction between I^{ml} and I^p is different from the choice of a specific decoding algorithm, that is, how stimulus likelihoods are estimated from the responses. Common decoding algorithms include Bayesian decoding (Földiák, 1993), population vector methods (Georgopoulos et al., 1986), template matching (Wilson & McNaughton, 1993), biologically plausible decoding (Seung & Sompolinsky, 1993; Rolls, Treves, & Tovée, 1997), but only two examples will be used in this article. The first is aimed at maximally efficient information reconstruction and therefore uses Bayesian decoding based on the responses probabilities. The second is Euclidean distance decoding (Rolls, Treves, Robertson, Georges-François, & Panzeri, 1998), which estimates the likelihood of any stimulus as a function of the Euclidean distance between the response vector of a test trial and the mean response vector to that stimulus, and is aimed instead at understanding how much information can be decoded by biologically plausible operations.

In principle, optimal decoding uses Bayes' rule:

$$P(s'|\mathbf{n}) = \frac{P(\mathbf{n}|s')P(s')}{P(\mathbf{n})}\,, \tag{2.9}$$

but this requires knowledge of the response probabilities $P(\mathbf{n}|s)$. In practice, this means fitting $P(\mathbf{n}|s)$ to a model function. Obviously, probability models that are far from the actual probabilities may lead to information loss. However, in the short time limit, the choice of a response probability model is not important, because the response probabilities in this limit depend only on the mean firing rates, not on any other detail of the distribution.

To avoid biasing the estimation of conditional probabilities, the responses used in estimating $P(\mathbf{n}|s)$ (called the *training* responses for what is a cross-validation procedure) should not include the particular test trial for which $P(s'|\mathbf{n})$ is going to be derived. Summing over different test trial responses to the same stimulus s, one can extract the probability that by presenting

stimulus s, the neuronal response is interpreted as having been elicited by stimulus s',

$$P(s'|s) = \sum_{\mathbf{n}\in test} P(s'|\mathbf{n})P(\mathbf{n}|s)\,, \tag{2.10}$$

Note that in equation 2.10 we have used the identity $P(s'|\mathbf{n}, s) = P(s'|\mathbf{n})$, which simply states that stimulus decoding is made only on the basis of the current response, without any regard to which stimulus was actually presented. Although there is growing evidence that simple neural networks can perform efficient stimulus estimation (Pouget, Zhang, Deneve, & Latham, 1998), it is interesting to consider decoding algorithms that make use of only simple neurophysiologically plausible operations that could be performed by downstream neurons, such as dot product summations (which might be followed by thresholding, scaling, and other single cell nonlinearities). An example of this approach, which is alternative to the Bayesian optimal decoding and is meaningful in the limit of short times is Euclidean distance (ED) decoding (Rolls et al., 1998). This algorithm estimates the likelihood of each stimulus as a function (e.g., exponentially decreasing) of the Euclidean distance between the response vector $\mathbf{n}$ during the test presentation and $\mathbf{n}(s)$, the mean response vector to stimulus s during the training trials:

$$P(s|\mathbf{n}) \propto \exp\left(-\frac{|\mathbf{n} - \mathbf{n}(s)|^2}{2\sigma^2}\right), \tag{2.11}$$

where σ is the standard deviation of the responses across all trials and stimuli. This decoding step is biologically plausible in that it might be performed by a cell that receives the test vector as a set of input firings and produces an output that depends on its synaptic weight vector, which might represent the average response vector to a stimulus. A simpler version of ED decoding is a decoding procedure based on the scalar (or dot) product of the test response vector with the average response vectors to each of the stimuli (Rolls, Treves, & Tovée, 1997). We will not discuss dot product decoding except to note that in the short time limit, it becomes the same as ED decoding, provided that a sensible rule for stimulus prediction is assigned when decoding the **0** response.

Having estimated the probabilities that the test trial response has been elicited by each of the stimuli, 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 fraction of correct decodings f_{cor}, or alternatively by the mutual information extracted from the probability table $Q(s^p|s)$,

$$I^{ml}(t) = \sum_{s,s_p\in\mathcal{S}} P(s)Q(s^p|s)\log_2\frac{Q(s^p|s)}{Q(s^p)}\,, \tag{2.12}$$

where $Q(s^p|s)$ is 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 not only reflects, like the percentage correct, the number of times the decoding is exact, but also the distribution of wrong decodings. Of course, the matrix of decodings $Q(s^p|s)$, and therefore the information I^{ml}, depend on the decoding algorithm used.

The mutual information I^p is given by[4]

$$I^p(t) = \sum_{s,s' \in \mathcal{S}} P(s,s') \log_2 \frac{P(s,s')}{P(s)P(s')} \tag{2.13}$$

I^p reflects also the degree of certainty with which each single trial has been decoded (Treves, 1997).

3 Analytical Results

3.1 Maximum Likelihood Information from Short Windows. The results obtained for Bayesian decoding, equation 2.9, are considered first and extended to ED decoding, equation 2.11, at the end of this subsection. To first order in t, all that is needed for Bayesian decoding are the conditional probabilities of posited stimuli $P(s|\mathbf{n})$ for the $C + 1$ possible first-order responses $\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_C$. The conditional probabilities $P(s|\mathbf{n})$ can be explicitly calculated by substituting the response probabilities (see equation 2.6) into Bayes' rule (see equation 2.9). $P(s|\mathbf{n})$ and the most likely stimulus depend only on the mean firing rates of the cells in response to the different stimuli and on the probability of presentation of the stimuli themselves:

- Call the most likely stimulus when response $\mathbf{0}$ is observed the "worst stimulus": *if* all stimuli are equiprobable, then by equation 2.6, the most likely stimulus s^p for response $\mathbf{0}$ is the stimulus that elicits the smallest population response, that is, the stimulus s that minimizes $\sum_c \overline{r}_c(s)$. Suppose that this worst stimulus has a degeneracy D, that is, there are D distinct stimuli with either the very same minimum response (if equiprobable) or with the responses in the exact proportion to compensate the extra $P(s)$ factor (if not). Denote these stimuli as sw_a, with the additional index a labeling the degenerate stimuli, $a = 1, \ldots, D$.

[4] The difference between the $Q(s^p|s)$ and $P(s|s')$ can be appreciated by noting that each vector comprising a given trial contributes (before normalization by dividing by the number of trials) 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. As a consequence, I^{ml} must be corrected with the correction term corresponding to the "quantized" case, equation 2.2, whereas I^p must be corrected with the term derived for the "smoothed" case, see (Panzeri & Treves, 1996).

- Call the most likely stimulus when response $\mathbf{e}_c$ is observed the "preferred" (or "best") stimulus for cell c: *if* all stimuli are equiprobable, then by equation 2.7 the most likely stimulus s^p for the response $\mathbf{e}_c$ is the stimulus that maximizes the mean response $\overline{r}_{s;c}$ of the cell c that fired. Denote the best stimulus for cell c as $sb(c)_a$, with the subscript a again labeling the possibly D_c degenerate best stimuli for that cell, $a = 1, \ldots, D_c$.

It is important to note that the stimuli decoded by the $C + 1$ events $\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_C$ may not all be different.[5] The number of the stimuli that have a nonzero probability to be decoded is a number that we call $D+K$, where D as noted is the "worst stimulus degeneracy" and K is the number of stimuli that are predicted by any of the $\mathbf{e}_c$ responses and are distinct from one another and from the worst stimulus. $D + K$ may be, to first order in t, either greater or smaller than the number of events $C + 1$ (depending on the degeneracies and on the overlapping of preferred stimuli from different cells). Since the ordering of the stimuli is arbitrary, one can assign to the (degenerate) worst stimuli sw_a the index $s = 0, \ldots, D - 1$. Similarly, call $s = D, \ldots, D + K - 1$ the K distinct stimuli predicted by an $\mathbf{e}_c$ response. The set of cells that have $s = k(k = 0, \ldots, D + K - 1)$ as a preferred stimulus is denoted $\mathcal{C}(k)$.

The maximum likelihood information (see equation 2.12) cannot exceed the information contained in the neuronal responses (see equation 2.1), as noted above. On the other hand, if the stimulus reconstruction is performed with minimal information loss, then equation 2.12 should be very close to equation 2.1. Expanding the maximum likelihood information as a power series in t, $I^{ml} = t\ I_t^{ml} + O(t^2)$, the information rate I_t^{ml} estimated through maximum likelihood information may be compared with the full information rate I_t contained in the neuronal responses (see equation 2.8). The analysis, together with the examples considered in section 4, shows that in the short time limit, the two information quantities can be equal: $I_t^{ml} = I_t$.

The table $Q(s^p|s)$ can be calculated. If s^p is one of the (degenerate) worst stimuli, $s^p = 0, \ldots, D - 1$, then s^p is predicted whenever we observe a $\mathbf{0}$ response or an $\mathbf{e}_c$ response $[c \in \mathcal{C}(s^p)]$. The stimuli $s^p = D, \ldots, D + K - 1$ (i.e., s^p is a preferred stimulus for some cells and it is not one of the worst stimuli) are predicted whenever we observe a corresponding $\mathbf{e}_c$ response $[c \in \mathcal{C}(s^p)]$. The remaining possible stimuli $s = D + K, \ldots, S - 1$ are never predicted.

Therefore the matrix containing the fractions of decodings has the form:

$$Q(s^p|s) = \sum_{c \in \mathcal{C}(s^p)} \frac{P(\mathbf{e}_c|s)}{D_c} + \frac{1}{D} P(\mathbf{0}|s) \qquad s^p = 0, \ldots, D - 1$$

[5] As an example, two cells c_1 and c_2 may share one of the preferred stimuli, $sb(c_1)_a = sb(c_2)_b$. Alternatively, one of the (degenerate) preferred stimuli for cell c_3 may coincide with one of the (degenerate) worst population responses, $sw_a = sb(c_3)_b$.

$$Q(s^p|s) = \sum_{c\in\mathcal{C}(s^p)} \frac{P(\mathbf{e}_c|s)}{D_c} \qquad s^p = D, \ldots, D+K-1$$
$$Q(s^p|s) = \quad 0 \qquad\qquad\qquad\quad s^p = D+K, \ldots, S-1. \tag{3.1}$$

The estimated information rate I_t^{ml} can be computed by first inserting the probabilities equations 3.1 into 2.12 and then expanding 2.12 in powers of t (using the well-known expansion for the logarithm: $\ln(x) \simeq -1 + x$ for $x \to 0$). The result is as follows:

$$I_t^{ml} = \sum_s P(s) \sum_{k=D}^{D+K-1} \left(\sum_{c\in\mathcal{C}(k)} \frac{\overline{r}_c(s)}{D_c} \right) \log_2 \left[\frac{(\sum_{c\in\mathcal{C}(k)} \overline{r}_c(s)/D_c)}{(\sum_{c\in\mathcal{C}(k)} \overline{r}_c/D_c)} \right] . \tag{3.2}$$

Notice that the "worst" stimuli do not contribute to the sum over predicted stimuli in equation 3.2. One can show that due to the usual log-sum inequality, the maximum likelihood information rate I_t^{ml} is bounded from above by the true value of the rate of information contained in the neuronal responses, $I_t^{ml} \leq I_t$. The difference between $I_t - I_t^{ml}$ precisely quantifies (once multiplied by t) the information loss due to the decoding procedure to first order in t. When is all the information contained in the neuronal responses preserved after decoding, independent of the number of cells considered? The inequality becomes an equality only if the following conditions are met. First, there must be no overlap between the preferred stimuli of some of the cells and the worst population responses. Second, for each of the preferred stimuli k that are distinct from one another and from the worst population responses (i.e., $k = D, \ldots, D+K-1$), the ratio $\overline{r}_c(s)/\overline{r}_c$ must be constant across all cells $c \in \mathcal{C}(k)$ for each predicted stimulus k and for each actual stimulus s. In other words, if each of the $C+1$ events $\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_C$ predicts a different stimulus, then all the information present in neuronal responses is fully decodable to first order in t. When there is overlap between stimuli predicted by the $C+1$ events $\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_C$, then all the information is fully decodable *if and only if* there is no overlap between the preferred stimuli of some of the cells and the worst population responses *and*, if two or more cells share the same preferred stimulus, they have the same response profile (up to a proportionality constant) to all the different stimuli in the set.

It is interesting to note that according to equation 3.2, the difference between the true and the maximum likelihood information I^{ml} is in general expected to be very small if one or two cells are considered and to increase progressively as the number of cells C increases: with many cells, overlapping between predicted stimuli by different cells becomes more likely. This is indeed what is found when estimating information with I^{ml} not only in the short time limit, but also for longer time windows, as shown by the simulations presented below. There is a theoretical explanation for this analysis and expectation being confirmed for intermediate times: it is possible to show, by the very same formalism used here, that if one extends the analysis to

any arbitrary order in the t expansion, a sufficient condition for no information loss in the decoding is that each event predicts a different stimulus. The number of possible population responses at any order in t (and thus the probability of overlapping predicted stimuli) increases with the number of cells in the population, and therefore I^{ml} tends to underestimate the true information more for larger sets of cells, even for intermediate times.

Now replacing Bayesian decoding with the biologically plausible ED decoding, equation 2.11, exactly the same results are found. In fact it is possible to show that the most likely stimulus predicted by equation 2.11 when response **0** is observed is, as in the Bayesian case, the worst population response; the most likely stimulus predicted by equation 2.11 when response $\mathbf{e}_c$ is observed is again the best stimulus for cell c. Therefore all the information that can be extracted (through I^{ml}) with the Bayesian decoding procedure in short times can also be extracted by more crude, neuronal-like decoding algorithms. This finding has also been confirmed by computer simulations in the case of intermediate times (see section 4).

3.2 Probability Information. Turning now to I^p and to the relevant table, $P(s'|s)$, it can be shown, by using equations 2.10, 2.6, and 2.7, that to first order in t, $P(s'|s)$ can be written as:

$$P(s'|s) = P(s')\left[1 + t\sum_{c=1}^{C}\frac{(\overline{r}_c - \overline{r}_{s;c})(\overline{r}_c - \overline{r}_{s';c})}{\overline{r}_c}\right] + O(t^2)\,, \tag{3.3}$$

By substituting equation 3/3 into the definition of I^p, it follows that the first derivative of I^p is *always* zero:

$$I^p(t) \approx O(t^2)\,. \tag{3.4}$$

This means that I^p cannot estimate information transmission rates, and it gives poor estimates of information for relatively short times.

This result applies not only when information is decoded from several cells in short time windows but generalizes to other situations, such as the information contained in the response profile of a cell when its spike emission is temporally sparse.[6] This may account for some of the inconsistencies in the results presented by Heller et al. (1995), where the binary vector code (in which the presence or not of a spike in each 1 ms bin of the response constitutes a component of a 320-dimensional vector) contains much less (I^p) information than other simpler codes.

[6] If we divide the total recording time into successive time windows of length Δt, as $\Delta t \to 0$ the correlations between occurrence of spikes in different bins should shrink to zero, analogous with equation 2.5. Therefore, an analysis similar to ours can be applied in this case, the weakly correlated variable being the number of spikes in very short (e.g., 1 ms) time bins rather than the number of spikes emitted in a single time interval by different cells.

To understand how to use I^p to evaluate the redundancy in the information conveyed by different cells, as done, for example, by Gawne et al. (1996) and Rolls, Treves, & Tovée (1997), the dependence of I^p on the number of cells must be considered. The redundancy of a population is defined as one minus the ratio between the information carried by the population responses and the sum of the information carried by the individual cells (Gawne et al., 1996; Rolls, Treves, & Tovée, 1997). By expanding I^p in powers of the number of cells C instead in powers of t one would obtain, in analogy to equation 3.4, $I^p \propto C^2$. Therefore, using I^p may lead to an underestimation of the true redundancy, and one might find (for a few cells) an apparently synergistic representation where in fact there are no real synergistic effects.

Thus, although I^p certainly suffers less from limited sampling distortions (Panzeri & Treves, 1996), it tends to underestimate I more seriously than I^{ml} does. Note that I^{ml} is usually expected to contain more information than I^p in any case (since the decoding table based on the fraction of predicted stimuli should be more peaked along the diagonal than the table containing the probability of confusing two stimuli), although situations where $I^p > I^{ml}$ are certainly possible.[7] The $t \to 0$ analysis shows that for very short times, I^{ml} is dramatically more efficient at estimating the true information I.

3.3 Percentage Correct Predictions and the Metric Content. The percentage of correct decodings can be calculated directly as the trace of the matrix $Q(s^p|s)P(s)$ representing the fraction of trials in which a stimulus s is presented and a stimulus s^p is decoded. From equation 3.1 an expression for the fraction of correct guesses f_{cor} is obtained, which we present for the case of equiprobable stimuli:

$$f_{cor} \equiv \sum_s Q(s|s)\, P(s) = \frac{\left(1 + t\left(\sum_{c=1}^{C} (\overline{r}_{sb(c);c} - \overline{r}_{sw;c})\right)\right)}{S} + O(t^2). \quad (3.5)$$

This result is independent of any degeneracy and overlapping between maximally likely stimuli. The fraction of correct decodings is greater than $1/S$, because the term $\propto t$ in equation 3.5 is always nonnegative and equal to zero (and thus $f_{cor} = 1/S$) only if the information in the firing rates is zero. For a given set of stimuli, the value f_{cor} is not affected by the amount of degeneracy among decoded stimuli, or by overlaps in the response profiles of different cells.

[7] An example of $I^p > I^{ml}$ is the following. Suppose there are two stimuli. When the first stimulus is presented, half of the responses predict the first stimulus with probability 1.0, and the other half of the responses predict the second stimulus with probability 0.6. When the second stimulus is presented, half of the responses predict the first stimulus with probability 0.6, and the other half of the responses predict the second stimulus with probability 1.0. In this case, the percentage correct is equal to chance, $I^{ml} = 0$, but $I^p > 0$.

From the mutual information I^{ml} (see equation 3.2) and the fraction of correct decodings f_{cor} (see equation 3.5), it is possible to extract the metric content of the neuronal representation (Treves, 1997; Treves, Panzeri, Robertson, Georges-François, & Rolls, 1996) in short time windows. The metric content measure is based on the observation that for a given f_{cor}, the information may take a range of values depending on the amount of structure in the data. The information may range from a minimum $I_{\min}$, when incorrect decodings are distributed equally among all incorrect stimuli (thus all stimuli are encoded as equisimilar to each other), up to $I_{\max}$, when the stimuli fall into clusters or classes and the incorrect decodings are distributed with minimum entropy within the correct cluster. The expression for $I_{\max}$ for equiprobable stimuli (Treves, 1997) and its short time limit is:

$$I_{\max} = \log_2 S + \log_2 f_{cor} \simeq t \sum_{c=1}^{C} (\overline{r}_{sb(c);c} - \overline{r}_{sw;c}) + O(t^2). \tag{3.6}$$

Similarly

$$\begin{aligned} I_{\min} &= \log_2 S + f_{cor} \log_2 f_{cor} + (1 - f_{cor}) \log_2((1 - f_{cor})(S - 1)) \\ &= 0 + O(t^2). \end{aligned} \tag{3.7}$$

The metric content is (Treves, 1997; Rolls & Treves, 1998)

$$\lambda_m = \frac{I^{ml} - I_{\min}}{I_{\max} - I_{\min}}. \tag{3.8}$$

In the short time window limit, this becomes

$$\lambda_m = \frac{I_t^{ml}}{\sum_{c=1}^{C} (\overline{r}_{sb(c);c} - \overline{r}_{sw;c})} + O(t). \tag{3.9}$$

Treves et al. (1996) found the metric content to grow with the time window used to evaluate it, which they interpret as the gradual emergence of meaningful structure in neuronal activity. Equation 3.9 indicates that there is residual structure in the neuronal activity in very short time windows, and this is related to the rate of information transmission by the neuronal ensemble about the structured stimulus set. In fact, given that when $I_t^{ml} = I_t$ both derivatives reduce to sums of single cell contributions, λ_m can be seen from equation 3.9 to take a finite value even for single cells in the $t \to 0$ limit. Using populations simply allows better averaging (and modulation of the metric content by correlation effects), but a nontrivial λ_m value can be obtained even with single cells.

3.4 Cross-Validation. The study of stimulus decoding with short time windows is based on the assumption that the true firing rates of the cells are well determined (from a set of "training" trials used to establish the statistics of the data) and that the "test" trials follow the same probability distribution as the training trials. When the number of trials available is finite, there are finite sampling distortions on both firing-rate estimation and the distribution of test trials. Finite sampling distortions in the distribution of test trials lead, given a particular training set, to an average overestimation of the information that scales as the inverse number of test trials and can be corrected by the finite sampling corrections; the effect of the distortions on parameter estimation depends instead crucially on the length of the time window considered, the firing-rate separations, and the method used for cross-validation. In general, a cross-validation procedure that makes an efficient use of the data is a jackknife (Efron, 1982) cross-validation consisting of using only one response as the test trial and the remaining as training data, and then averaging over all the possible choices of that test trial. In extreme cases, though, when the training set is not large and the firing rates are very low (or equivalently the time window very short), and the temporary exclusion of a particular trial (which, for example, contains the only spike recorded in response to a given stimulus) from the training set leads to a substantial trial-by-trial redistribution of preferred and worst stimuli, the use of jackknife cross-validation can lead to systematic errors in the estimation of information and percentage correct. In fact, this applies not only to jackknife cross-validation but to any other cross-validation method where the training set changes with the particular trial considered. Although other cross-validation methods like dividing the data into two separate sets of test data and (test-trial-independent) training data can be safer in this case, they cannot always be applied because they require more data. Therefore one practical approach, when decoding the information transmitted in very short time windows, is to check if the results are affected by those problems. In particular, the analysis developed here allows some checks for inconsistencies in the information estimation. First, for time windows short enough that those problems may be important, the analytical approximations (see equations 2.8 and 3.2) to the information should be reliable, and therefore the application of decoding procedures can be checked against analytical formulas. Second, for such short time windows, one can also evaluate the information for up to a few cells directly, making use of finite sampling corrections.

4 Computer Simulations

Simulations were based on samples of few cells firing independent Poisson responses (see Figures 2 and 3) or with correlated firing (see Figure 4), with stimulus-dependent mean firing rates. Time windows of between 25 and

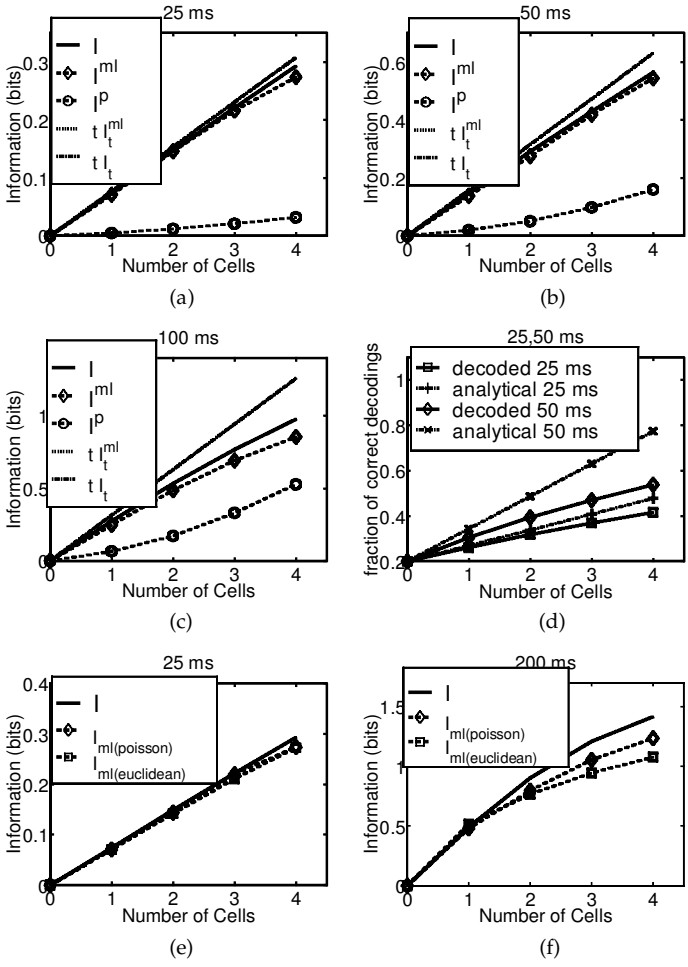

Figure 2: Perfect $t \to 0$ decoding. Four Poisson firing cells were simulated, each with a different nondegenerate preferred stimulus, and, in addition, a fifth stimulus, which elicited the worst population response. (a–c) Information estimators for different time windows, with Bayesian decoding. I_t and I_t^{ml} coincide. Also for short times, I^{ml} yields an excellent approximation to I; small losses in I^{ml} are due to second-order effects. I^p instead approaches zero for $t \to 0$, and note the artifactual superlinear growth with the number of cells. (d) Comparison of the percentage correct decoding with its $t \to 0$ analytical approximation, which is seen to be accurate over shorter t and C ranges than the linear (first-order) approximations to I^{ml} and I. (e–f) Comparison of Bayesian with ED decoding. The Poisson model included in the Bayesian algorithm matches, by construction, the statistics of the simulations. Nevertheless, even the more biologically plausible ED algorithms yield a reasonable estimate of the full I, at least for short times (the two algorithms are seen analytically to be equivalent in the $t \to 0$ limit).

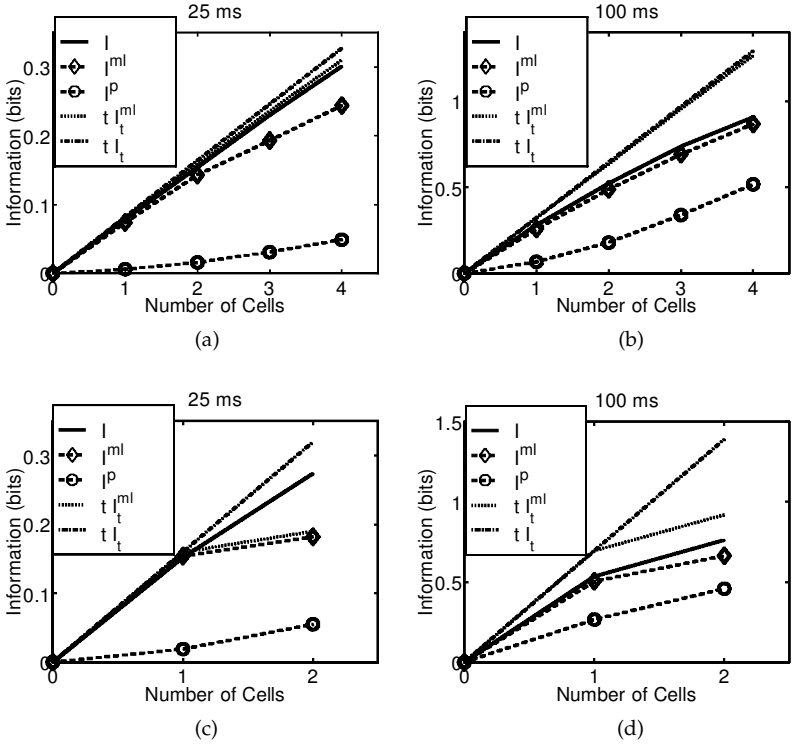

Figure 3: Mismatches between cells and stimuli decrease decoding efficiency. (a–b) When three of the four cells have the same nondegenerate preferred stimulus and the fourth has a different preferred stimulus, the information loss $I - I^{ml}$ is more marked, but only at short windows. $t\ I_t^{ml}$ is slightly smaller than the first-order full information $t\ I_t$. Here the worst stimulus was, again, different from each preferred stimulus. (c–d) Two cells responding to just two stimuli. Although the cells have different preferred stimuli, one of the preferred stimuli coincides with the worst stimulus. As expected, for the shorter window there is a large decoding loss, $I - I^{ml}$, when the two cells are considered together. Interestingly, the loss is minor for the longer window, indicating that higher-order effects (in t) may contribute positively to decoding efficiency. Bayesian Poisson decoding throughout the figure.

200 milliseconds were generated. Firing rates in response to stimuli ranged from 0 Hz to a peak firing rate of 15 Hz in order to operate in the same regime as real hippocampal spatial view cells (analyzed in the next section) with peak rates of 10 to 20 Hz and near-zero spontaneous activity (Rolls, Robertson, & Georges-François, 1997; Rolls et al., 1998). One hundred presentations were generated for each of the equiprobable stimuli in the set. Mean firing

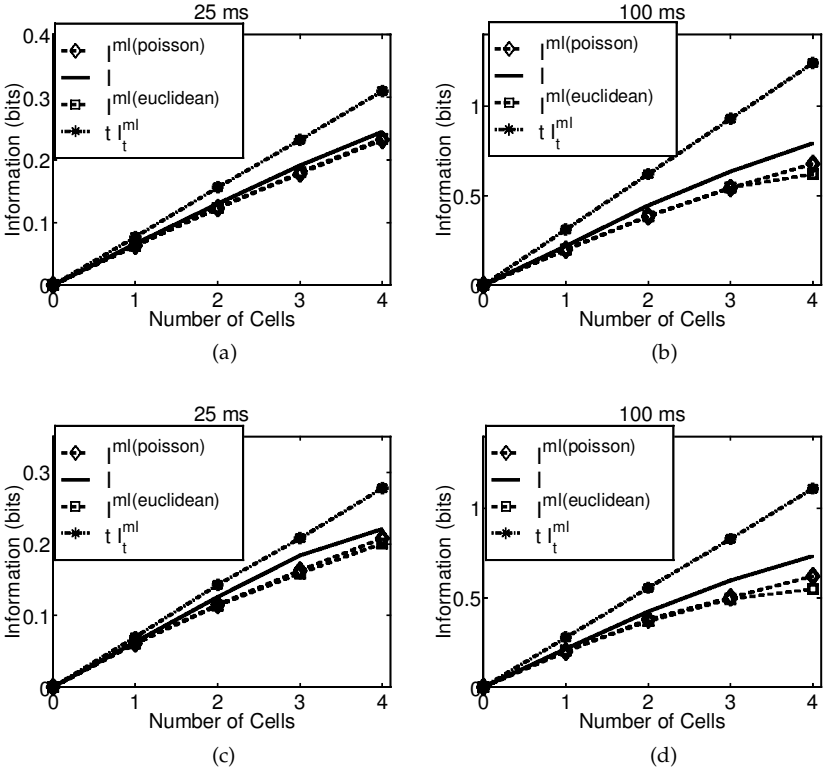

Figure 4: Decoding in short times is relatively insensitive to correlations. Responses were generated with the same mean firing rates to different stimuli as Figure 2; the firing, however, was now correlated across different cells, as follows. For any cell, the instantaneous probability of generating a spike at any 1 ms time interval $[t_0, t_0 + t]$ was still independent of the occurrence of other spikes emitted at different times, but was facilitated by the emission of a spike (by any other different cell) in the same very short time interval, as quantified by equation 2.5. Firing activity in longer time windows was generated by using equation 2.5 for many consecutive 1 ms intervals. (a–b) Scaled cross-correlation $\gamma = 2.0$. (c–d) $\gamma = 20.0$. The I^{ml} measures are not greatly affected by correlations, while the first-order approximation $t\, I^{ml}$ overestimates the information by a greater amount than for the pure Poisson data. This accords with intuition: the effect of the pairwise correlations is to induce a negative term at second order in t, which is being ignored in this approximation.

rates to each stimulus were chosen such that the stimuli predicted were nondegenerate and the mean rates were well separated so that any possible problem related to jackknife cross-validation was unimportant. After gen-

erating the responses, the stimuli were decoded with a Bayesian algorithm based on a Poisson model of responses and independence of responses of different cells, and with ED decoding for comparison. Then the maximum likelihood information I^{ml} and probability information I^{p} were calculated (a jackknife cross-validation was used), as were the first-order approximations I_t^{ml} and I_t to the maximum likelihood and the true information. The true information I, equation 2.1, was also computed, for comparison, from the underlying probabilities. Finite sampling corrections (Panzeri & Treves, 1996) were applied to all the quantities of interest. The figures show how the simulations confirmed analytical results and, moreover, indicated their range of validity. Although the first time derivatives describe precisely the true information only for short windows and smaller number of cells, we find that I^{ml} is in all the cases considered a much more precise quantification of the true neuronal information than I^{p}, as predicted by our analysis.

5 Application to Real Data

The responses of two pyramidal cells simultaneously recorded in the parahippocampal gyrus (PHG) and of three cells simultaneously recorded in the CA3 region of the hippocampus of a monkey (Rolls, Robertson, & Georges-François, 1997; Rolls et al., 1998) were analyzed with the same procedures described for the simulations, with the only obvious difference that the underlying probability distributions were now unknown. These cells were found by Rolls, Robertson, & Georges-François (1997) to be selective for "spatial views"; they responded mainly when the monkey looked at one part of the environment but not at another. The information about spatial views conveyed by these two small sets was calculated, after discretizing all possible views in 16 bins (see Rolls et al., 1998, for a full discussion of this procedure), for a time window 100 ms long. The number of trials (time windows) available per each stimulus was in the range 20 to 100. The full information I carried by the real neuronal responses was estimated directly, as in equation 2.1.[8] ED decoding outperformed Bayesian Poisson decoding for these cells. I^{p} yielded poorer estimates of I, exactly as with computer simulations. To check if trial-by-trial (i.e., noise) correlations between the simultaneously recorded responses carry information and affect the decoding, I and I^{ml} were also calculated after randomly shuffling, independently for each cell, the order of presentations of each stimulus. Those shuffled information measures are control quantities that represent the information

[8] For this purpose, two response bins per cell were used for the CA3 cells (after checking that at the single-cell level, this binarization of responses did not lead to significant information loss) and four response bins for the two PHG cells, which had higher firing rates (again after checking that the binning had no effect).

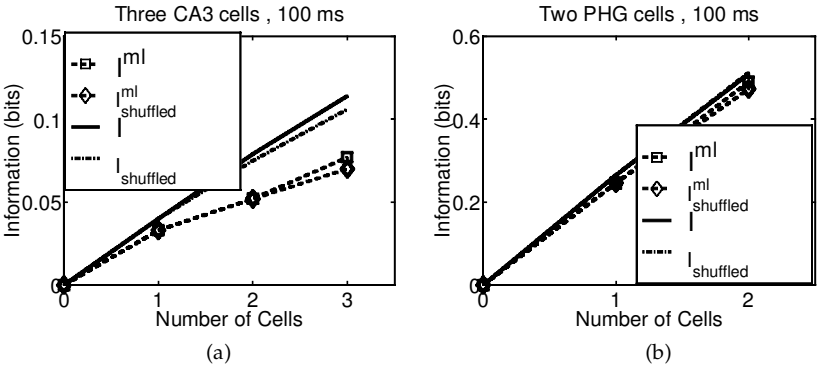

Figure 5: Estimates of I and I^{ml} and the control values obtained by shuffling responses, for real cells. (a) The result for three CA3 cells. In this case the mean response profiles of two of the cells were very similar (in particular, they had the same preferred stimulus and the same worst stimulus), and this leads, as expected, to less decoding efficiency as soon as the pair is included in the set. The small difference between shuffled and simultaneous information values shows instead that the correlation in response variability has little impact on the information transmitted by these cells. Thus, as predicted by the short time analysis, the loss of information in decoding is largely due to the similarity of the mean response profile of the cells, while the effects of trial-to-trial correlations (which appear only at the second order in t) are not evident from these 100 ms windows. (b) The result for a pair of PHG cells. In this case the two cells had different preferred and worst stimuli, and therefore the information loss in I^{ml} is small compared to the CA3 triplet. As before, trial-by-trial correlations do not appear to affect much either I or I^{ml}.

carried by cells with the same response profile as the original ones but fire independently. The results, shown in Figure 5, further confirm the analytical results described here and also show that the correlation in the response variability has little effect in a real sample of simultaneously recorded single cells.

6 Conclusions

- The decoded information I^{ml} can be an excellent approximation to I, the full information contained in the responses. The analysis valid in the $t \to 0$ limit indicates that this is the case whenever the response profiles of different cells adequately span, without much overlap, the range of stimuli used. Simulations and real data from very small ensembles of cells show that when response profiles match stimuli, I^{ml} continues to approximate I rather well even for intermediate time windows, of

the order of an interspike interval. The impossibility of measuring I directly from large ensembles prevents an explicit check of whether this result extends to more meaningful population sizes.

- On the other hand, I^p grows only quadratically with t, and similarly for a fixed short window it grows only quadratically with population size. Although I^p is less affected by limited sampling and easier to measure, any estimate of the information I contained in small populations of cells, and using very sparse data (or equivalently short windows), is expected to be strongly underestimated and therefore useless if based on I^p. It is not clear from the $t \to 0$ analysis what happens with larger populations, but data reported elsewhere indicate that I^{ml} and I^p tend to get closer in value as C becomes large.
- The relation between percentage correct and decoded (I^{ml}) information is nontrivial at first order in t and for very small ensembles, and therefore the metric content index of a representation can be estimated in this condition, with larger ensembles, of course, allowing better averaging.

Another finding, which is so far just empirical but deserves a better understanding, is that when information is estimated for a time window long enough that the responses are effectively graded and not binary, the decoding procedure often seems to be reasonably accurate. This is found in the simulations when considering longer windows: up to 200 ms, the information loss in the examples is, even with correlated firing, at most 10%. The accurate estimates of information through I^{ml} obtained with this longer windows do not simply result from using for the stimulus decoding the same (Poisson) probability distribution that elicited the simulated responses, because another simpler decoding procedure (ED decoding) gives very similar results and also because the decoding worked well with the simulation of correlated cells. This generally reliable estimation of information for longer times, even using simple decoding procedures or simple models for the firing-rate distributions, has also been suggested by analyses of real data. Examples include primate visual cortex data (Rolls, Treves, & Tovée, 1997; Gershon, Wiener, Latham, & Richmond, 1998); primate hippocampus and neighboring areas (Rolls et al., 1998); the precise stimulus reconstruction possible from the activity of rat hippocampal cells (Zhang et al., 1998); the relatively good performance of the neural network decoder of Hertz and coworkers on a set of lateral geniculate nucleus responses simulated by D. Golomb (Golomb, Hertz, Panzeri, Treves, & Richmond, 1997). This may be due to the fact that the information from single cells can be decoded, even with windows as long as 500 ms, with just a few levels of firing rates (Panzeri, Biella, Rolls, Skaggs, & Treves, 1996), and therefore even crude models of firing-rate distributions can be fed into a decoding procedure without significant information loss.

Thus the $t \to 0$ limit, on which the analytical results reported here are based, may not be a critical limitation. Moreover, there is substantial evidence that in many cases information is transmitted by neuronal activity in very short times, suggesting that it may also be decoded in short times. Therefore, the short time limit is interesting in itself. The fact that correlations are not included in first-order terms of the Taylor expansion does not seem a major limitation either; in any case, their effect on information transmission is evaluated in Panzeri et al. (1999), which analyzes second-order terms. The main limitation of our approach is likely to be instead in its applicability to short times *and* large ensembles. With large ensembles, any explicit check of decoding efficiency is not feasible, and although analytical results describe the conditions allowing efficient decoding, it remains unclear how to verify quantitatively the extent to which those conditions hold in real-life situations.

Appendix: Extension to Continuous Stimuli

The main results can be generalized to the case of a continuous distribution of stimuli $p(s)$ (we denote with $P(\cdot)$ a discrete distribution, and with $p(\cdot)$ a continuous probability density function (PDF)). In this case the conditional response probability $P(\mathbf{n}|s)$ is still discrete because neuronal responses are discrete anyway. The most likely stimulus s^p and the posited stimulus s' now belong to a continuous space, and $q(s^p|s)$ and $p(s'|s)$ are PDFs, although since the responses are discrete, only a discrete set of s^p can be predicted, and therefore $q(s^p|s)$ is in fact a sum of Dirac delta distributions, not a function. The expressions for $I(t)$,I^{ml},I^p are the same as for the case of discrete stimuli, the only difference being that the various sums over stimuli must be replaced by integrals.

Suppose that only one stimulus maximizes $p(s'|\mathbf{n})$ for each response $\mathbf{n}$ (in other words, predicted stimuli s^p are not degenerate). This is for simplicity, but also because it is unlikely and artificial to suppose that the response function of a neuron to a continuous stimulus has a large, flat maximum with exactly the same value of likelihood. The discrete sample of predicted stimuli can be studied as before, with the same notation as in section 3.1. The only difference is that now degeneracy need not be considered (and therefore there are $K + 1$ decoded stimuli, K being the number of predicted stimuli different from the worst stimulus and from one another). In order to avoid the entropy of the continuous stimulus set becoming infinite (i.e., the stimuli being measured, or predicted, with infinite precision), it is possible, for example, to regularize the distribution of s^p by convolving it with a gaussian of (small) standard deviation ϵ. ϵ thus corresponds to the finite resolution of the measurement of the stimulus parameters; the limit $\epsilon \to 0$ corresponds to the case when the distribution of s^p becomes a sum of delta

functions. The conditional distribution $q(s^p|s)$ becomes:

$$q(s^p|s) = \frac{1}{\sqrt{2\pi}\epsilon}\left[p(\mathbf{0}|s) + \sum_{c\in\mathcal{C}(0)} p(\mathbf{e}_c|s)\right]\exp -\frac{(s^p - sw)^2}{2\epsilon^2} + \frac{1}{\sqrt{2\pi}\epsilon}\sum_{c\in\mathcal{C}(k)} p(\mathbf{e}_c|s)\exp -\frac{(s^p - sb(c))^2}{2\epsilon^2}. \quad \text{(A.1)}$$

Taking the $t \to 0$ limit, and then the infinite stimulus resolution limit $\epsilon \to 0$ we find:

$$I_t^{ml} = \int ds p(s) \sum_{k=1}^{K}\left[\sum_{c\in\mathcal{C}(k)} \overline{r}_{s;c}\right]\log_2 \frac{\sum_{c\in\mathcal{C}(k)} \overline{r}_{s;c}}{\sum_{c\in\mathcal{C}(k)} \overline{r}_c}, \quad \text{(A.2)}$$

that is, essentially the same result as in the discrete case. The main difference is that with continuous stimuli, it is unlikely that two responses from a discrete set predict exactly the same value of s^p (which belongs instead to a continuous space), and therefore in general no information loss is expected to first order in t, apart from that arising from finite stimulus resolution ($\epsilon > 0$). The "probability information" I^p behaves exactly as with discrete stimuli case, in that I_t^p is again zero.

Brunel and Nadal (1998) have shown that in the limit of a large number of neurons coding for a low-dimensional, continuous stimulus, the mutual information between the population response and the stimulus becomes equal to the mutual information between the stimulus and an efficient gaussian prediction of the stimulus itself (efficient in this context means that the estimator has a variance equal to the Fisher information). These results, while interesting, are based on the assumption that the estimator s^p has a gaussian distribution around the correct value. While this is the case in the limits discussed in Brunel and Nadal (1998), in general the distribution of the estimator s^p is far from being gaussian around the true stimulus value s, and in the short time limit, it is, moreover, strongly biased toward the "worst" stimulus (see equation A.1). Another advantage of the analysis presented here is that since it does not require the use of a metric in the stimulus set, it can be applied in cases when the Fisher information cannot be calculated and therefore can be the complement of analyses based on Fisher information (Seung & Sompolinsky, 1993; Zhang et al., 1998) in the case of nonmetric stimuli.

Acknowledgments

We are grateful to F. Battaglia, W. Bialek, N. Brunel, M. Elliffe, N. Parga, and R. Petersen for interesting discussions. This research was supported by an EC Marie Curie Research Training grant ERBFMBICT972749 (S. P.), a studentship from the Oxford McDonnell-Pew Centre for Cognitive Neuroscience (S. S.), by MRC PG8513790, and by HCM.

References

Abbott, L. F. (1994). Decoding neuronal firing and modelling neural networks. *Quarterly Review of Biophysics, 27*, 291–331.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R., & Warland, D. (1991). Reading a neural code. *Science, 252*, 1854–1857.

Brunel, N., & Nadal, J. P. (1998). Mutual information, Fisher information and population coding. *Neural Comp, 10*, 1731–1757.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: SIAM.

Földiák, P. (1993). The "ideal homunculus": Statistical inference from neural population responses. In F. H. Eeckman and J. M. Bower (Eds.), *Computation and neural systems* (pp. 55–60). Norwell, MA: Kluwer.

Gawne, T. J., Kjaer, T. W., Hertz, J. A., & Richmond, B. J. (1996). Adjacent visual cortical complex cells share about 20% of their stimulus-related information. *Cerebral Cortex, 6*, 482–489.

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

Gershon, E. D., Wiener, M. C., Latham, P. E., & Richmond, B. J. (1998). Coding strategies in monkey V1 and inferior temporal cortices. *J. Neurophysiol., 79*, 1135–1144.

Gochin, P. M., Colombo, M., Dorfman, G. A., Gerstein, G. L., & Gross, C. G. (1994). Neural ensemble encoding in inferior temporal cortex. *J. Neurophysiol., 71*, 2325–2337.

Golomb, D., Hertz, J., Panzeri, S., Treves, A., & Richmond, B. (1997). How well can we estimate the information carried in neuronal responses from limited samples? *Neural Comp., 9*, 649–655.

Heller, J., Hertz, J. A., Kjaer, T. W., & Richmond, B. J. (1995). Information flow and temporal coding in primate pattern vision. *J. Comput. Neurosci., 2*, 175–193.

Miller, G. A. (1955). Note on the bias on information estimates. *Information Theory in Psychology: Problems and Methods, II-B*, 95–100.

Optican, L. M., Gawne, T. J., Richmond, B. J., & Joseph, P. J. (1991). Unbiased measures of transmitted information and channel capacity from multivariate neuronal data. *Biological Cybernetics, 65*, 305–310.

Oram, M. W., & Perrett, D. I. (1992). Time course of neuronal responses discriminating different views of face and head. *J. Neurophysiol., 68*, 70–84.

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

Panzeri, S., Schultz, Treves, A., & Rolls, E. T. (1999). Correlations and the encoding of information in the nervous system. *Proc. R. Soc. Lond. B* 266:1001–1012.

Panzeri, S., & Treves, A. (1996). Analytical estimates of limited sampling biases in different information measures. *Network, 7*, 87–107.

Pouget, A., Zhang, K., Deneve, S., & Latham, P. E. (1998). Statistically efficient estimation using population coding. *Neural Comp., 10*, 373–401.

Rieke, F., Warland, D., de Ruyter van Steveninck, R. R., & Bialek, W. (1996). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Rolls, E. T., Robertson, R. G., & Georges-François, P. (1997). Spatial views cells in the primate hippocampus. *European J. Neurosci., 9*, 1789–1794.

Rolls, E. T., & Treves, A. (1998). *Neural networks and brain function*. Oxford: Oxford University Press.

Rolls, E. T., Treves, A., Robertson, R. G., Georges-François, P., & Panzeri, S. (1998). Information about spatial views in an ensemble of primate hippocampal cells. *J. Neurophysiol., 79*, 1797–1813.

Rolls, E. T., Treves, A., & Tovée, M. J. (1997). The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Exp. Brain Res., 114*, 149–162.

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences of the USA, 90*, 10749–10753.

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

Skaggs, W. E., McNaughton, B. L., Gothard, K., & Markus, E. (1993). An information theoretic approach to deciphering the hippocampal code. In S. Hanson, J. Cowan, & C. Giles (Eds.), *Advances in neural information processing systems, 5* (pp. 1030–1037). San Mateo, CA: Morgan Kaufmann.

Snippe, H. P. (1996). Parameter extraction from population codes: A critical assessment. *Neural Comp., 8*, 511–529.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520–522.

Tovée, M. J., Rolls, E. T., Treves, A., & Bellis, R. J. (1993). Information encoding and the responses of single neurons in the primate temporal visual cortex. *J. Neurophysiol., 70*, 640–654.

Treves, A. (1997). On the perceptual structure of face space. *BioSystems, 40*, 189–196.

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

Treves, A., Panzeri, S., Robertson, R., Georges-François, P., & Rolls, E. (1996). The emergence of structure in neuronal representations. *Society for Neuroscience Abstracts, 22*, 281.

Victor, J. D., & Purpura, K. P. (1996). Nature and precision of temporal coding in visual cortex: A metric space analysis. *J. Neurophysiol., 76*, 1310–1326.

Wilson, M. A., & McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science, 261*, 1055–1058.

Zhang, K., Ginzburg, I., McNaughton, B., & Sejnowski, T. J. (1998). Interpreting neuronal population activity by reconstruction: A unified framework with application to hippocampal place cells. *J. Neurophysiol., 79*, 1017–1044.

Received April 13, 1998; accepted October 15, 1998.