Data-Robust Tight Lower Bounds to the Information Carried by Spike Times of a Neuronal Population

G. Pola
g_pola@yahoo.co.uk
Department of Pure and Applied Mathematics, University of L'Aquila, I-67010 L'Aquila, Italy

R. S. Petersen
r.petersen@manchester.ac.uk
University of Manchester, Faculty of Life Sciences, Manchester M60 1QD, U.K.

A. Thiele
alex.thiele@ncl.ac.uk
M. P. Young
m.p.young@ncl.ac.uk
Psychology, Brain, and Behaviour, University of Newcastle upon Tyne Newcastle upon Tyne, NE2 4HH, U.K.

S. Panzeri
s.panzeri@manchester.ac.uk
The University of Manchester, Faculty of Life Sciences, Moffat Building, PO Box 88, Manchester M60 1QD, U.K.

We develop new data-robust lower-bound methods to quantify the information carried by the timing of spikes emitted by neuronal populations. These methods have better sampling properties and are tighter than previous bounds based on neglecting correlation in the noise entropy. Our new lower bounds are precise also in the presence of strongly correlated firing. They are not precise only if correlations are strongly stimulus modulated over a long time range. Under conditions typical of many neurophysiological experiments, these techniques permit precise information estimates to be made even with data samples that are three orders of magnitude smaller than the size of the response space.

1 Introduction

A fundamental problem in systems neuroscience is to understand the nature of the code used by neuronal populations to transmit sensory information. A traditional hypothesis is that information is carried by the total number of spikes emitted by individual neurons over a relatively long time window.

Total spike counts typically vary across a stimulus set, such that spike counts afford some degree of discriminability concerning which stimulus has occurred (Adrian, 1926; Tovee, Rolls, Treves, & Bellis, 1993; Shadlen & Newsome, 1998). However, recent studies have revealed that information may also be carried by precise spike timing. The spike timing code may take the simple form of precise timing of individual spikes (Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1996; Furukawa, Xu, & Middlebrooks, 2000; Panzeri, Petersen, Schultz, Lebedev, & Diamond, 2001; DeWeese, Wehr, & Zador, 2003) or the more complex form of patterns of spikes whose emission times are statistically correlated (Abeles, Bergman, Margalit, & Vaadia, 1993; Vaadia et al., 1995; Dan, Alonso, Usrey, & Reid, 1998).

An increasingly popular way to characterize quantitatively the relative role of spike timing and spike counts is to quantify and compare the amounts of information carried by different neuronal codes (Rieke et al., 1996; Borst & Theunissen, 1999; Dimitrov and Miller, 2001; Panzeri, Petersen, et al., 2001). However, a problem with this approach is that quantifying reliably the information conveyed by spike timing often requires the collection of unpractically large samples of data. This is mainly because spike times are not statistically independent: they are correlated (Mastronarde, 1983; Gawne & Richmond, 1993; de Oliveira, Thiele, & Hoffman, 1997; Averbeck & Lee, 2004). If such correlations did not exist, then the statistics of spike times would be completely characterized by the time-dependent firing rate of each neuron. However, one needs to measure also the correlations among all possible groups of spikes. A complete characterization of these correlations requires a number of parameters that are difficult to sample with a realistic amount of neuronal data. Thus, spike timing information measures suffer from an upward sampling bias problem (Panzeri & Treves, 1996).

One useful approach to the sampling bias problem has been to seek data-robust lower bounds to the spike timing information that neglect part of the spike timing correlations (Reich, Mechler, Purpura, & Victor, 2000; Reich, Mechler, & Victor, 2001). These information lower bounds have been used to provide convincing characterizations of the role of spike timing in cortical coding (Reich et al., 2000; Reich et al., 2001; Panzeri, Petersen, et al., 2001). However, these lower bounds are not tight if neurons are strongly correlated. To overcome this limitation, in this letter we develop new data-robust lower bounds to spike timing information that have better sampling properties and are tighter than previous estimators. Our new lower bounds are very tight also in the presence of strongly correlated firing. They fail only if correlations are strongly stimulus modulated over a long time range. They permit precise and reliable estimates of the information conveyed by spike times even when dealing with data samples that are relatively small with respect to the size of the response space. Under appropriate circumstances, the new techniques allow investigation of the spike timing information in

long time windows with samples made of tens to hundreds of trials in cases where direct estimation of mutual information would require thousands to millions of trials.

The letter is organized as follows. We first review basic concepts of information theory applications to spike trains; we then critically evaluate previous lower-bound techniques. Next we introduce our new lower bounds, testing and illustrating them by means of applications to both simulated data and real neuronal spike trains.

2 The Information Carried by Neuronal Population Responses

We consider a time period of duration T, associated with a dynamic or static sensory stimulus s (chosen with probability $P(s)$ from a stimulus set $\mathcal{S}$ with S elements), during which the activity of C cells is observed. We assume that the spike arrival times are binned with a timing precision Δt and transformed into a sequence of spike counts in each time bin. L denotes the number of time bins (i.e., $T = L\Delta t$). The neuronal population response is denoted by a one-dimensional array $\mathbf{r} = \{\mathbf{r}(1), \mathbf{r}(2), \mathbf{r}(3), \ldots, \mathbf{r}(L)\}$, where $\mathbf{r}(t) = \{r_1(t), r_2(t), r_3(t), \ldots, r_C(t)\}$ is the population response in the tth time bin; $r_c(t)$ is the number of spikes emitted by the cth neuron in the tth time bin. The maximum number of spikes that can be observed in a single time bin in any trial is denoted by M. (If Δt is very short, M is 1 and $r_c(t)$ is binary.) We indicate the response space with $\mathcal{R}$. ($\mathcal{R}$ contains $(M+1)^{LC}$ elements.)

Following Shannon (1948), we write the mutual information transmitted by the population response about the whole set of stimuli as

$$I(\mathcal{R};\mathcal{S}) = H(\mathcal{R}) - H(\mathcal{R}|\mathcal{S}), \tag{2.1}$$

where $H(\mathcal{R})$ and $H(\mathcal{R}|\mathcal{S})$ are the response entropy (stimulus unconditional) and the noise entropy (stimulus conditional), respectively, of the response variables. They are defined (Cover & Thomas, 1991) as

$$H(\mathcal{R}) = -\sum_{\mathbf{r}\in\mathcal{R}} P(\mathbf{r}) \log_2 P(\mathbf{r}), \tag{2.2}$$

$$H(\mathcal{R}|\mathcal{S}) = -\sum_{s\in\mathcal{S}} P(s) \sum_{\mathbf{r}\in\mathcal{R}} P(\mathbf{r}|s) \log_2 P(\mathbf{r}|s). \tag{2.3}$$

In equations 2.2 and 2.3, the summation over $\mathbf{r}$ stands for the sum over all possible population responses. The summation over s indicates a sum over all stimuli s. $P(\mathbf{r}|s)$ is the probability of simultaneously observing a particular response $\mathbf{r}$ conditional to stimulus s, and $P(\mathbf{r}) = \langle P(\mathbf{r}|s)\rangle_s$ is its average across all stimulus presentations (the angular brackets indicate the average over different stimuli, $\langle F(s)\rangle_s \equiv \sum_{s\in\mathcal{S}} P(s)F(s)$). In practice, $P(\mathbf{r}|s)$

is determined experimentally by repeating each stimulus in exactly the same way on many trials, while recording the neuronal responses. The probability $P(s)$ is usually chosen by the experimenter.

Estimating the information carried by spike times of real neuronal populations is difficult because each stimulus-response probability has to be measured from limited amounts of data. The statistical errors in estimating the response probabilities lead to downward systematic errors (biases) in both noise and response entropy (Miller, 1955) and in an overall upward bias when estimating mutual information (Panzeri & Treves, 1996). This makes it difficult to estimate the information directly from equation 2.1, especially for long time windows or precise spike time discretizations (large L) and large neuronal populations (large C).

3 Independent and Correlated Stimulus-Response Probabilities

The full description of the stimulus-response relationship is given by $P(\mathbf{r}|s)$. Estimating this probability, which has $(M+1)^{LC}-1$ free parameters for each stimulus s, requires extensive data samples. However, if spike times were statistically independent events, the stimulus-response probability would simply be characterized by the probability of a spike in each individual time bin. One way to alleviate the sampling problem is thus to work with probability models that assume that spikes emitted in response to a certain stimulus are independent of each other. Since by definition, stochastic variables are statistically independent if their joint response probabilities equal the product of the individual probabilities, we define the independent probability model $P_{ind}(\mathbf{r}|s)$ as the product of $P(r_c(t)|s)$, the stimulus-conditional marginal probabilities of responses of individual cells in each time bin:

$$P_{ind}(\mathbf{r}|s) = \prod_{c=1}^{C}\prod_{t=1}^{L} P(r_c(t)|s). \tag{3.1}$$

While estimating $P(\mathbf{r}|s)$ requires an evaluation of $(M+1)^{LC}-1$ parameters for each stimulus s, estimating $P_{ind}(\mathbf{r}|s)$ needs only MLC parameters for each stimulus.

In this letter, when we say that the spike trains are correlated, we mean that for some stimulus s, the real stimulus-response probability $P(\mathbf{r}|s)$ is different from $P_{ind}(\mathbf{r}|s)$. Thus, when we refer to correlations, we refer to correlations at fixed stimulus. These correlations are usually called *noise correlations* (Gawne & Richmond, 1993; Nirenberg & Latham, 2003; Pola, Thiele, Hoffmann, & Panzeri, 2003), and are the main subject of this letter. For simplicity, in the rest of this letter, when we use the term *correlation*, we intend "noise correlation."

One way to parameterize noise correlations is (Pola et al., 2003) to introduce a generalized correlation coefficient $\gamma(\mathbf{r}|s)$ quantifying how much the

real response probability $P(\mathbf{r}|s)$ deviates from $P_{ind}(\mathbf{r}|s)$:

$$\gamma(\mathbf{r}|s) = \frac{P(\mathbf{r}|s)}{P_{ind}(\mathbf{r}|s)} - 1, \quad \text{if } P_{ind}(\mathbf{r}|s) \neq 0,$$
$$\gamma(\mathbf{r}|s) = 0, \quad \text{if } P_{ind}(\mathbf{r}|s) = 0. \tag{3.2}$$

This generalized correlation coefficient varies in the range $-1 \leq \gamma(\mathbf{r}|s) < \infty$. Negative values indicate anticorrelation; positive values indicate correlation (Pola et al., 2003).

4 A Lower Bound That Neglects Correlations in the Noise Entropy

Let us now consider in detail the effect of ignoring correlations in the stimulus-conditional response probability on both noise entropy and response entropy.

Neglecting correlations by using $P_{ind}(\mathbf{r}|s)$ instead of the true distribution $P(\mathbf{r}|s)$ necessarily increases the noise entropy:

$$H_{ind}(\mathcal{R}|\mathcal{S}) \geq H(\mathcal{R}|\mathcal{S}), \tag{4.1}$$

where $H_{ind}(\mathcal{R}|\mathcal{S})$ is defined as

$$H_{ind}(\mathcal{R}|\mathcal{S}) = -\sum_{s\in\mathcal{S}} P(s) \sum_{\mathbf{r}\in\mathcal{R}} P_{ind}(\mathbf{r}|s) \log_2 P_{ind}(\mathbf{r}|s). \tag{4.2}$$

The inequality in equation 4.1 can be proved rewriting the difference between $H_{ind}(\mathcal{R}|\mathcal{S})$ and $H(\mathcal{R}|\mathcal{S})$ as

$$H_{ind}(\mathcal{R}|\mathcal{S}) - H(\mathcal{R}|\mathcal{S}) = D(P(\mathbf{r}|s)||P_{ind}(\mathbf{r}|s)), \tag{4.3}$$

where $D(P(\mathbf{r}|s)||P_{ind}(\mathbf{r}|s))$ is the conditional Kullback-Leibler (KL) distance between $P(\mathbf{r}|s)$ and $P_{ind}(\mathbf{r}|s)$. The conditional KL distance between two distributions $P(\mathbf{r}|s)$ and $Q(\mathbf{r}|s)$ is defined as (see Cover & Thomas, 1991)

$$D(P(\mathbf{r}|s)||Q(\mathbf{r}|s)) = \sum_{s\in\mathcal{S}} P(s) \sum_{\mathbf{r}\in\mathcal{R}} P(\mathbf{r}|s) \log_2 \frac{P(\mathbf{r}|s)}{Q(\mathbf{r}|s)}. \tag{4.4}$$

$D(P(\mathbf{r}|s)||Q(\mathbf{r}|s))$ is nonnegative, and it is zero if and only if $P(\mathbf{r}|s) = Q(\mathbf{r}|s)$ for every $\mathbf{r}$ and s. Thus, $H_{ind}(\mathcal{R}|\mathcal{S}) = H(\mathcal{R}|\mathcal{S})$ if and only if $P(\mathbf{r}|s) = P_{ind}(\mathbf{r}|s)$ for every $\mathbf{r}$ and s; otherwise, $H_{ind}(\mathcal{R}|\mathcal{S}) > H(\mathcal{R}|\mathcal{S})$.

$H_{ind}(\mathcal{R}|\mathcal{S})$ is much easier to sample than $H(\mathcal{R}|\mathcal{S})$ because it can be expressed as a sum of entropies of the marginal distributions in individual time bins,

$$H_{ind}(\mathcal{R}|\mathcal{S}) = \sum_{c=1}^{C} \sum_{t=1}^{L} H(\mathcal{R}_{c,t}|\mathcal{S}), \tag{4.5}$$

where

$$H(\mathcal{R}_{c,t}|\mathcal{S}) = -\sum_{s\in\mathcal{S}} P(s) \sum_{r_c(t)=0}^{M} P(r_c(t)|s) \log_2 P(r_c(t)|s). \tag{4.6}$$

In contrast to the noise entropy, the response entropy $H(\mathcal{R})$ can be either reduced or increased by using the independent probability model (Schultz & Panzeri, 2001). Hence, replacing $P(\mathbf{r}|s)$ with $P_{ind}(\mathbf{r}|s)$ in the definition of mutual information, equation 2.1, does not provide a lower bound on the mutual information.

The idea of Reich and collaborators (Reich et al., 2000, 2001) was to produce a sampling-robust lower bound by neglecting correlations only in the noise entropy, as follows,[1]

$$I_{LB1} = H(\mathcal{R}) - H_{ind}(\mathcal{R}|\mathcal{S}), \tag{4.7}$$

where $H(\mathcal{R})$ and $H_{ind}(\mathcal{R}|\mathcal{S})$ are defined in equations 2.2 and 4.2, respectively. The lower bound introduced in this way is much more data robust than the mutual information because only $P_{ind}(\mathbf{r}|s)$ and $P(\mathbf{r})$, but not $P(\mathbf{r}|s)$, need to be estimated. Since $H_{ind}(\mathcal{R}|\mathcal{S})$ is a sum of low-dimensional entropies (see equation 4.5), the most biased term in equation 4.7 is $H(\mathcal{R})$. Numerical investigations of these sampling properties will be presented below.

It is useful to rewrite the lower-bound I_{LB1} in the following equivalent way:

$$I_{LB1} = I(\mathcal{R};\mathcal{S}) - D(P(\mathbf{r}|s)||P_{ind}(\mathbf{r}|s)). \tag{4.8}$$

Since the conditional KL distance is always nonnegative, and it is zero if and only if $P(\mathbf{r}|s) = P_{ind}(\mathbf{r}|s)$ for every $\mathbf{r}$ and s, the very presence of correlations causes I_{LB1} to become less than the mutual information, even if these correlations are not carrying any information about the stimuli. Indeed, I_{LB1} can sometimes become negative in the presence of strong enough correlations.

[1] Like the mutual information $I(\mathcal{R};\mathcal{S})$, I_{LB1} is also defined over the stimulus and response space and should thus have $(\mathcal{R};\mathcal{S})$ as argument; however, we drop this argument for simplicity of notation.

It is also useful to point out that I_{LB1} is always less than the sum of single time bin information,

$$I_{LB1} \leq \sum_{c=1}^{C} \sum_{t=1}^{L} I(\mathcal{R}_{c,t}; \mathcal{S}), \tag{4.9}$$

where $I(\mathcal{R}_{c,t}; \mathcal{S})$ is the information conveyed by the spikes emitted by the cth cell in the tth time bins and is defined as

$$I(\mathcal{R}_{c,t}; \mathcal{S}) = H(\mathcal{R}_{c,t}) - H(\mathcal{R}_{c,t}|\mathcal{S}), \tag{4.10}$$

where $H(\mathcal{R}_{c,t}|\mathcal{S})$ is defined in equation 4.6, and

$$H(\mathcal{R}_{c,t}) = -\sum_{r_c(t)=0}^{M} P(r_c(t)) \log_2 P(r_c(t)). \tag{4.11}$$

Thus, I_{LB1} cannot reveal the presence of synergistic encoding between spikes emitted at different times or by different neurons.

5 A Tighter Lower Bound That Ignores Stimulus-Modulated Correlations

I_{LB1}, though much more data robust than the mutual information, is not tight in the presence of correlations between spike times. The main purpose of this article is to introduce new data-robust lower bounds that can be tight even in the presence of strong correlations. This new lower bound is based on a recently developed information breakdown formalism (Pola et al., 2003) that separates out the contribution of different coding mechanisms. These components of the mutual information breakdown have different magnitudes and sampling properties. In this section, we show how this information breakdown provides the basis for better data-robust lower bounds.

5.1 The Information Breakdown. The information breakdown method was originally introduced to investigate the role of correlated cortical neuronal firing in transmitting sensory information. Here, we briefly review it and slightly extend it to quantify the information content of all correlation between spike times (and not only of cross-cell correlations as in Pola et al., 2003).

This formalism consists in breaking down the total mutual information into a sum of components, each of which can be associated with a different

coding mechanism (Pola et al., 2003; Golledge et al., 2003):

$$I(\mathcal{R};\mathcal{S}) = I_{lin} + I_{sig-sim} + I_{cor}. \tag{5.1}$$

The mathematical expression and its interpretation in terms of coding mechanisms, and the sampling properties of each component will be discussed below. We will present the coding components in a mathematical form similar to that reported in appendix A of Pola et al. (2003) and in Golledge et al. (2003).

5.1.1 The Linear Component I_{lin}. The first term of the information breakdown is the information that would be obtained if the spikes emitted in different time bins and cells all conveyed independent information. In this case, the total information transmitted by the population would just be a linear sum of the information provided by each time bin and each cell:

$$I_{lin} = \sum_{c=1}^{C}\sum_{t=1}^{L}[H(\mathcal{R}_{c,t}) - H(\mathcal{R}_{c,t}|\mathcal{S})]. \tag{5.2}$$

This component has very good sampling properties because it only requires sampling the entropies in single bins separately.

Deviations from independent information transmission (i.e., synergy or redundancy) are expressed by the terms considered next.

5.1.2 The Signal-Similarity Term $I_{sig-sim}$. This term quantifies the redundancy (or information loss) arising from signal similarity (Gawne & Richmond, 1993)—similarity across stimuli of the mean probability of spike emission in each time bin.[2] Its expression is:

$$I_{sig-sim} = H_{ind}(\mathcal{R}) - \sum_{c=1}^{C}\sum_{t=1}^{L} H(\mathcal{R}_{c,t}), \tag{5.3}$$

where $H_{ind}(\mathcal{R})$ is the stimulus-unconditional independent entropy:

$$H_{ind}(\mathcal{R}) = -\sum_{\mathbf{r}} P_{ind}(\mathbf{r}) \log_2 P_{ind}(\mathbf{r}) \tag{5.4}$$

and $P_{ind}(\mathbf{r}) = \sum_s P(s) P_{ind}(\mathbf{r}|s)$. $I_{sig-sim}$, which is always less than or equal to zero, was first introduced by Pola et al. (2003) as a generalization of

[2]Signal similarity is more often called signal correlation, (see Gawne & Richmond, 1993; Nirenberg and Latham, 2003). Here we use the word *similarity* because we use *correlation* to refer only to noise correlation.

the series expansion of Panzeri and Schultz (2001), and later discussed by Schneidman, Bialek, and Berry (2003) (with an overall change in sign) under the name "ΔI_{signal}." $I_{sig-sim}$ does not depend on spike time correlations, but only on the marginal distributions. For this reason, it has extremely good sampling properties.

The sum of I_{lin} and $I_{sig-sim}$ equals I_{ind}, the information that would be obtained if there were no correlations (i.e., if $P(\mathbf{r}|s) = P_{ind}(\mathbf{r}|s)$ for every $\mathbf{r}$ and s):

$$I_{ind} = I_{lin} + I_{sig-sim}. \tag{5.5}$$

5.1.3 The Total Impact of Correlation I_{cor}. The next term in the information breakdown, indicated as I_{cor}, is the only term that depends on the correlation coefficient $\gamma(\mathbf{r}|s)$ and quantifies the total impact of correlation in information encoding. It is defined as the difference between the information $I(\mathcal{R}; \mathcal{S})$ in the presence of correlations and the information I_{ind} in absence of correlation:

$$I_{cor} = I(\mathcal{R}; \mathcal{S}) - I_{ind}. \tag{5.6}$$

This quantity was introduced in Hatsopoulos, Ojakangas, Paninski, and Donoghue (1998) and Nirenberg and Latham (1998) and later refined and used in Panzeri, Golledge, Zheng, Tovee, and Young (2001), and Pola et al. (2003) to study the role of correlations in sensory and motor coding. I_{cor} can be further broken down into two components, which we will term "correlational," that reflect two different ways in which correlations may contribute to coding (Pola et al., 2003; Golledge et al., 2003):

$$I_{cor} = I_{cor-ind} + I_{cor-dep}. \tag{5.7}$$

The two correlational components $I_{cor-ind}$ and $I_{cor-dep}$ will be briefly introduced next. Further considerations on the meaning and scope of I_{cor}, $I_{cor-ind}$ and $I_{cor-dep}$ can be found in Panzeri, Pola, Petroni, Young, & Petersen, (2002), Nirenberg and Latham (2003), Pola et al. (2003), and Schneidman et al.(2003). These considerations will not be reviewed here because they have no direct bearing on the work presented in this article, as the information breakdown is used merely as a tool to obtain data-robust lower bounds rather than separating out different correlational coding mechanisms.

5.1.4 The Stimulus-Independent Correlational Component $I_{cor-ind}$. Even if not stimulus modulated, correlations can still affect the neuronal code (Abbott & Dayan, 1999; Oram, Földiák, Perrett, & Sengpiel, 1998). The

component of information associated with stimulus-independent correlations is

$$I_{cor-ind} = \chi(\mathcal{R}) - H_{ind}(\mathcal{R}), \tag{5.8}$$

where $\chi(\mathcal{R})$ is defined as

$$\chi(\mathcal{R}) = -\sum_{\mathbf{r}} P(\mathbf{r}) \log_2 P_{ind}(\mathbf{r}). \tag{5.9}$$

$I_{cor-ind}$ is positive (synergistic) when spike timing correlation and signal similarity have opposite signs $I_{cor-ind}$ is instead negative (redundant) when they have the same sign (see Oram et al., 1998; Pola et al., 2003).

Neurophysiological studies revealed that $I_{cor-ind}$ can have a substantial impact on cortical information encoding. It is positive for both within-cell correlations in rat S1 cortex (+18% of the total information; Panzeri, Petersen, et al., 2001) and cross-cell correlations in both monkey S2 cortex (Romo, Hernandez, Zainos, & Salinas, 2003) and rat prefrontal cortex (Jung, Qin, Lee, & Mook-Jung, 2000). It is negative (−20%) for both cross-cell correlations between nearby neurons in rat S1 cortex (Petersen, Panzeri, & Diamond, 2001) and monkey IT cortex (−12% ; Rolls, Franco, Aggelopoulos, & Reece, 2003).

$I_{cor-ind}$ has good sampling properties because it depends on only $P(\mathbf{r})$ and $P_{ind}(\mathbf{r})$, not on $P(\mathbf{r}|s)$ and $P_{ind}(\mathbf{r}|s)$.

5.1.5 The Stimulus-Dependent Correlational Component $I_{cor-dep}$. The final term of the information breakdown, $I_{cor-dep}$, is associated with stimulus modulation of correlations:

$$I_{cor-dep} = D(P(s|\mathbf{r})||P_{ind}(s|\mathbf{r})) \equiv \sum_{\mathbf{r}} P(\mathbf{r}) \sum_{s} P(s|\mathbf{r}) \log_2 \frac{P(s|\mathbf{r})}{P_{ind}(s|\mathbf{r})}. \tag{5.10}$$

This term is always positive or zero. It is associated with stimulus-dependent correlations because it equals zero if and only if the correlation coefficient $\gamma(\mathbf{r}|s)$ does not depend on s for every response $\mathbf{r}$. If a neuronal population carries information by emitting patterns of correlated spikes that "tag" each stimulus, $I_{cor-dep}$ is greater than zero.

Equation 5.10 was introduced by Nirenberg, Carcieri, Jacobs & Latham (2001) (and named ΔI); Pola et al. (2003) then showed that this is a generalization of an analogous quantity introduced by Panzeri and Schultz (2001) in the short-time approximation. An alternative interpretation of $I_{cor-dep}$ stems from the fact that $I_{cor-dep}$ is zero if and only if $P(s|\mathbf{r}) = P_{ind}(s|\mathbf{r})$ for every s and $\mathbf{r}$ (Cover & Thomas, 1991; Nirenberg et al., 2001). Thus, $I_{cor-dep}$ is zero if and only if the decoding dictionary obtained using $P_{ind}(\cdot)$ is the same one obtained using the true distribution $P(\cdot)$. Nirenberg & Latham (2003)

interpreted it as a cost function measuring "how much harder it is to decode neural responses when correlations are ignored than when they are taken into account."

Although in general, a K-L distance can be infinite in some cases (Cover & Thomas, 1991), it is important to note that $I_{cor-dep}$ is always finite (Pola et al., 2003). This is because if $P_{ind}(s|\mathbf{r})$ is zero, then so is $P(s|\mathbf{r})$, and in this case the quantity $P(s|\mathbf{r}) \log \frac{P(s|\mathbf{r})}{P_{ind}(s|\mathbf{r})}$ would be zero.[3]

A number of studies have made estimates of $I_{cor-dep}$ from neurophysiological data. With the one exception of Dan et al. (1998),[4] all of these studies have reported small values of $I_{cor-dep}$ as follows. Panzeri, Petersen, et al. (2001) and Petersen et al. (2001) found that it contributes 3% to coding of whisker position in rat S1 cortex. Nirenberg et al. (2001) found it to be negligible for the vast majority of nearby cells in mouse retina ($I_{cor-dep}/I$ was more than 10% for only one pair out of over four hundred). Rolls et al. (2003) found it to be less than 2% of the total information about faces carried by the firing of monkey IT neurons. Golledge et al. (2003) found that $I_{cor-dep}$ was $\approx$5% of the total information about visual objects carried by the firing of cat V1 neurons. Current data thus suggest that $I_{cor-dep}$ typically contributes little to the total information available in the neuronal response.

$I_{cor-dep}$ is by far the most biased of all terms entering the information breakdown in equation 5.1. It is as biased as the total mutual information itself. This is because its evaluation requires measuring the full correlational structure for each stimulus.

5.2 The New Tighter Lower Bound. Since the stimulus-dependent correlational component $I_{cor-dep}$ is nonnegative, is the only component that presents significant sampling problems, and has been found in most cases to account for a small proportion of the total information, a data-robust and tight lower bound can be obtained by computing the information ignoring $I_{cor-dep}$:

$$I_{LB2} = I_{lin} + I_{sig-sim} + I_{cor-ind}. \tag{5.11}$$

Using equations 5.2, 5.3, and 5.8, I_{LB2} can be written as

$$I_{LB2} = \chi(\mathcal{R}) - H_{ind}(\mathcal{R}|\mathcal{S}). \tag{5.12}$$

[3] In fact, $P_{ind}(s|\mathbf{r}) = 0$ implies that either P(s)=0 (which implies that $P(s|\mathbf{r}) = 0$) or that $P_{ind}(\mathbf{r}|s) = 0$. In the latter case, at least one of the marginals of $P(\mathbf{r}|s)$ entering the product in equation 3.1 is zero, which implies that $P(\mathbf{r}|s) = 0$ and $P(s|\mathbf{r}) = 0$.

[4] It should be however noted that the experiment of Dan et al. (1998) quantified information through a stimulus reconstruction method rather by means of a more direct approach, and this makes the comparison with the work presented here difficult.

As we shall see in section 6, I_{LB2} has very good sampling properties compared to both $I(\mathcal{R};\mathcal{S})$ and I_{LB1}.

$I_{cor-dep}$ is zero if and only if $\gamma(\mathbf{r}|s)$ is not stimulus modulated for every response $\mathbf{r}$, whatever the overall strength of $\gamma(\mathbf{r}|s)$. Hence, even if the spike trains are strongly correlated but these correlations are stimulus independent, the lower bound I_{LB2} is still tight. Thus, under many circumstances, I_{LB2} is a significantly tighter lower bound than I_{LB1}, which is tight only in the total absence of correlations.

As for I_{LB1}, there are hypothetical cases where stimulus modulation of correlations is particularly strong and individual spikes code for little information; here, I_{LB2} can become negative and thus not useful. However, no such situation has yet been reported in information analysis of experimentally recorded spike trains.

Is I_{LB2} always tighter than I_{LB1}? The following inequality can be proved:

$$I_{LB2} - I_{LB1} = \sum_{\mathbf{r}} P(\mathbf{r}) \log_2 \frac{P(\mathbf{r})}{P_{ind}(\mathbf{r})} = D(P(\mathbf{r})||P_{ind}(\mathbf{r})) \geq 0, \tag{5.13}$$

where in the above equation $D(P(\mathbf{r})||P_{ind}(\mathbf{r}))$ is a KL distance (see Cover & Thomas, 1991, p. 18, equation 2.26). Thus, I_{LB2} is always tighter than I_{LB1}:

$$I_{LB1} \leq I_{LB2} \leq I(\mathcal{R};\mathcal{S}), \tag{5.14}$$

with $I_{LB1} = I_{LB2}$ if and only if $P(\mathbf{r})$ equals $P_{ind}(\mathbf{r})$ for each $\mathbf{r}$.

In order to clarify and illustrate the differences between the two lower-bound estimators, we applied the method to synthetic spike trains. We simulated a neuronal pair responding to two stimuli, reflecting different ways of encoding information through correlations. The analysis in this section was performed using a large number of trials; the behavior of the two estimators with small data samples will be discussed in the next section.

We considered three situations: uncorrelated spike trains, correlated spike trains with weak stimulus modulation of correlation, and correlated spike trains with strong stimulus modulation of correlation.

We started with the uncorrelated case. We generated, independently for each cell, simulated data according to a stationary Poisson process with the mean rates reported in Figure 1A (left panel). For illustration, we show in Figure 1A (central panel) that the cross-correlogram (CCG) between the two spike trains was flat. Consistent with the above mathematical analysis, both lower-bound estimators were exactly equal to the true mutual information (see Figure 1A, right panel). Thus, in the absence of correlations, both lower bounds provide equally precise estimates.

We then modeled a neuronal pair with the same mean firing rates as above, but with the addition of correlated activity (see Figure 1B). The

amount of correlation was strong, but only very weakly stimulus modulated (see Figure 1B, central panel). In this case, I_{LB2} was extremely accurate, but I_{LB1} underestimated the true information by 39% (see Figure 1B, right panel).

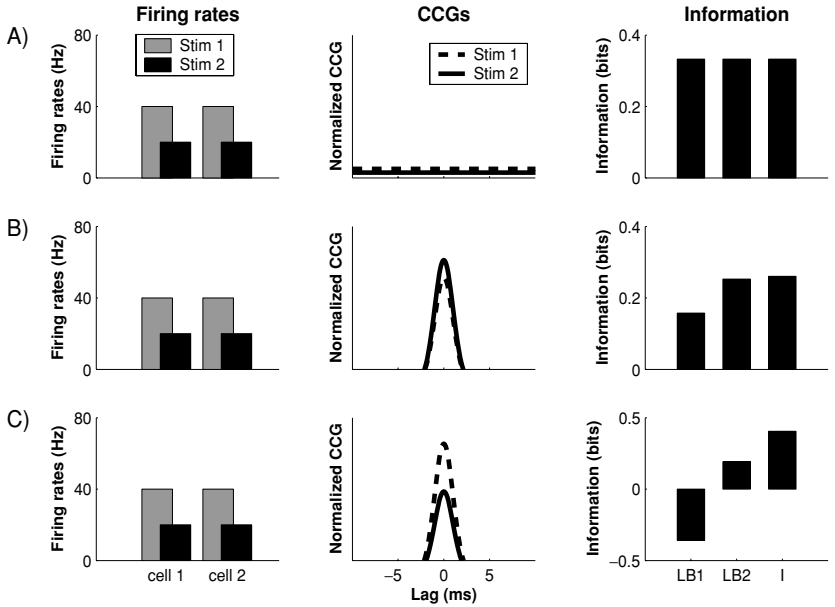
Finally, we simulated a case of strong stimulus modulation of the correlation (see Figure 1C). In this case, I_{LB2} was no longer tight; however, it still performed much better than I_{LB1}, which in this case was strongly negative.

Thus, unlike I_{LB1}, I_{LB2} can provide accurate and robust measures of information even in the presence of correlated activity, so long as correlations are not stimulus modulated. In section 7, we will show how to improve on this.

6 Sampling Bias Properties

In practice, the spike timing information and its lower-bound approximations must be estimated from experimental probabilities obtained from a limited number N of repeated presentations of all stimuli. This leads to a systematic error (or bias) in the estimate of the mutual information and of its lower bounds, the size of the bias decreasing when increasing the number of trials Panzeri & Treves, 1996). In this section, we focus more explicitly on the sampling bias properties of the two lower-bound estimators.

Figure 1: The performance of the lower-bound estimators LB1 and LB2 in the presence and absence of correlated activity. The lower-bounds estimates are compared to the total mutual information $I(\mathcal{R};\mathcal{S})$ in simulated neuronal pairs responding to two stimuli. The simulated spike trains were analyzed in the 0–60 ms poststimulus window, using a timing resolution Δt of 10 ms to bin the responses (thus, there were six time bins per cell). We considered three ways of encoding information through correlations: (1) absence of correlation, (2) stimulus-independent correlation, and (3) stimulus-dependent correlation. Data were generated as described next. We first created, independently for each cell, spikes from a Poisson process with a certain firing rate r, as follows. For each time bin, we generated at random a spike with a generation probability equal to $r\,\Delta t$, the generated spike being given a random time within the time bin. To simplify the terminology, we call this process "Poisson" to emphasize that it generates spike independently in each time bin. We then generated spikes from a third Poisson process, and these spikes were added to both cells in order to create cross-correlation. To avoid synchronization with infinite time precision, the shared spike times added to the second cell were shifted together in time by a random amount chosen anew for each trial from a zero-mean gaussian distribution with standard deviation of 1 ms. In all cases, a joint spike timing code over a 60 ms long window with a time precision of 10 ms was considered.



The left panels plot the mean firing rate of the two cells; the central panels plot the cross-correlogram (CCG; computed analytically from knowledge of the simulated processes); the right panels report the values of the lower bounds I_{LB1} and I_{LB2} and the true information. In this information estimation, we considered a large number of trials per stimulus in order to focus only on the asymptotic estimations (see main text). (A) Uncorrelated spike trains. A Poisson process is used for each cell, and no shared spikes are added. Both I_{LB1} and I_{LB2} give good estimations. (B) Correlated spikes with weak stimulus modulation of correlation. The ratio of independent versus shared spikes was approximately the same for both stimuli, and hence the correlation strength $\gamma(\cdot)$ was only weakly stimulus modulated. The presence of stimulus-independent correlation makes I_{LB1} markedly different from I. However, I_{LB2} remained a precise estimator of information. The simulation parameters were as follows. For both cells, the mean rate of the independently generated spikes was 32 Hz for the first stimulus and 16 Hz for the second stimulus. The mean rate of the shared spikes was 8 Hz to the first stimulus and 4 Hz to the second one. (C) Correlated spikes with strong stimulus modulation of correlation. The ratio of independent versus shared spikes was much higher for the first stimulus than for the second. As a consequence, the $\gamma(\cdot)$ coefficients were stimulus dependent. The stimulus modulation of correlation made I_{LB2} smaller than I; I_{LB1} was negative. The data were generated as follows. For both cells, the mean rate of the independently generated spikes was 15 Hz for the first stimulus and 19 Hz for the second stimulus. The mean rate of the shared spikes was 25 Hz to the first stimulus and 1 Hz to the second one.

We first report derivations of approximate analytical estimates of the sampling biases of I_{LB1} and I_{LB2}. These bias equations provide useful insights into the relative sampling properties of each information quantity, and they can also be used to quantify and subtract out the bias in real experimental conditions. After considering the analytical approximations to the bias, we perform numerical simulations to test independently both the validity of the analytical estimates and the performance of bias removal procedures based on the subtraction of the above analytical estimates. The extent to which these corrected estimates might be improved further by applying alternative methods to control the bias (Victor, 2002; Nemenman, Shafee, & Bialek, 2002; Paninski, 2003) will not be investigated here.

6.1 Analytical Estimates of the Magnitude of Bias Properties of Lower-Bound Estimators. The mutual information is $H(\mathcal{R}) - H(\mathcal{R}|\mathcal{S})$. I_{LB1} is $H(\mathcal{R}) - H_{ind}(\mathcal{R}|\mathcal{S})$, and I_{LB2} is $\chi(\mathcal{R}) - H_{ind}(\mathcal{R}|\mathcal{S})$. Thus, the relative sampling properties of the information and its two lower-bound estimators can be established by considering the sampling properties of the four quantities $H(\mathcal{R}|\mathcal{S})$, $H_{ind}(\mathcal{R}|\mathcal{S})$, $H(\mathcal{R})$, and $\chi(\mathcal{R})$.

Our approximations to the bias of these quantities were derived on the assumption that the number of trials per stimulus is large, so that the probability of each response is empirically sampled on the basis of many available trials (Panzeri & Treves, 1996). We will use the symbol $\approx$ to indicate that we report only the leading term of the perturbative evaluation in $1/N$ of the bias (N being the total number of trials across all stimuli).

The bias of a given functional of the probability distributions is defined as the difference between the trial-averaged value of the functional when the probability distributions are computed from N trials and the value of the functional computed with the true probability distributions.

We first consider the noise entropy $H(\mathcal{R}|\mathcal{S})$. The analytical expression for its bias is

$$Bias[H(\mathcal{R}|\mathcal{S})] \approx -\frac{1}{2N \ln 2} \sum_s (\tilde{R}(s) - 1), \tag{6.1}$$

where N is the number of trials across all stimuli presentations and $\tilde{R}(s)$ denotes the number of "relevant" responses of the stimulus conditional response probability distribution $P(\mathbf{r}|s)$, that is, the number of different responses $\mathbf{r}$ with nonzero probability of being observed when stimulus s is presented (Panzeri & Treves, 1996). Like all other entropy quantities, the noise entropy is biased downward when sampled with limited trials. Since $H(\mathcal{R}|\mathcal{S})$ depends on $P(\mathbf{r}|s)$, the number of relevant responses in the numerator of equation 6.1 is of order $(M+1)^{LC}$ for each stimulus in the summation. It follows that for this bias to be small, N should be bigger than $S \times (M+1)^{LC}$.

The analytical expression for the bias of $H(\mathcal{R})$ is

$$Bias[H(\mathcal{R})] \approx -\frac{1}{2N\ln 2}(\tilde{R}-1), \tag{6.2}$$

where $\tilde{R}$ is the number of relevant responses of $P(\mathbf{r})$. $H(\mathcal{R})$ is still difficult to sample because $\tilde{R}$ is still of order $(M+1)^{LC}$. However, since $H(\mathcal{R})$ depends on only $P(\mathbf{r})$, its bias is approximately S times smaller than the bias of $H(\mathcal{R}|\mathcal{S})$. This is an advantage when many different stimuli are presented.

The bias of the mutual information $I(\mathcal{R};\mathcal{S})$ is the difference between the biases of $H(\mathcal{R})$ and $H(\mathcal{R}|\mathcal{S})$. As the most biased term is $H(\mathcal{R}|\mathcal{S})$, the mutual information is upward biased Panzeri & Treves, 1996).

Estimating the number of relevant responses to compute the bias in equations 6.1 and 6.2 from small data samples is nontrivial. Panzeri & Treves, (1996) have proposed a "Bayes" procedure to estimate these parameters empirically. This approach works well when there are at least two to four times as many trials per stimulus as the number of parameters describing the responses, that is, $(M+1)^{LC}$ (Panzeri & Treves, 1996; Pola et al., 2003). This bias estimate can be subtracted from the raw information estimate to get accurate and unbiased results. Throughout this letter, we will use this procedure.

Let us consider now $H_{ind}(\mathcal{R}|\mathcal{S})$. Since it can be expressed as the sum of simpler entropies (see equation 4.5), its bias has the following expression:

$$Bias[H_{ind}(\mathcal{R}|\mathcal{S})] \approx -\frac{1}{2N\ln 2}\sum_{c=1}^{C}\sum_{t=1}^{L}\sum_{s}(\tilde{R}_{ct}(s)-1), \tag{6.3}$$

where $\tilde{R}_{ct}(s)$ is the number of relevant responses of the marginal distributions $P(r_c(t)|s)$. As $\tilde{R}_{c,t}(s)$ is of order $M+1$, the bias of $H_{ind}(\mathcal{R}|\mathcal{S})$ is proportional to MLC and is thus much smaller than that of $H(\mathcal{R}|\mathcal{S})$. As a consequence, the bias of I_{LB1} (which is simply the difference between the biases of $H(\mathcal{R})$ and $H_{ind}(\mathcal{R}|\mathcal{S})$) is dominated by the bias of $H(\mathcal{R})$ and is therefore smaller by a factor of S than the bias of $I(\mathcal{R};\mathcal{S})$.

Like $H(\mathcal{R})$, $\chi(\mathcal{R})$ depends on only the stimulus unconditional probability distributions. However, it has a feature that makes its bias properties much better than $H(\mathcal{R})$. Bias arises from the logarithmic form of entropy functionals. The log in $\chi(\mathcal{R})$ depends on $P_{ind}(\mathbf{r})$. Since $P_{ind}(\mathbf{r})$ is better sampled than $P(\mathbf{r})$, $\chi(\mathcal{R})$ has less bias than $H(\mathcal{R})$, whose log depends on $P(\mathbf{r})$. As a consequence, the bias of $\chi(\mathcal{R})$ is much smaller than the bias of $H(\mathcal{R})$. In particular, the bias of $\chi(\mathcal{R})$ (whose expression is reported in equation A.4 in the appendix) scales approximately quadratically with LC, whereas that of $H(\mathcal{R})$ scales exponentially. Differences in sampling properties between $\chi(\mathcal{R})$ and $H(\mathcal{R})$ get more pronounced if the response space is high dimensional.

The bias of I_{LB2} is the difference between the biases of $\chi(\mathcal{R})$ and $H_{ind}(\mathcal{R};\mathcal{S})$. Since $\chi(\mathcal{R})$ is less biased than $H(\mathcal{R})$, the bias of I_{LB2} is much

smaller than that of I_{LB1} (and thus of that of $I(\mathcal{R};\mathcal{S})$). From the above mathematical considerations, it is also expected that the improvement of the bias of I_{LB2} does not come at the expense of an increase in variance.

6.2 Investigation of the Bias Properties of Lower Bounds by Mean of Computer Simulations. In this section, we perform numerical simulations to validate the analytical estimates of the bias reported above and test the performance of bias removal procedures based on the subtraction of the analytical estimates. We performed extensive simulations with both correlated and uncorrelated data; however, for simplicity, we report only results from typical simulations, which summarize the general findings.

The first computer simulation (see Figure 2) consists of a neuronal pair. The response time window was 60 ms long, and spikes were digitized with a time precision of 10 ms, with each time bin containing 0 or 1 spikes (i.e., $LC = 12$ and $M = 1$). We considered four stimulus conditions. The firing rate of each cell in the different stimulus conditions was in the range 12 to 48 Hz. The spike trains were designed to have weak stimulus modulation of the correlated activity.

We started by studying the bias of the entropy quantities $\chi(\mathcal{R})$, $H(\mathcal{R})$, $H_{ind}(\mathcal{R}|\mathcal{S})$, and $H(\mathcal{R}|\mathcal{S})$ as a function of the number of trials per stimulus. In Figure 2A we report the values for the above quantities when obtained by a direct evaluation of equations 5.9, 2.2, 4.2, and 2.3, without application of any bias correction procedure. In agreement with the analytical result obtained above, $\chi(\mathcal{R})$ and $H_{ind}(\mathcal{R}|\mathcal{S})$ were much less downward biased than $H(\mathcal{R})$ and $H(\mathcal{R}|\mathcal{S})$.

We then studied the sampling properties of the mutual information $I(\mathcal{R};\mathcal{S})$ and its two lower-bound estimators I_{LB1} and I_{LB2} (see Figure 2B). I_{LB1} had better bias properties than the full mutual information $I(\mathcal{R};\mathcal{S})$. However, it was two orders of magnitude less data robust than I_{LB2}. Even without any sampling bias correction procedure, I_{LB2} was well estimated with 2×10^2 trials per stimulus, while we needed, respectively, 3×10^4 and more than 10^5 to estimate I_{LB1} and $I(\mathcal{R};\mathcal{S})$ with similar accuracy.

In Figures 2C and 2D, we report the data sampling behavior of both entropy and information quantities after subtracting the bias estimates. The probabilities entering the bias of $\chi(\mathcal{R})$, equation A.4, were estimated directly from the experimental probabilities, whereas the number of relevant responses entering the entropy expression was estimated using the procedure of Panzeri & Treves, (1996). Convergence to the asymptotic value of the corrected estimates (see Figures 2C and 2D) was much better than in the "raw" case: I_{LB2} was well estimated with only 50 trials per stimulus, whereas we needed 3×10^3 and 10^4 trials per stimulus to estimate I_{LB1} and $I(\mathcal{R};\mathcal{S})$, respectively. With $LC = 12$, the number of possible different responses was $2^{12} = 4096$ in this simulation. Thus, approximately two to four times more trials per stimulus than response classes were needed to obtain precise estimates of $I(\mathcal{R};\mathcal{S})$, whereas I_{LB2} was well sampled with a number of trials

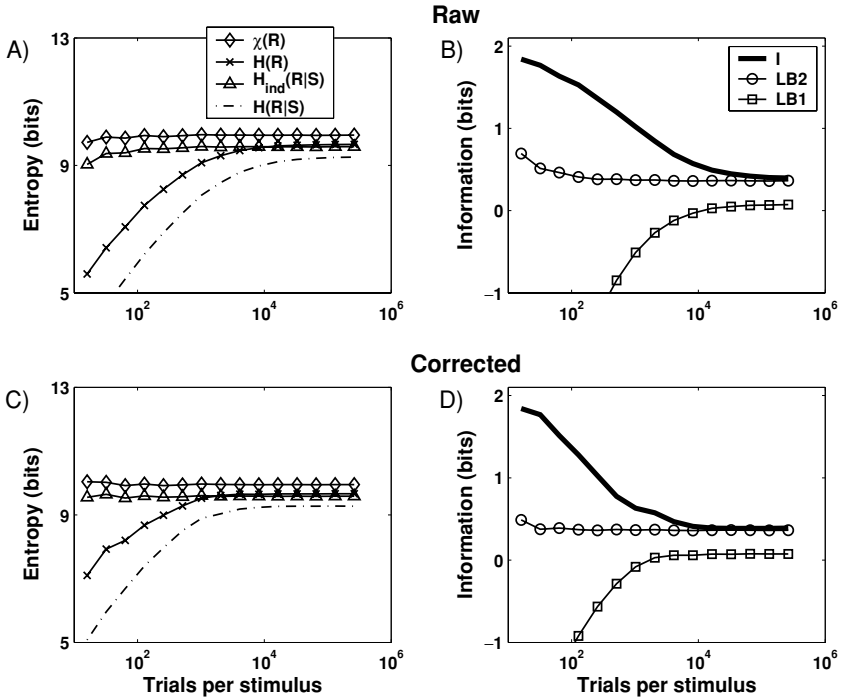


Figure 2: Sampling behavior of the lower-bound estimators I_{LB1} and I_{LB2}. We generated correlated simulated data with weak stimulus modulation of correlation and tested how the estimates depend on the data size. We considered a neuronal pair responding to four different stimuli in a time window of 60 ms and a time precision of 10 ms. The simulated spike trains have been produced as in Figure 1B. The simulation parameters were as follows. For both cells, the mean rate of the independently generated spikes was, respectively, for the four stimulus conditions 32 Hz, 24 Hz, 16 Hz, and 8 Hz. The mean rate of the shared spikes was 16 Hz, 12 Hz, 8 Hz, and 4 Hz. In A and B, we report the raw values of the information estimators, entropy, and like-entropy terms where in C and D the values are corrected for finite sampling (see section 6). Results were averaged over repetitions of the simulation decreasing as the number of trials per stimulus available increases. (a) Raw values of $\chi(R)$, $H(R)$, $H_{ind}(R|S)$, and $H(R|S)$ obtained without using any bias corrections. (b) Raw values of $I(\mathcal{R};\mathcal{S})$, I_{LB1}, and I_{LB2}. (c) Values of $\chi(R)$, $H(R)$, $H_{ind}(R|S)$, and $H(R|S)$ obtained after subtracting the bias corrections described in the text. (d) Corrected values of $I(\mathcal{R};\mathcal{S})$, I_{LB1}, and I_{LB2}. Each value of the plot is obtained averaging over random repetitions of the same simulation.

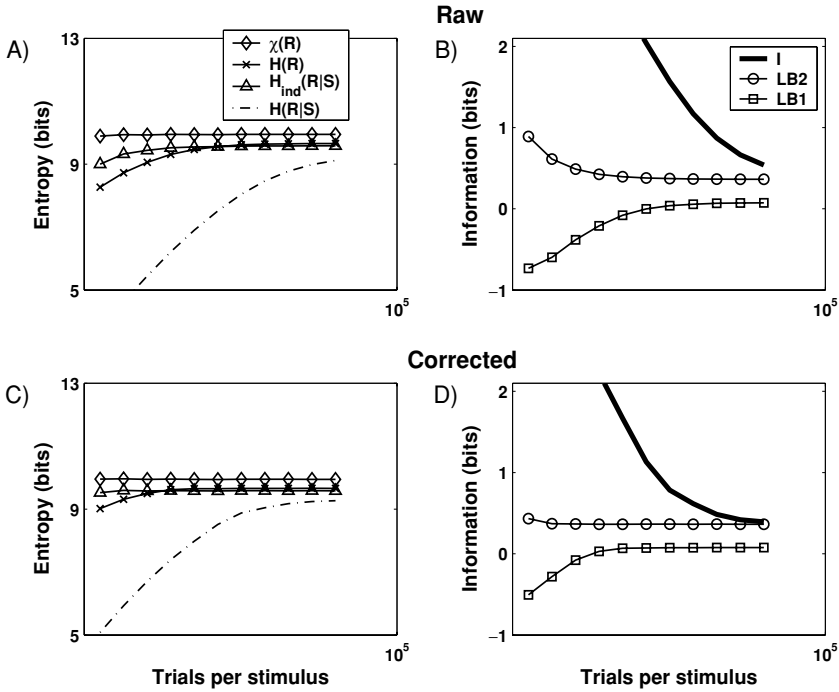


Figure 3: Sampling behavior of the lower-bound estimators I_{LB1} and I_{LB2}. Conventions are as in Figure 2. We generated spike trains as in Figure 2; however, this time we considered 64 stimuli.

per stimulus that was approximately 80 times smaller than the number of possible responses.

In order to study the effect of increasing the number of stimuli while keeping the number of trials per stimulus fixed, we considered next (see Figure 3) the same simulated responses as in Figure 2, but using 64 stimulus conditions rather than 4. Figure 3 shows that $H(\mathcal{R})$ was better sampled than with 4 stimuli. As a consequence, I_{LB1} worked much better. However, the sampling behavior of $\chi(\mathcal{R})$ also improved. Both I_{LB1} and I_{LB2} improved their sampling properties when increasing the stimulus set size, but I_{LB2} retained overall better sampling properties. The noise entropy $H(\mathcal{R}|\mathcal{S})$ was still badly sampled when the stimulus size was increased but the number of trials per stimulus was kept fixed. As a consequence, the mutual information $I(\mathcal{R};\mathcal{S})$ did not improve its sampling properties (see Figure 3D).

The above results on the effect of using a large stimulus set have some implications for particular stimulation paradigms, such as a dynamic stimulus (Nirenberg et al., 2001) or an m-sequence (Reich et al., 2001) in which thousands of different stimuli are available. In such cases, both $P(\mathbf{r})$ and $P_{ind}(\mathbf{r})$ will be extremely well sampled, whereas the stimulus-conditional

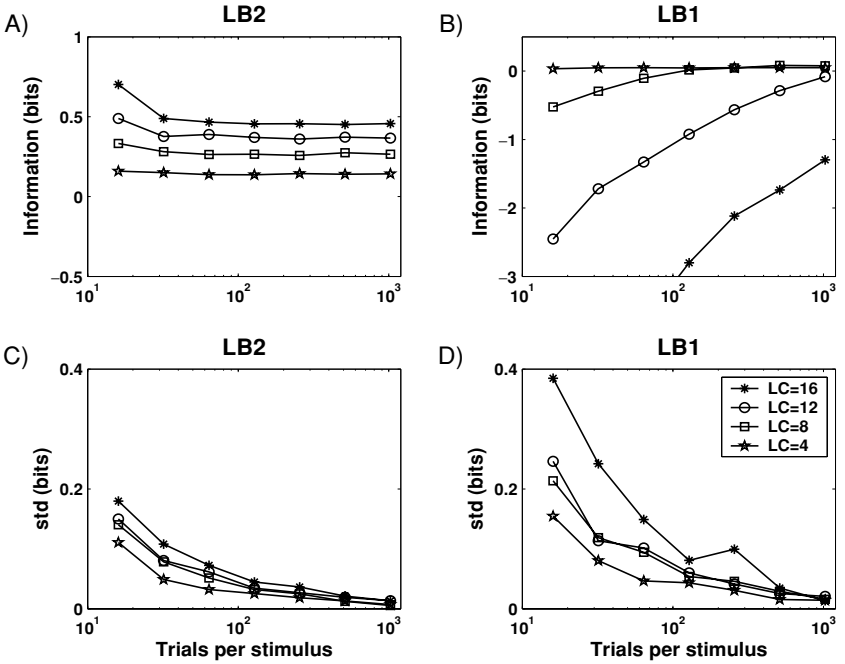


Figure 4: Lower-bounds values and standard deviations of I_{LB1} and I_{LB2}. We plotted the mean values and the standard deviations of the lower bounds as a function of the number of trials per stimulus available. The values of the estimates and the standard deviations are computed averaging over repetitions of the same simulation (see main text). We simulated a neuronal pair as in Figure 2. We considered a time precision of 10 ms and time window, respectively, of 20 ms, 40 ms, 60 ms, and 80 ms; thus, LC was 4, 8, 12, and 16 for this neuronal pair. In these plots, we considered the values corrected for finite sampling as described in the main text. (a) I_{LB2} estimates. (b) I_{LB1} estimates. (c) I_{LB2} standard deviations. (d) I_{LB1} standard deviations.

probabilities will not be as well sampled. Thus, both I_{LB1} and I_{LB2} will be generally well sampled and relatively bias free when using such very large stimulus set (with I_{LB2}, however, retaining an advantage in tightness), whereas $I(\mathcal{R};\mathcal{S})$ will not be as well sampled.

In Figure 4 we report a study of the performance of the estimators when increasing the number of time bins L. We studied both the convergence of the estimators to the true asymptotic values and the standard deviation of the estimates on the available number of trials per stimulus. As in Figure 2, we simulated responses of two neurons to four different stimulus conditions, and we considered windows of length $L = 2, 4, 6, 8$. In this plot we considered only bias-subtracted values. In Figure 4A, we show the results for I_{LB2}. We found that approximately 50 trials per stimulus were enough

to accurately estimate I_{LB2}, even up to $LC = 16$ (50 trials is more than three orders of magnitude less than the size of the response space, that is, 2^{16}). A direct estimation of $I(\mathcal{R};\mathcal{S})$ through equation 2.1 required at least 3×10^5 trials per stimulus. To get accurate estimates of I_{LB1} (see Figure 4B), we needed many more data than for I_{LB2}. The downward bias of I_{LB1} made it negative in conditions of undersampling (see Figure 4B). In Figures 4C and 4D, we show the standard deviations of I_{LB2} and I_{LB1}, obtained over different random repetitions of the same simulated process. As the number of trials per stimulus increased, the standard deviation decreased. Increasing the number of time bins affected only weakly the standard deviations of the estimates. The standard deviation of I_{LB1} was larger than that of I_{LB2}, suggesting that I_{LB2} is not only a less biased estimator of the information than I_{LB1} but is also less variable.

7 A Tighter Lower Bound Including Short-Time-Range Stimulus-Dependent Correlations

I_{LB2} is data robust and can lead to precise and tight information estimates even in the presence of strong, nonstimulus-modulated, spike timing correlation. However, when there is significant stimulus modulation of correlations and $I_{cor-dep}$ is not negligible, then I_{LB2} may not quantify precisely the transmitted information. Although neurophysiological experiments reported so far have all found $I_{cor-dep}$ to be a small proportion of the total information, it is conceivable that stimulus modulations of correlation may convey substantial information in some neural systems or under specific stimulus conditions. A relevant question is thus how the total information can be better approximated with a data-robust bound in this case. This section addresses this issue by suggesting a new strategy, which consists of neglecting only stimulus modulations of long time-range correlations.

7.1 The Markov Approximation to Model the Stimulus-Response Probability. The reason that both $I_{cor-dep}$ and $I(\mathcal{R};\mathcal{S})$ are strongly biased is that they depend on $P(\mathbf{r}|s)$, and the latter takes into consideration the complete history of firing: the probability of the neuronal response $\mathbf{r}(t)$ in the tth time bin is affected by the neural responses in all the previous time bins. This is made explicit by expressing $P(\mathbf{r}|s)$ using the chain rule (Cover & Thomas, 1991):

$$\begin{aligned} P(\mathbf{r}|s) &= P(\mathbf{r}(1), \mathbf{r}(2), \ldots, \mathbf{r}(t), \ldots, \mathbf{r}(L-1), \mathbf{r}(L)|s) \\ &= P(\mathbf{r}(1)|s)P(\mathbf{r}(2)|\mathbf{r}(1), s)P(\mathbf{r}(3)|\mathbf{r}(1), \mathbf{r}(2), s)\ldots \\ &\quad \times P(\mathbf{r}(t)|\mathbf{r}(1), \ldots, \mathbf{r}(t-1), s)\ldots P(\mathbf{r}(L)|\mathbf{r}(1), \ldots, \mathbf{r}(L-1), s). \\ &= P(\mathbf{r}(1)|s)\prod_{t=2}^{L} P(\mathbf{r}(t)|\mathbf{r}(1), \ldots, \mathbf{r}(t-1), s). \end{aligned} \tag{7.1}$$

However, in many neural systems, correlations are significant only between spikes that are separated by a short time lag, in the range of 1 to 15 ms (Gray, König, Engel, & Singer, 1989; Brosch, Bauer, & Eckhorn, 1997; Dan et al., 1998; Nirenberg et al., 2001; Golledge et al., 2003). In such cases, to preserve the entire information, it is sufficient to take into account only correlations extending over a short lag. Our approach will be to approximate the real probability of current response $\mathbf{r}(t)$ given the past firing with a finite-memory Markov model that looks back to only q time steps, as follows:

$$P(\mathbf{r}(t)|\mathbf{r}(1), \mathbf{r}(2), \ldots, \mathbf{r}(t-1), s) \to \tilde{P}_q(\mathbf{r}(t)|\mathbf{r}(t-q), \ldots, \mathbf{r}(t-1), s). \tag{7.2}$$

The latter is computed from the experimental probabilities via Bayes' rule:

$$\tilde{P}_q(\mathbf{r}(t)|\mathbf{r}(t-q), \ldots, \mathbf{r}(t-1), s) = \frac{P(\mathbf{r}(t-q), \ldots, \mathbf{r}(t-1), \mathbf{r}(t)|s)}{P(\mathbf{r}(t-q), \ldots, \mathbf{r}(t-1)|s)}. \tag{7.3}$$

$P(\mathbf{r}(t-q), \ldots, \mathbf{r}(t-1), \mathbf{r}(t)|s)$ and $P(\mathbf{r}(t-q), \ldots, \mathbf{r}(t-1)|s)$ are marginal distributions of the full model $P(\mathbf{r}|s)$. They can be computed by integrating away the dependence on all the response variables that do not enter in their argument:

$$P(\mathbf{r}(t-q), \ldots, \mathbf{r}(t-1), \mathbf{r}(t)|s) = \sum_{\mathbf{r}(1),\ldots,\mathbf{r}(t-q-1)} \sum_{\mathbf{r}(t+1),\ldots,\mathbf{r}(L)} P(\mathbf{r}|s), \tag{7.4}$$

$$P(\mathbf{r}(t-q), \ldots, \mathbf{r}(t-1)|s) = \sum_{\mathbf{r}(1),\ldots,\mathbf{r}(t-q-1)} \sum_{\mathbf{r}(t),\ldots,\mathbf{r}(L)} P(\mathbf{r}|s). \tag{7.5}$$

By using the above equation and the chain rule, one arrives at the following equation for the q-length Markov probability of the stimulus-conditional response probability:

$$\begin{aligned} \tilde{P}_q(\mathbf{r}|s) &= P(\mathbf{r}(1)|s) \prod_{t=2}^{L} \tilde{P}_q(\mathbf{r}(t)|\mathbf{r}(t-q), \ldots, \mathbf{r}(t-1), s), \\ &\quad \text{if } q = 1, \ldots, L-1, \\ \tilde{P}_0(\mathbf{r}|s) &= \prod_{t=1}^{L} P(\mathbf{r}(t)|s) \quad \text{if } q = 0 \end{aligned} \tag{7.6}$$

(in particular, $\tilde{P}_{L-1}(\mathbf{r}|s) = P(\mathbf{r}|s)$). $\tilde{P}_q(\mathbf{r}|s)$ preserves all correlations (both cross-cell and within-cell) extending up to q time bins in the past, and it

Table 1: Number of Free Parameters Required to Specify the Probability Models for Each Stimulus Configuration s, Computed Assuming That There Is at Most One Spike per Time Bin ($M = 1$).

$P_{ind}(\mathbf{r}\|s)$	LC
$\tilde{P}_0(\mathbf{r}\|s)$	$L(2^C - 1)$
$\tilde{P}_1(\mathbf{r}\|s)$	$2^C - 1 + (2^C - 1)2^C(L - 1)$
...	...
$\tilde{P}_q(\mathbf{r}\|s)$	$2^{qC} - 1 + (2^C - 1)2^{qC}(L - q)$
...	...
$P(\mathbf{r}\|s)$	$2^{LC} - 1$

Notes: P, P_{ind}, and $\tilde{P}_q$ denote the full probability model, the independent model, and the q-step Markov model, respectively. C and L denote the number of cells and time bins, respectively.

neglects all correlations of range longer than q. Thus, it is a perfect description of neuronal firing if correlations extend to a lag shorter than or equal to q time bins. The longer is the range of correlations q included in the model, the closer is the Markov model to the real distribution $P(\mathbf{r}|s)$. The notion of closeness can be proved rigorously as follows. First, the q-length Markov model preserves all marginals of the original probability distribution extending up to $q + 1$ consecutive time bins:

$$\tilde{P}_q(\mathbf{r}(t-q), \ldots, \mathbf{r}(t-1), \mathbf{r}(t)|s) = P(\mathbf{r}(t-q), \ldots, \mathbf{r}(t-1), \mathbf{r}(t)|s), \quad (7.7)$$

for $t = q + 1, \ldots, L$. Second, it can be shown that the conditional Kullback-Leibler distance between $P(\mathbf{r}|s)$ and $\tilde{P}_q(\mathbf{r}|s)$ decreases as q increases,

$$D(P(\mathbf{r}|s) \parallel \tilde{P}_q(\mathbf{r}|s)) \geq D(P(\mathbf{r}|s) \parallel \tilde{P}_{q+1}(\mathbf{r}|s)), \quad (7.8)$$

for every $q = 0, \ldots, L - 2$.

As q increases, the probability model gets more and more complex, and it needs bigger data samples to be well estimated. The number of free parameters needed to specify the Markov model $\tilde{P}_q(\mathbf{r}|s)$ (reported in Table 1 for the case of $M = 1$—up to one spike in each time bin) grows with q. It can be seen that in this case, the number of free parameters of the Markov model is in between the LC parameters needed to describe the independent probability model $P_{ind}(\mathbf{r}|s)$ and the $2^{LC} - 1$ parameters describing the general model $P(\mathbf{r}|s)$. Thus, Markov models with larger q are more accurate, whereas Markov models with small q are more data robust.

7.2 Tighter Lower Bounds I^q_{LB3} to Include the Contribution of Stimulus Modulation of Correlation. The Markov probabilities can be used to

compute tighter data-robust lower bounds. For each $q = 0, \ldots, L-1$, we define the following information lower bound:

$$I^q_{LB3} = \chi^q(\mathcal{R}) - H^q(\mathcal{R}|\mathcal{S}), \tag{7.9}$$

where

$$\begin{aligned} \chi^q(\mathcal{R}) &= -\sum_{\mathbf{r}\in\mathcal{R}} P(\mathbf{r}) \log_2 \tilde{P}_q(\mathbf{r}), \\ H^q(\mathcal{R}|\mathcal{S}) &= -\sum_{s\in\mathcal{S}} P(s) \sum_{\mathbf{r}\in\mathcal{R}} \tilde{P}_q(\mathbf{r}|s) \log_2 \tilde{P}_q(\mathbf{r}|s), \end{aligned} \tag{7.10}$$

and $\tilde{P}_q(\mathbf{r}) = \sum_s P(s)\tilde{P}_q(\mathbf{r}|s)$. These lower bounds have a very similar expression to I_{LB2}, the difference being that $P_{ind}(\mathbf{r}|s)$ is replaced by $\tilde{P}_q(\mathbf{r}|s)$. The fact that I^q_{LB3}, for $q = 0, 1, \ldots, L-1$, are lower bounds to the total information is proved by the following:

$$I(\mathcal{R};\mathcal{S}) - I^q_{LB3} = D(P(s|\mathbf{r}) \parallel \tilde{P}_q(s|\mathbf{r})) \geq 0. \tag{7.11}$$

To understand the conditions in which I^q_{LB3} is a tight information lower bound, we note that $D(P(s|\mathbf{r}) \parallel \tilde{P}_q(s|\mathbf{r}))$, the difference between the total information and I^q_{LB3}, can be interpreted (see equation. 5.10) as the stimulus-dependent correlational component related to a correlation coefficient defined as

$$\begin{aligned} \gamma^q(\mathbf{r}|s) &= \frac{P(\mathbf{r}|s)}{\tilde{P}_q(\mathbf{r}|s)} - 1, \quad \text{if } \tilde{P}_q(\mathbf{r}|s) \neq 0, \\ \gamma^q(\mathbf{r}|s) &= 0, \qquad\qquad\quad \text{if } \tilde{P}_q(\mathbf{r}|s) = 0. \end{aligned} \tag{7.12}$$

$D(P(s|\mathbf{r}) \parallel \tilde{P}_q(s|\mathbf{r}))$ is always positive or zero, and it is zero if and only if $\gamma^q(\mathbf{r}|s)$ does not depend on s for every response $\mathbf{r}$. Since $\gamma^q(\mathbf{r}|s)$ is nonzero only when there are correlations extending over a time range greater than q, I^q_{LB3} is not tight only if there are stimulus-modulated correlations occurring over a time range spanning more than q time steps.

7.3 Bias Properties of I^q_{LB3}. The bias of I^q_{LB3} is given by the difference between the biases of $\chi^q(\mathcal{R})$ and $H^q(\mathcal{R}|\mathcal{S})$.

The bias of I^q_{LB3} is largely characterized by that of $H^q(\mathcal{R}|\mathcal{S})$, because $\chi^q(\mathcal{R})$ is better sampled than $H^q(\mathcal{R}|\mathcal{S})$. In fact, (1) $\chi^q(\mathcal{R})$ depends on $\tilde{P}_q(\mathbf{r})$ and $P(\mathbf{r})$, while $H^q(\mathcal{R}|\mathcal{S})$ is a functional of the stimulus-conditional probabilities $\tilde{P}_q(\mathbf{r}|s)$, and (2) $\chi^q(\mathcal{R})$ depends linearly on $P(\mathbf{r})$, the argument of the logarithmic part being the much better sampled distributions $\tilde{P}_q(\mathbf{r}|s)$.

The expression for the bias of $H^q(\mathcal{R}|\mathcal{S})$ is reported in the appendix (see equation A.2). As shown in the appendix, $H^q(\mathcal{R}|\mathcal{S})$ can be decomposed into

a sum of lower-dimensional entropies of the marginal probability distributions of up to $q+1$ time bins together. For this reason, the bias of $H^q(\mathcal{R}|\mathcal{S})$ is smaller than that of $H(\mathcal{R}|\mathcal{S})$, and it is larger for larger q values. Thus, the larger the range of the Markov model, the larger the samples needed to get accurate estimates.

The least biased of all I^q_{LB3} estimators is I^0_{LB3}. In this case, the bias of $H^0(\mathcal{R}|\mathcal{S})$ is reported in equation A.3. $\chi^0(\mathcal{R})$ also has a small bias (see equation A.18 in the appendix), which scales approximately as L^2.

We investigated numerically the properties of I^0_{LB3} by applying this analysis to synthetic spike trains.[5] We simulated a neuronal pair in a poststimulus time window of 80 ms with spike times digitized with a precision of 10 ms (thus, $LC = 16$). In this simulation, the stimulus-dependent correlational component was about 60% of the total information. The cross-correlations were short-ranged; data were generated in such a way that the CCG (not shown) had a gaussian shape with width 1 ms. Thus, given bin sizes of 10 ms, we would expect to recover all information by using I^0_{LB3}. Results of the simulations are reported in Figure 5. Since $I_{cor-dep}$ was a substantial fraction of the total information, I_{LB2} was not a tight estimator of the total information. However, I^0_{LB3} recovered all the information not captured by I_{LB2}. The behavior of $I(\mathcal{R};\mathcal{S})$, I_{LB2}, and I^0_{LB3} when varying the number of trials per stimulus shows that I^0_{LB3} is much more data robust than $I(\mathcal{R};\mathcal{S})$ and almost as data robust as I_{LB2}. After correcting for the bias, an accurate estimate of $I(\mathcal{R};\mathcal{S})$ required about 3×10^5 trials per stimulus. Both I_{LB2} and I^0_{LB3} required only 50 to 100 trials per stimulus. Given that there were $2^{16} = 65,536$ possible responses, this is extremely good sampling behavior.

8 Application to Neurophysiological Data

To illustrate and evaluate their possible practical applications, we apply the new lower-bound methods to real neuronal recordings from somatosensory cortex of anesthetized rats and from cortical visual area MT of awake macaques.

8.1 Spike Timing, Spike Count, and Short-Time-Range Correlations in Rat Somatosensory Cortex. We first apply the new method to spike trains recorded from the whisker representation in somatosensory ("barrel") cortex of rats anesthetized with urethane.

In this example, we analyze and compare two different data sets of S1 neuronal activity recorded with different techniques under the same stimulation paradigm. The first data set (kindly provided to us by M. Lebedev and M. Diamond; see Lebedev, Mirabella, Erchova, & Diamond, 2000, for

[5] A detailed numerical investigation of I^q_{LB3} with $q > 0$ will appear elsewhere (Panzeri, 2005).

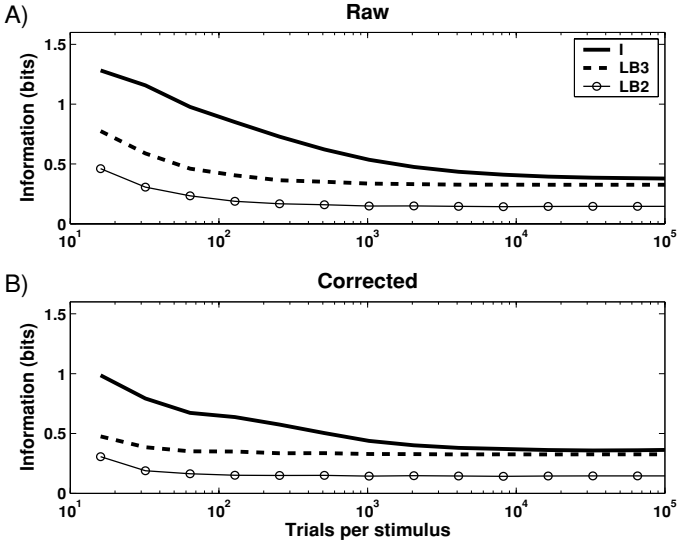


Figure 5: Sampling behavior of the lower-bound estimators I_{LB2} and I^0_{LB3}. The mutual information and the lower-bound estimates are plotted as a function of the number of trials per stimulus available; the values of the estimates are computed averaging over repetitions of the same simulation. We simulated a neuronal pair responding to four stimuli. We considered a time precision of 10 ms and a time window of 80 ms; thus, LC was equal to 16 for this pair. We considered a simulation with a strong stimulus modulation of the correlation. In this case, the ratio of independent versus shared spikes was modulated with the stimulus set. As a consequence, the $\gamma(\cdot)$ coefficients were stimulus dependent. The data were generated as follows. For both cells, the mean rate of the independently generated spikes was, respectively, for the four stimuli 16 Hz, 36 Hz, 8 Hz, and 8 Hz; the mean rate of the shared spikes was 32 Hz, 0 Hz, 16 Hz, and 4 Hz. As in the simulations in Figure 1, the shared spikes in the second neuron were shifted in time by a random amount chosen anew for each trial from a zero-mean gaussian distribution with standard deviation of 1 ms; this makes neurons nearly synchronous. (a) Lower-bounds estimates compared with the true mutual information. The values are not corrected for finite sampling. (b) Bias-corrected values of the lower-bounds estimates and the mutual information.

further details) consisted of single-neuron activity, each neuron being from a different tungsten electrode. The second data set consisted of multiunit activity (MUA) that was recorded from each electrode of a silicon electrode array (see Petersen & Diamond, 2000, for further details). For this MUA data set, it has been estimated that each electrode captured the spikes of a small cluster of neurons (≈2–5; see Petersen & Diamond, 2000). In both data sets,

neural activity was recorded in response to individual stimulation of one of nine different whiskers (whisker D2 and its eight nearest neighbors); individual whiskers were stimulated near their base by a piezoelectric wafer, controlled by a voltage generator. The stimulus was an up-down step function of 80 μm amplitude and 100 msec duration, delivered once per second. The trials per stimulus available were 50 for the single-unit data set and 500 for the MUA data set.

The interest in comparing the two data sets arises from the fact that the single-unit and the MUA data set have very different numbers of trials per stimulus available (50 versus 500) and that the MUA activity is more correlated than the single-unit activity. In fact, previous analysis(Panzeri, Petersen, et al., 2001; Petersen et al., 2001; Petersen & Diamond, 2000) on these S1 neurons has shown that spikes from the same cell have a weak negative autocorrelation (a period of refactoriness or inhibition follows a spike from the same cell), whereas nearby neurons have a substantial positive, near-synchronous cross-correlation (Lebedev et al., 2000). Thus, MUA contains also a positive correlation (resulting from correlations between nearby neurons) that is not present in the single-unit spike trains. It is interesting to study how our methods behave in these two different conditions of sampling and correlation sources.

The time course of the estimates of the information transmitted about stimulus location by spike times of single cells (averaged over all 10 single cells in this dataset) is reported in Figure 6A. We used the procedure of Panzeri and Treves (1996) described in section 6 to correct the information estimates for finite sampling. We increased the poststimulus time windows from 0 to 80 ms, using 10 ms time bins to digitize the spike train. We compared the time course of the full spike timing information $I(\mathcal{R};\mathcal{S})$ to that of the lower bounds I_{LB1} and I_{LB2}. To quantify whether spike timing added extra information to that conveyed by spike counts alone, we also computed the time course of the spike count information $I_c(\mathcal{R};\mathcal{S})$, the latter being computed from equation 2.1 after quantifying neuronal responses as the total number of spikes emitted in each trial in the poststimulus window considered. The full spike timing information increased smoothly until 30 to 40 ms and then diverged rapidly. This is due to failure of removing the sampling bias, consistent with the rules of thumb for sampling correction, given in section 6, which predict that, with 50 trials per stimulus available, the mutual information should be well estimated only up to three to four time bins.

In contrast to the total spike timing information $I(\mathcal{R};\mathcal{S})$, its two lower-bound estimators did not diverge over time. This indicated that I_{LB1} and I_{LB2} were better sampled than $I(\mathcal{R};\mathcal{S})$, consistent with the above simulation results. The spike count information $I_c(\mathcal{R};\mathcal{S})$ also varied smoothly with time and was well sampled given that the firing rates in this data set were low (Lebedev et al., 2000). I_{LB2} was very close to the total spike timing information $I(\mathcal{R};\mathcal{S})$ for the whole 0 to 40 ms range, in which both quantities

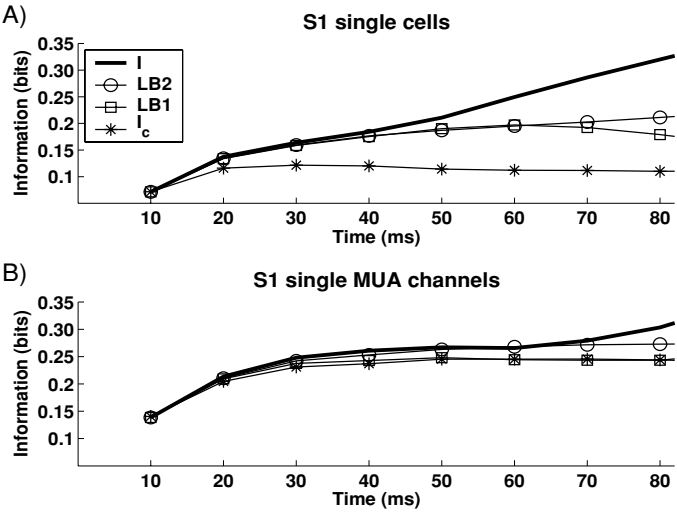


Figure 6: Lower-bounds I_{LB1} and I_{LB2} to estimate the spike timing information in rat somatosensory cortex. We report the information analysis performed on (A) 10 single units and (B) 7 single MUA spike trains. In both cases, the spike trains were analyzed one at a time and then averaged across the population. The total spike timing information I, I_{LB1}, I_{LB2}, and spike count information I_c is plotted.

were well sampled, indicating that I_{LB2} is a tight estimator. The use of I_{LB2} demonstrated that spike timing conveys information above and beyond that carried by spike counts: at 0 to 80 ms poststimulus, the spike timing information computed with I_{LB2} was 90% higher than the spike count information. I_{LB1} was close to I_{LB2} up to 60 ms poststimulus and then dropped by 30% at 80 ms, consistent with the simulation predictions that I_{LB1} is less data robust than I_{LB2} and that it tends to be downward biased in conditions of undersampling (see Figure 4).

The time course of the estimates of the information transmitted about stimulus location carried by spike times recorded from a single MUA channel are reported in Figure 6B. Seven single channels were analyzed separately and then averaged. With respect to the single-unit case above, the tenfold increase in the number of trials improved the sampling of all information estimates: $I(\mathcal{R};\mathcal{S})$ now increased smoothly up to 60 ms, and both I_{LB1} and I_{LB2} varied smoothly with time in the whole 0 to 80 ms window considered. This is consistent with the simulation results in Figure 4. I_{LB2} was close to the total information in the range 0 to 60 ms, suggesting that I_{LB2} was tight. Despite the good sampling, I_{LB1} remained very close to the spike count information and was unable to reveal the presence of any extra

spike timing information. At 0 to 80 ms, I_{LB2} was 28% higher than both the spike count information and I_{LB1}, thus demonstrating that spike timing conveys information above and beyond that carried by spike counts.

There are two interesting differences between single-unit and MUA results. For MUA, we found that (1) I_{LB1} was not tight and (2) I_{LB2} was still tight, but there was less extra spike timing information than for single cells. Both facts can be accounted for by the fact that MUA contains more positive correlations between spikes. The loss of tightness of I_{LB1} is expected from the addition of correlations between local neurons introduced by MUA (the fewer correlation sources there are, the tighter I_{LB1}). The loss of timing information is accounted for by the fact that 20% of the information carried by single S1 neurons is due to stimulus-indepedent negative autocorrelations (Panzeri, Petersen, et al., 2001) and that the addition of cross-correlation between nearby neurons introduces stimulus-independent positive cross-correlations that decrease the spike timing information considerably (Petersen et al., 2001).

Overall, these examples show that I_{LB2} can reliably reveal the presence of genuine spike timing information, even in cases when there would not be enough data to compute the full spike timing information and the use of I_{LB1} would fail to reveal it. They also show that the performance of the estimators on real data sets with different characteristics is consistent with the analytical and numerical results derived in previous sections.

To study how the performance of the estimators varies when increasing the population size, we next considered pairs of S1 MUA recording channels. We analyzed only pairs located in the same barrel column recorded with silicon electrodes spaced by $\approx$0.4 mm. Their activity is known to be cross-correlated with short time lag (Lebedev et al., 2000; Petersen & Diamond, 2000). In Figure 7, we report the time course of the information analysis performed on the three same-column S1 pairs available (results averaged across pairs). We increased the poststimulus time windows in steps of 10 ms from 0 to 80 ms, and we used 10 ms time bins to digitize the spike times. Since we also wanted to investigate the role of short-time-range stimulus modulations of cross-channel correlations, we considered I^0_{LB3} alongside I_{LB1} and I_{LB2}. According to the numerical and analytical considerations above, both I_{LB2} and I^0_{LB3} were well sampled in the time range considered. Consistent with this prediction, they behaved smoothly as a function of time. I_{LB1} was significantly smaller than both I_{LB2} and I^0_{LB3}, and it decreased dramatically after 50 ms. I_{LB1} performed worse for longer windows when analyzing pairs than when analyzing single channels. This is because I_{LB1} was less data robust than the other two estimators. In particular, after five time bins ($LC = 10$), the response entropy in I_{LB1} (see Figure 4) gets strongly downward biased, giving rise to the pattern in the figure. I^0_{LB3} was consistently higher than I_{LB2} (it was 9% higher at 0–80 ms poststimulus). Since I^0_{LB3} considers only the effects of stimulus modulation of cross-correlations within the same time bin, these results show that stimulus modulations

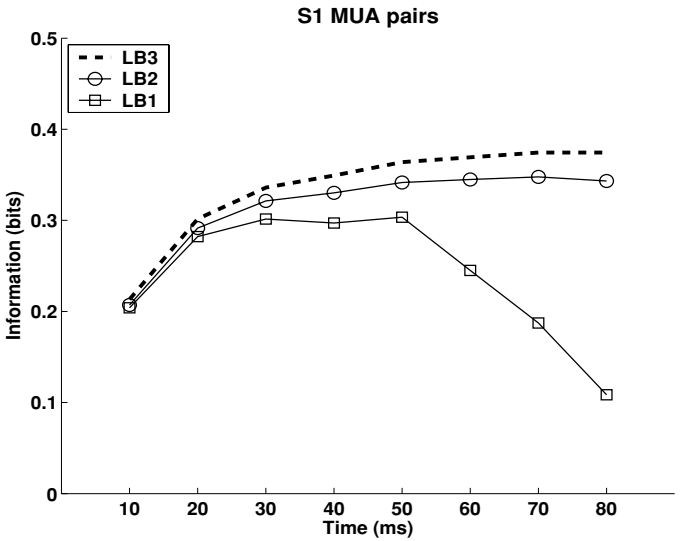


Figure 7: Lower-bounds I_{LB1}, I_{LB2}, and I_{LB3} to estimate the spike timing information conveyed by the MUA activity recorded from paired electrodes in rat somatosensory cortex. We report the information analysis performed on pairs of MUA spike trains recorded simultaneously from two different electrodes (both located in the same barrel column; see main text). We analyzed three such paired spike trains; they were analyzed one at a time and then averaged across the population. Data were averaged across the population. The bias-corrected values of I_{LB1}, I_{LB2}, and I^0_{LB3} are plotted.

of short-lag correlations contribute to information transmission. This was consistent with the finding that the strength of cross-correlation was weakly modulated by the stimulus: the Pearson cross-correlation coefficient of the joint spike counts (computed in the same 10 ms long sliding windows as for the information, and then averaged across all windows and cell pairs) was slightly smaller in response to stimulation of the principal whisker than in response to other whiskers (0.04 and 0.06, respectively). The contribution of stimulus modulation of correlations in this MUA pair data set was higher than in the corresponding analysis of pairs of single units (reported by Petersen et al., 2001). One possible explanation is that MUA pairs effectively sample a larger population and that correlations play a more important contribution for larger populations.

In the case of MUA pairs, computation of the full mutual information was possible only for the first two or three time bins, after which the estimate diverged dramatically (data not shown). This illustrates that the lower-bound techniques developed here significantly extend the time range over

which the spike timing information carried by neuronal populations can be analyzed. When analyzing pairs, our new approach allowed us to estimate the spike timing information with the 500 trials per stimulus, as opposed to the hundreds of thousands of trials that would have been required to estimate the total spike timing information.

8.2 Spike Timing and Coding of Motion Direction in the MT Visual Cortex of the Awake Macaque. In this section, we apply our new analysis to multiunit recordings collected from the MT visual area of a behaving monkey, and we show how our techniques could be successfully used to investigate whether neurons (such as those in MT) encode information about stimuli (such as motion direction) by means of spike timing.

The MT neuronal responses analyzed in this section were recorded as follows. MUA was recorded through electrodes placed in area MT of a macaque monkey. MUA was subjected to thresholding to discriminate spike times. Although in general it was not possible to isolate spikes emitted by individual cells, in a few cases we were confident that there was only one single unit in the MUA (on the basis of standard clustering and autocorrelogram analysis criteria). The monkey was trained on a direction discrimination task and, during recording of neuronal activity, was fixating a screen centrally (fixation window ± 0.5 degree). A structured background was back-projected onto the screen. After a randomized period of time, a grating was projected onto the screen moving in one of four possible directions along the cardinal axes, positioned over the receptive fields of the neurons under study. Luminance contrast (i.e., visibility) of the stimulus was randomized (0%, 2%, 4%, 17%). The monkey performed a reaction time task and indicated the perceived direction of motion by a hand movement to one of four touch bars located in front of the chest. Neural data analyzed in this article were from high-luminance stimulus conditions. (For more detailed information, see Thiele, Distler, & Hoffmann, 1999). Forty to 50 trials per stimulus were available. In this example analysis, we considered two different MUA recordings from MT (see Figures 8 and 9), which illustrate different ways in which information about motion direction might be encoded by MT neurons.

An analysis of the first example of MUA single-channel recording is shown in Figure 8. The poststimulus time histograms (PSTHs) to different stimulus conditions (see Figure 8A) showed that neuronal activity from this electrode was strongly modulated by motion direction. In particular, responses to up and down motion directions were particularly strong, with large response peaks between 50 and 100 ms poststimulus. Although the peaks to up and down motion were of similar magnitude ($\approx$150 Hz), there was a latency difference between responses to up and down motion (80 ms versus 70 ms, respectively). Responses to left and right motions were smaller in magnitude and occurred with longer latencies. These stimulus-related latency differences suggest that spike timing may convey important

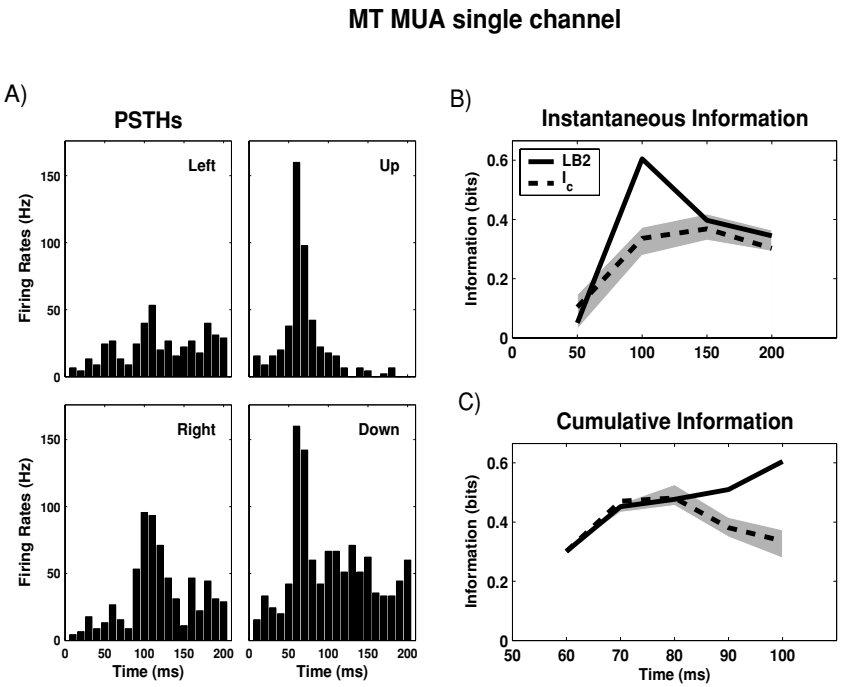


Figure 8: Information about motion direction transmitted by MUA in MT visual cortex. We quantified the information that spike times and spike counts convey about motion direction by applying the information analysis MUA recorded from one electrode in MT of an awake behaving monkey. Information conveyed by spike timing (computed with our I_{LB2} lower bound) is compared to the information in the spike count. The recording site considered here presents strong extra spike timing information not available in the spike count. The gray areas represent a (bootstrap-computed) confidence band for the spike count information: if the spike timing information lies in this area, it is likely (with a probability of 0.95) that the spike timing information equals the spike count information; hence, there is not extra information in the timing of spikes not provided in the count. (A) PSTHs of the neuron. (B) spike timing (I_{LB2}) and spike count information in sliding windows from 0 to 200 ms, increased in steps of 50 ms. (C) Spike timing (I_{LB2}) and spike count information in cumulative windows from 50 to 100 ms, increased in steps of 10 ms.

information about motion direction in this case. We investigated this hypothesis by using the lower-bound estimate I_{LB2}, which, on the basis of the $\approx$50 trials per stimulus available, was well enough sampled up to 12 time bins, and compared it to the information conveyed by spike counts. We first examined the information transmitted in sliding windows of 50 ms

ranging from 0 to 200 ms.[6] We found (see Figure 8B) that in the 50 to 100 ms time interval, there was twice as much information in spike times as in spike counts (0.6 bits versus 0.3 bits). This difference in information was highly significant, as the spike timing information I_{LB2} was far above the ($P < 0.05$) bootstrap-computed confidence interval of the spike count information (the gray area in Figure 8B). The time course of the information conveyed by the spikes in the 50 to 100 ms time window was magnified in Figure 8C, where we report the cumulative plot of information in the time windows [50,60] ms, [50,70] ms, [50,80] ms, [50,90] ms, and [50,100] ms. There was rapidly increasing extra information in spike timing after 80 ms poststimulus (the time when an observer of neuronal activity could use the latency difference between up and down motion direction responses to discriminate the stimulus).

In Figure 9 we considered a second example of MUA activity. In this case, from PSTH inspection, it is likely that the MUA channel contains a smaller neuronal population than that in the previous example. Also in this case, PSTHs to different motion directions (see Figure 9A) showed that neuronal activity was strongly modulated by motion direction. Responses to left and right motion directions were particularly strong. However, responses were more tonic than in the previous case, and there was no marked latency difference between the motion directions eliciting stronger response. Thus, in this case, we expected that spike times did not add much information about motion direction to that provided by spike counts alone. Results of the information analysis using our new lower bound (see Figures 9B and 9C) confirmed this expectation: the lower-bound analysis could not find any evidence that knowledge of spike times further increased the information provided by spike counts.

In both examples, we could not estimate the full information reliably out of 50 trials per stimulus for time windows as long as that analyzed here.

This application is useful in showing that the new lower bounds reliably and consistently pick up spike timing information originating by stimulus-related differences in the temporal shape of PSTHs, and they can achieve this by using the number of data that can be collected from a behaving animal. This shows that our bounds could become a useful tool to probe the importance of spike timing in neural coding in awake behaving animals, when it is usually not possible to record responses to hundreds of repetitions of the same stimulus.

We would like to stress that we report this example only as a demonstration of the applicability of the new method. We do not draw from it general conclusions about the role of spike timing in coding of visual information

[6] This means that we computed the information estimations in four time intervals: [0,50] ms, [50,100] ms, [100,150] ms, and [150,200] ms. The information in each of these four time periods was computed after digitizing the spike trains with 10 ms precision.

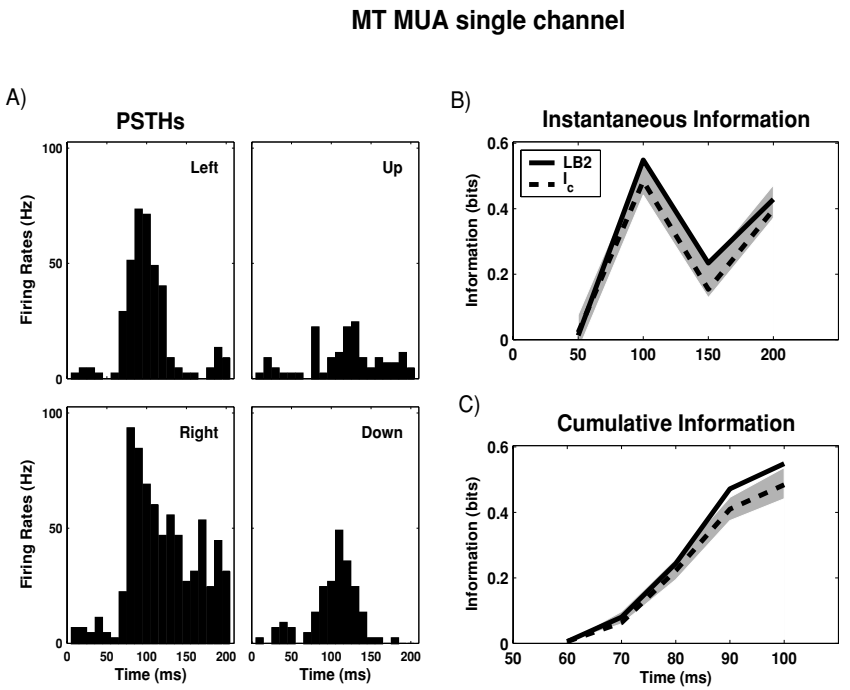


Figure 9: Information about motion direction transmitted by MUA in MT visual cortex. In this second example, this recording site presented here does not convey any extra information by spike timing. Conventions are as in Figure 8.

in area MT, which still has to be determined. In particular, the use of MUA recording rather than single units may affect the spike timing results (as shown in the previous section). For examples the MUA channel in Figure 8 may carry significant spike timing information because it contains more than one cell with different latencies. In general, we found that in all but a few cases analyzed so far in which a single unit could be reliably discriminated, there was very little extra information carried by spike timing. For the present purpose, we preferred to report only the two above examples of MUA activity in MT, because those two examples show very clearly the relation between stimulus-related PSTH temporal properties and I_{LB2} spike timing information.

9 Discussion

Multiple electrodes are a widely used tool in neuroscience research that makes it possible to study the simultaneous activity of neuronal populations (Brown, Kass, & Mitra, 2004). The information-theoretic analysis of

such simultaneously recorded neuronal activity offers a principled way to study how spike timing and correlation contribute to neuronal population coding of sensory information (Nirenberg & Latham, 2003; Pola et al., 2003; Schneidman et al., 2003; Averbeck & Lee, 2004). However, despite the fact that information theory has had widespread use in single-neuron analysis (Optican & Richmond, 1987; Rieke et al., 1996; Borst & Theunissen, 1999), relatively few studies have used it for analysis of population spike trains (e.g., Petersen et al., 2001; Nirenberg et al., 2001; Rolls et al., 2003). The main reason for this has been the unpractically large numbers of data that are often required for information-theoretic analysis of populations (Brown et al., 2004). This article presents several advances that can help in alleviating this problem and extending the range of applicability of information theory to multi-spike-train analysis.

First, this study alleviates the limited sampling problem by providing data-robust quantities that approximate precisely the information under very general conditions. This new approach complements other recent advances on the sampling problem (Victor, 2002; Paninski, 2003; Nemenman, Bialek, & de Ruyter van Steveninck, 2004).

Second, the estimators introduced here approximate the mutual information from below: this means that any such estimate of information contained by spike timing does not contain any spurious information due to sampling artifacts. Thus, a demonstration, obtained with these methods, that population spike timing conveys substantial information not available in spike counts would be very robust and statistically significant. The examples reported in Figures 8 and 9 show that this is possible even with the numbers of data recoded from awake-behaving animals. Thus, the method opens up the possibility of quantitative investigations of the role of spike timing in some cognitive and perceptual tasks.

Third, the information estimators developed here lend themselves to investigations of the role of correlated firing in coding. In fact, all estimators I_{LB2} and I^q_{LB3} can take into account the effect of stimulus-independent correlations. Moreover, the estimators I^q_{LB3} also take into account the effect of stimulus modulations of correlations between spikes that are separated by q time steps or fewer. By varying q within the range allowed by the data size available in a particular experiment, one could use the I^q_{LB3} estimates to obtain a quantitative characterization of the timescales over which correlations contribute to population coding. This approach will succeed in determining the timescales over which correlations carry information if all the informative stimulus-dependent correlations are short-time-ranged. However, a practical problem is that determining with our approach the presence or absence of long time-range correlations may require large numbers of data. Thus, when working with typical neurophysiological data sets, it is useful to complement the q-time-steps analysis presented here with a rigorous assessment of long-range correlations based on other statistical methods (e.g., Oram et al., 2001).

It is important to note that the finding that the new estimators I_{LB2} and I^q_{LB3} are less biased and more data robust than the mutual information is general and due to the intrinsic properties of these functionals. In fact, these estimators depend on (1) entropies that are "lower dimensional" than the full spike timing response entropy and are thus more data robust, and (2) other nonentropy quantities (such as χ) that are intrinsically less biased than entropies. However, the actual performance of each estimator may depend on the particular method used to remove the sampling bias. Here we corrected for the bias using an analytical approximation based on the assumption that the number of trials N used to compute the probabilities was bigger than the number of possible responses R (Panzeri & Treves, 1996). Although the bias subtraction method presented here performs extremely well for nonentropy quantities (such as χ) and relatively well for the entropy quantities, it is possible that estimating entropy quantities on which I_{LB2} and I^q_{LB3} depend by using recent advances on the entropy sampling problem (Paninski, 2003; Nemenman et al., 2004; Victor, 2002) may push the performance of the approach presented here much further. We are currently investigating in a systematic way, by means of computer simulations, how various bias elimination methods perform when applied to the probability functionals developed and studied here (Panzeri, 2005).

Appendix: Bias Expressions of $\chi(\mathcal{R})$, $\chi^0(\mathcal{R})$, and $H^q(\mathcal{R}|\mathcal{S})$

This appendix reports explicit expressions for the bias approximations of the quantities $H^q(\mathcal{R}|\mathcal{S})$, $\chi(\mathcal{R})$, and $\chi^0(\mathcal{R})$ that were not reported in the main text. These approximations to the bias can be subtracted from the estimates of the functional obtained from limited sampled probabilities to correct for the sampling problem.

The bias of a given functional of the probability distributions is defined as the difference between the trial-averaged value of the functional when the probability distributions are computed from N trials only and the value of the functional computed with the true probability distributions (obtained from an infinite number of observations). There are several ways to derive the bias correction: we have followed a simple procedure equivalent to that used and detailed in appendix B of Pola et al. (2003). In brief, we used a Taylor series expansion of the functional around the true probability value and then averaged over all possible outcomes of the N trials. We considered only the first two terms in the expansion (corresponding to the effect of mean and variance of the estimates of the probabilities obtained with N trials). This corresponds to computing the bias to first order in $\frac{1}{N}$ (see Pola et al., 2003). Thus, all bias equations reported in this appendix are valid to the $\frac{1}{N}$ order. This approximation is good if there are enough experimental trials N so that fluctuations of the estimated probability distributions around the asymptotic value are small. In this appendix, we report schematically the

derivation of only the bias expressions that were not derived in Pola et al. (2003).

A.1 Bias of $H^q(\mathcal{R}|\mathcal{S})$. The bias for $H^q(\mathcal{R}|\mathcal{S})$ $q = 1, \ldots, L-1$ can be calculated by expressing it as a sum of lower-dimensional noise entropies:

$$H^q(\mathcal{R}|\mathcal{S}) = H(\mathcal{R}_1, \ldots, \mathcal{R}_q|\mathcal{S}) + \\ + \sum_{t=q+1}^{L} [H(\mathcal{R}_{t-q}, \ldots, \mathcal{R}_t|\mathcal{S}) - H(\mathcal{R}_{t-q}, \ldots, \mathcal{R}_{t-1}|\mathcal{S})], \quad \text{(A.1)}$$

where, in the above, $H(\mathcal{R}_1, \ldots, \mathcal{R}_q|\mathcal{S})$ is the noise entropy of the marginal probabilities $P(\mathbf{r}(1), \ldots, \mathbf{r}(q)|s)$. By applying equation. 6.1 (whose derivation is reported in Panzeri & Treves, 1996, and Pola et al., 2003) to all noise entropies in the above, we obtain:

$$Bias[H^q(\mathcal{R}|\mathcal{S})] \approx -\frac{1}{2N\log 2}\sum_s(\tilde{R}_{1,\ldots,q}(s) - 1) \\ - \frac{1}{2N\log 2}\sum_{t=q+1}^{L}\sum_s[\tilde{R}_{t-q,\ldots,t}(s) - \tilde{R}_{t-q,\ldots,t-1}(s)], \quad \text{(A.2)}$$

where $\tilde{R}_{1,\ldots,q}(s)$, $\tilde{R}_{t-q,\ldots,t}(s)$, and $\tilde{R}_{t-q,\ldots,t-1}(s)$ stand, respectively, for the number of relevant responses of the probability distributions $P(\mathbf{r}(1), \ldots, \mathbf{r}(q)|s)$, $P(\mathbf{r}(t-q), \ldots, \mathbf{r}(t)|s)$, and $P(\mathbf{r}(t-q), \ldots, \mathbf{r}(t-1)|s)$. As for the case of the full probability distribution $P(\mathbf{r}|s)$, the determination of the number of the number of relevant bins of these marginal probabilities may not be straightforward when data are scarce. As discussed in the main text, an approach to this problem was presented by Panzeri and Treves (1996).

The bias of $H^0(\mathcal{R}|\mathcal{S})$ can be derived in an analogous way and has the following simpler expression:

$$Bias[H^0(\mathcal{R}|\mathcal{S})] \approx -\frac{1}{2N\ln 2}\sum_s\sum_t(\tilde{R}_t(s) - 1), \quad \text{(A.3)}$$

where $\tilde{R}_t(s)$ is the number of relevant responses of the marginal distributions $P(\mathbf{r}(t)|s)$.

A.2 Bias of $\chi(\mathcal{R})$. The derivation of the bias of $\chi(\mathcal{R})$ follows almost exactly the one reported in appendix B of Pola et al. (2003) and is very

similar to the derivation of the bias of $\chi^0(\mathcal{R})$ (reported below). Thus, for conciseness, here we report only the full result:

$$Bias[\chi(\mathcal{R})] \approx \frac{\Lambda + \Gamma LC + \Theta L^2 C^2}{2N \ln 2}, \tag{A.4}$$

The coefficients Λ, Γ, and Θ are functionals of the stimulus conditional probability distributions $P(\mathbf{r}|s)$, the marginal distributions $P(r_c(t)|s)$, and P(s). Their values are given by

$$\begin{aligned}\Lambda = 1 &- \widehat{\sum_{\mathbf{r}}} \frac{P(\mathbf{r})}{P_{ind}(\mathbf{r})} \sum_s P_{ind}(\mathbf{r}|s)\beta(\mathbf{r}|s) \\ &+ \widehat{\sum_{\mathbf{r}}} \frac{P(\mathbf{r})}{P^2_{ind}(\mathbf{r})} \left\langle \left(1 + \frac{\alpha(\mathbf{r}|s)}{P_{ind}(\mathbf{r}|s)} + \beta(\mathbf{r}|s)\right) P^2_{ind}(\mathbf{r}|s) \right\rangle_s \\ &- \widehat{\sum_{\mathbf{r}}} \frac{2}{P_{ind}(\mathbf{r})} \left\langle P(\mathbf{r}|s)[P_{ind}(\mathbf{r}|s) + \alpha(\mathbf{r}|s)] \right\rangle_s,\end{aligned} \tag{A.5}$$

$$\Gamma = \widehat{\sum_{\mathbf{r}}} \frac{1}{P_{ind}(\mathbf{r})} \sum_s P_{ind}(\mathbf{r}|s)[2P(s)P(\mathbf{r}|s) - P(\mathbf{r})], \tag{A.6}$$

$$\Theta = \widehat{\sum_{\mathbf{r}}} \frac{P(\mathbf{r})}{P^2_{ind}(\mathbf{r})} \sum_s P_{ind}(\mathbf{r}|s)[P_{ind}(\mathbf{r}) - P(s)P_{ind}(\mathbf{r}|s)], \tag{A.7}$$

where $\widehat{\sum_{\mathbf{r}}}$ is a summation restricted to the response variables $\mathbf{r}$ such that $P_{ind}(\mathbf{r}) \neq 0$. $\alpha(\mathbf{r}|s)$ and $\beta(\mathbf{r}|s)$ are defined as follows,

$$\alpha(\mathbf{r}|s) = \sum_{c,t_c} \frac{P_{ind}(\mathbf{r}|s)}{P(r_c(t_c)|s)}, \tag{A.8}$$

$$\beta(\mathbf{r}|s) = \sum_{(b,t_b)\neq(c,t_c)} \frac{P(r_b(t_b), r_c(t_c)|s)}{P(r_b(t_b)|s)P(r_c(t_c)|s)}, \tag{A.9}$$

where in the last equation, we sum up over every b, t_b, c, and t_c such that the couple of values (b, t_b) is different from (c, t_c). It is worth stressing that both $\alpha(\mathbf{r}|s)$ and $\beta(\mathbf{r}|s)$ are regular and finite quantities.

The bias coefficients Λ, Γ, and Θ are functionals of the full probability distributions. Thus, an interesting question is how to compute these terms from raw data. In this letter, we have performed this evaluation by plugging the empirically obtained probabilities into the above expression for Λ, Γ, and Θ. Although this may potentially introduce systematic inaccuracies in the determination of these terms, the numerical simulations presented in Figures 2 and 3 show that this problem is negligible and that a

very accurate estimation of the bias of $\chi(\mathcal{R})$ can in general be reached even with small data sets.

A.3 Bias of $\chi^0(\mathcal{R})$. We start the derivation with a reminder that $\chi^0(\mathcal{R})$ can be expressed as

$$\chi^0(\mathcal{R}) = -\widehat{\sum_{\mathbf{r}}} P(\mathbf{r}) \log_2 \tilde{P}_0(\mathbf{r}), \tag{A.10}$$

where the notation $\widehat{\sum}_{\mathbf{r}}$ reminds that the summation over responses is restricted to $\mathbf{r}$ such that $\tilde{P}_0(\mathbf{r}) \neq 0$. The first step to compute the bias of $\chi^0(\mathcal{R})$ is to perform a second-order series expansion of $\chi^0(\mathcal{R})$ around the true probability distributions,

$$\begin{aligned}
Bias[\chi^0(\mathcal{R})] \approx & \frac{1}{2} \sum_s \frac{\delta^2 \chi^0}{\delta P(s)^2} \sigma_N^2[P(s), P(s)] \\
& + \frac{1}{2} \sum_{s,s',s\neq s'} \frac{\delta^2 \chi^0}{\delta P(s) \delta P(s')} \sigma_N^2[P(s), P(s')] \\
& + \frac{1}{2} \sum_s \sum_t \sum_{\tilde{\mathbf{r}}(t)} \frac{\delta^2 \chi^0}{\delta P(\tilde{\mathbf{r}}(t)|s)^2} \sigma_N^2[P(\tilde{\mathbf{r}}(t)|s), P(\tilde{\mathbf{r}}(t)|s)] \\
& + \frac{1}{2} \sum_s \sum_{t,t',t\neq t'} \sum_{\tilde{\mathbf{r}}(t),\tilde{\mathbf{r}}(t')} \frac{\delta^2 \chi^0}{\delta P(\tilde{\mathbf{r}}(t)|s) \delta P(\tilde{\mathbf{r}}(t')|s)} \\
& \times \sigma_N^2[P(\tilde{\mathbf{r}}(t)|s), P(\tilde{\mathbf{r}}(t')|s)] \\
& + \sum_s \sum_t \sum_{\tilde{\mathbf{r}}(t)} \sum_{\hat{\mathbf{r}}} \frac{\delta^2 \chi^0}{\delta P(\tilde{\mathbf{r}}(t)|s) \delta P(\hat{\mathbf{r}}|s)} \sigma_N^2[P(\tilde{\mathbf{r}}(t)|s), P(\hat{\mathbf{r}}|s)] \\
& + o\left(\frac{1}{N}\right),
\end{aligned} \tag{A.11}$$

where $\frac{\delta \chi^0}{\delta P}$ stands for the functional derivative of $\chi^0(\mathcal{R})$ with respect to response probability distributions (computed in the true asymptotic value obtained with an infinite amount of data). $\sigma_N^2[\cdot,\cdot]$ are the variances and covariances of the probability distributions. We introduced $\tilde{\mathbf{r}}(t)$ and $\hat{\mathbf{r}}$ to distinguish them from $\mathbf{r}$, which is a running variable used to define $\chi^0(\mathcal{R})$ (see equation 7.10). While $\tilde{\mathbf{r}}(t)$ corresponds to the neuronal population response in the tth time bin only, $\hat{\mathbf{r}}$ stands for the neuronal response in all the time bins. Computing the functional derivatives explicitly, and omitting for brevity all the terms that simplify away, we obtain the following expression for the leading term of the bias of $\chi^0(\mathcal{R})$:

$$
\begin{aligned}
Bias[\chi^0(\mathcal{R})] = {} & \frac{1}{2\ln 2}\sum_s \widehat{\sum_{\mathbf{r}}} \frac{\tilde{P}_0(\mathbf{r}|s)}{\tilde{P}_0(\mathbf{r})}\left(\frac{P(\mathbf{r})}{\tilde{P}_0(\mathbf{r})}\tilde{P}_0(\mathbf{r}|s) - 2P(\mathbf{r}|s)\right) \\
& \times \sigma_N^2[P(s),\, P(s)] \\
& + \frac{1}{2\ln 2}\sum_{s,s',s\neq s'} \widehat{\sum_{\mathbf{r}}}\left(\frac{P(\mathbf{r})}{\tilde{P}_0(\mathbf{r})^2}\tilde{P}_0(\mathbf{r}|s)\tilde{P}_0(\mathbf{r}|s')\right. \\
& \left. - \frac{P(\mathbf{r}|s)\tilde{P}_0(\mathbf{r}|s') + \tilde{P}_0(\mathbf{r}|s)P(\mathbf{r}|s')}{\tilde{P}_0(\mathbf{r})}\right)\sigma_N^2[P(s),\, P(s')] \\
& + \frac{1}{2\ln 2}\sum_s \sum_t \sum_{\tilde{\mathbf{r}}(t)} \widehat{\sum_{\mathbf{r}}} P^2(s)\frac{P(\mathbf{r})}{\tilde{P}_0(\mathbf{r})^2}\left(\frac{\tilde{P}_0(\mathbf{r}|s)}{P(\tilde{\mathbf{r}}(t)|s)}\right)^2 \\
& \times \delta_{[\mathbf{r}(t),\tilde{\mathbf{r}}(t)]}\sigma_N^2[P(\tilde{\mathbf{r}}(t)|s),\, P(\tilde{\mathbf{r}}(t)|s)] \\
& + \frac{1}{2\ln 2}\sum_s \sum_{t,t',t\neq t'} \sum_{\tilde{\mathbf{r}}(t),\tilde{\mathbf{r}}(t')} \widehat{\sum_{\mathbf{r}}} P(s) \\
& \times \delta_{[\mathbf{r}(t),\tilde{\mathbf{r}}(t)]}\delta_{[\mathbf{r}(t'),\tilde{\mathbf{r}}(t')]}\frac{\tilde{P}_0(\mathbf{r}|s)}{P(\tilde{\mathbf{r}}(t)|s)P(\tilde{\mathbf{r}}(t')|s)} \\
& \times \frac{P(\mathbf{r})}{\tilde{P}_0(\mathbf{r})}\left(\frac{P(s)\tilde{P}_0(\mathbf{r}|s)}{\tilde{P}_0(\mathbf{r})} - 1\right)\sigma_N^2[P(\tilde{\mathbf{r}}(t)|s),\, P(\tilde{\mathbf{r}}(t')|s)] \\
& - \frac{1}{\ln 2}\sum_s \sum_t \sum_{\tilde{\mathbf{r}}(t)} \widehat{\sum_{\mathbf{r}}} P^2(s)\delta_{[\mathbf{r}(t),\tilde{\mathbf{r}}(t)]}\frac{\tilde{P}_0(\mathbf{r}|s)}{P(\tilde{\mathbf{r}}(t)|s)\tilde{P}_0(\mathbf{r})} \\
& \times \sigma_N^2[P(\tilde{\mathbf{r}}(t)|s),\, P(\mathbf{r}|s)] + o\left(\frac{1}{N}\right), \qquad \text{(A.12)}
\end{aligned}
$$

where $\delta_{[\mathbf{r}(t),\tilde{\mathbf{r}}(t)]}$ is a Kronecker delta (i.e., $\delta_{[\mathbf{r}(t),\tilde{\mathbf{r}}(t)]} = 1$ if $\mathbf{r}(t) = \tilde{\mathbf{r}}(t)$ and $\delta_{[\mathbf{r}(t),\tilde{\mathbf{r}}(t)]} = 0$ if $\mathbf{r}(t) \neq \tilde{\mathbf{r}}(t)$). The values of the variances and covariances are as follows:

$$\sigma_N^2[P(s),\, P(s)] = \frac{P(s)(1-P(s))}{N} + o\left(\frac{1}{N}\right), \qquad \text{(A.13)}$$

$$\sigma_N^2[P(s),\, P(s')] = -\frac{P(s)P(s')}{N} + o\left(\frac{1}{N}\right), \qquad \text{(A.14)}$$

$$\sigma_N^2[P(\tilde{\mathbf{r}}(t)|s),\, P(\tilde{\mathbf{r}}(t)|s)] = \frac{P(\tilde{\mathbf{r}}(t)|s)(1-P(\tilde{\mathbf{r}}(t)|s))}{N_s} + o\left(\frac{1}{N_s}\right), \qquad \text{(A.15)}$$

$$
\begin{aligned}
\sigma_N^2[P(\tilde{\mathbf{r}}(t)|s),\, P(\tilde{\mathbf{r}}(t')|s)] = {} & -\frac{P(\tilde{\mathbf{r}}(t)|s)P(\tilde{\mathbf{r}}(t')|s)}{N_s} + \frac{P(\tilde{\mathbf{r}}(t),\tilde{\mathbf{r}}(t')|s)}{N_s} \\
& + o\left(\frac{1}{N_s}\right), \qquad \text{(A.16)}
\end{aligned}
$$

$$\sigma_N^2[P(\tilde{\mathbf{r}}(t)|s),\, P(\mathbf{r}|s)] = -\frac{P(\tilde{\mathbf{r}}(t)|s)P(\mathbf{r}|s)}{N_s} + \delta_{[\mathbf{r}(t),\tilde{\mathbf{r}}(t)]}\frac{P(\mathbf{r}|s)}{N_s} + o\left(\frac{1}{N_s}\right). \tag{A.17}$$

After replacing the explicit values of variances and covariances in equation A.12, and after performing some algebra, we obtain the following final expression for the bias of $\chi^0(\mathcal{R})$:

$$Bias[\chi^0(\mathcal{R})] \approx \frac{\Lambda^0 + L\Gamma^0 + L^2\Theta^0}{2N\ln 2}. \tag{A.18}$$

The coefficients Λ^0, Γ^0, and Θ^0 are functionals of the stimulus conditional probability distributions $P(\mathbf{r}|s)$ and of the marginal distributions $P(\mathbf{r}(t)|s)$, and $P(s)$. Their values are given by

$$\begin{aligned}\Lambda^0 = 1 &- \widehat{\sum_{\mathbf{r}}}\frac{P(\mathbf{r})}{\tilde{P}_0(\mathbf{r})}\sum_s \tilde{P}_0(\mathbf{r}|s)\beta^0(\mathbf{r}|s) \\ &+ \widehat{\sum_{\mathbf{r}}}\frac{P(\mathbf{r})}{\tilde{P}_0(\mathbf{r})^2}\sum_s P(s)\left(1 + \frac{\alpha^0(\mathbf{r}|s)}{\tilde{P}_0(\mathbf{r}|s)} + \beta^0(\mathbf{r}|s)\right)\tilde{P}_0(\mathbf{r}|s)^2 \\ &- \widehat{\sum_{\mathbf{r}}}\frac{2}{\tilde{P}_0(\mathbf{r})}\sum_s P(s)P(\mathbf{r}|s)[\tilde{P}_0(\mathbf{r}|s) + \alpha^0(\mathbf{r}|s)],\end{aligned} \tag{A.19}$$

$$\Gamma^0 = \widehat{\sum_{\mathbf{r}}}\frac{1}{\tilde{P}_0(\mathbf{r})}\sum_s \tilde{P}_0(\mathbf{r}|s)[2P(s)P(\mathbf{r}|s) - P(\mathbf{r})], \tag{A.20}$$

$$\Theta^0 = \widehat{\sum_{\mathbf{r}}}\frac{P(\mathbf{r})}{\tilde{P}_0(\mathbf{r})^2}\sum_s \tilde{P}_0(\mathbf{r}|s)[\tilde{P}_0(\mathbf{r}) - P(s)\tilde{P}_0(\mathbf{r}|s)], \tag{A.21}$$

where $\widehat{\sum_{\mathbf{r}}}$ is a summation restricted to the response variables $\mathbf{r}$ such that $\tilde{P}_0(\mathbf{r}) \neq 0$. $\alpha^0(\mathbf{r}|s)$ and $\beta^0(\mathbf{r}|s)$ are defined as follows:

$$\alpha^0(\mathbf{r}|s) = \sum_t \frac{\tilde{P}_0(\mathbf{r}|s)}{P(\mathbf{r}(t)|s)}, \tag{A.22}$$

$$\beta^0(\mathbf{r}|s) = \sum_{t,t',t\neq t'} \frac{P(\mathbf{r}(t), \mathbf{r}(t')|s)}{P(\mathbf{r}(t)|s)P(\mathbf{r}(t')|s)}. \tag{A.23}$$

Acknowledgments

We are grateful to M. E. Diamond and M. Lebedev for kindly making available to us the example data used in Figure 6A and to M. E. Diamond,

P. Latham, M. A. Montemurro and S. R. Schultz for many useful discussions. This research was supported by an MRC Research Fellowship (S.P.), Wellcome Trust 066372/Z/01/Z and GR070380, Royal Society and DFG SFB 509.

References

Abbott, L. F., & Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural Comp.*, *11*, 91–101.

Abeles, M., Bergman, H., Margalit, E., & Vaadia, E. (1993). Spatio-temporal firing patterns in the frontal cortex of behaving monkeys. *J. Neurophysiol.*, *70*, 1629–1638.

Adrian, E. D. (1926). The impulses produced by sensory nerve endings: Part I. *J. Physiol. (Lond.)*, *61*, 49–72.

Averbeck, B. B., & Lee, D. (2004). Coding and transmission of information by neural ensembles. *Trends in Neurosciences*, *27*, 225–230.

Borst, A., & Theunissen, F. E. (1999). Information theory and neural coding. *Nature Neuroscience*, *2*, 947–957.

Brosch, M., Bauer, R., & Eckhorn, R. (1997). Stimulus dependent modulations of correlated high frequency oscillations in cat visual cortex. *Cerebral Cortex*, *7*, 70–76.

Brown, E. N., Kass, R. E., & Mitra, P. P. (2004). Multiple neural spike train data analysis: State-of-the-art and future challenges. *Nature Neuroscience*, *7*, 456–461.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Dan, Y., Alonso, J.-M., Usrey, W. M., & Reid, R. C. (1998). Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus. *Nature Neuroscience*, *1*, 501–507.

de Oliveira, S. C., Thiele, A., & Hoffman, K.-P. (1997). Synchronization of neuronal activity during stimulus expectation in a direction discrimination task. *J. Neurosci.*, *17*, 9248–9260.

DeWeese, M. R., Wehr, M., & Zador, A. M. (2003). Binary spiking in auditory cortex. *J. Neurosci.*, *23*, 7940–7949.

Dimitrov, A. G., & Miller, J. P. (2001). Neural coding and decoding: Communication channels and quantization. *Network: Comput. Neural Syst.*, *12*, 441–472.

Furukawa, S., Xu, L., & Middlebrooks, J. C. (2000). Coding of sound-source location by ensembles of cortical neurons. *J. Neurosci.*, *20*, 1216–1228.

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

Golledge, H. D. R., Panzeri, S., Zheng, F., Pola, G., Scannell, J. W., Giannikopoulos, D. V., Mason, R. J., Tovee, M. J., & Young, M. P. (2003). Correlations, feature binding and population coding in primary visual cortex. *Neuroreport*, *14*, 1045–1050.

Gray, C. M., König, P., Engel, A. K., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, *338*, 334–337.

Hatsopoulos, N. G., Ojakangas, C. L., Paninski, L., & Donoghue, J. P. (1998). Information about movement direction obtained from synchronous activity of motor cortical neurons. *PNAS*, *95*(26), 15706–15711.

Jung, M. W., Qin, Y., Lee, D., & Mook-Jung, I. (2000). Relationships among discharges of neighboring neurons in the rat prefrontal cortex during spatial working memory tasks. *J. Neurosci.*, *20*, 6166–6172.

Lebedev, M. A., Mirabella, G., Erchova, I., & Diamond, M. E. (2000). Experience-dependent plasticity of rat barrel cortex: Redistribution of activity across barrel-columns. *Cerebral Cortex*, *10*, 23–31.

Mastronarde, D. N. (1983). Correlated firing of cat retinal ganglion cells. I. Spontaneously active inputs to x- and y-cells. *J. Neurophysiol.*, *49*, 303–324.

Miller, G. A. (1955). Note on the bias of information estimates. In H. Quastler (Ed.), *Information theory in psychology: Problems and methods* (vol. 2B, pp. 95–100). Glencoe, IL: Free Press.

Nemenman, I., Bialek, W., & de Ruyter van Steveninck, R. (2004). Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, *69*(5), 056111.

Nemenman, I., Shafee, F., & Bialek, W. (2002). Entropy and inference, revisited. In S. B. T. G. Dietterich & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14* (pp. 95–100). Cambridge, MA: MIT Press.

Nirenberg, S., Carcieri, S. M., Jacobs, A., & Latham, P. E. (2001). Retinal ganglion cells act largely as independent encoders. *Nature*, *411*, 698–701.

Nirenberg, S., & Latham, P. E. (1998). Population coding in the retina. *Current Opinion in Neurobiology*, *8*, 488–493.

Nirenberg, S., & Latham, P. E. (2003). Decoding neuronal spike trains: How important are correlations. *Proc. Natl. Acad. Sci. USA*, *100*, 7348–7353.

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

Oram, M. W., Földiák, P., Perrett, D. I., & Sengpiel, F. (1998). The "ideal homunculus": Decoding neural population signals. *Trends in Neurosciences*, *21*(6), 259–265.

Oram, M. W., Hatsopoulos, N., Richmond, B., & Donoghue, J. (2001). Excess synchrony in motor cortical neurons provides redundant direction information with that from coarse temporal measures. *J. Neurophysiol.*, *86*, 1700–1716.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, *15*, 1191–1253.

Panzeri, S. (2005). *A numerical evaluation of different approaches to measure spike timing information*. Manuscript in preparation.

Panzeri, S., Golledge, H. D. R., Zheng, F., Tovee, M. J., & Young, M. P. (2001). Objective assessment of the functional role of spike train correlations using information measures. *Visual Cognition*, *8*, 531–547.

Panzeri, S., Petersen, R. S., Schultz, S. R., Lebedev, M., & Diamond, M. E. (2001). The role of spike timing in the coding of stimulus location in rat somatosensory cortex. *Neuron*, *29*, 769–777.

Panzeri, S., Pola, G., Petroni, F., Young, M. P., & Petersen, R. (2002). A critical assessment of different measures of the information carried by correlated neuronal firing. *Biosystems*, *67*, 177–185.

Panzeri, S., & Schultz, S. (2001). A unified approach to the study of temporal, correlational and rate coding. *Neural Computation*, *13*, 1311–1349.

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

Petersen, R. S., & Diamond, M. (2000). Spatio-temporal distribution of whisker-evoked activity in rat somatosensory cortex and the coding of stimulus location. *J. Neurosci.*, *20*, 6135–6143.

Petersen, R. S., Panzeri, S., & Diamond, M. (2001). Population coding of stimulus location in rat somatosensory cortex. *Neuron*, *32*, 503–514.

Pola, G., Thiele, A., Hoffmann, K.-P., & Panzeri, S. (2003). An exact method to quantify the information transmitted by different mechanisms of correlational coding. *Network*, *14*, 35–60.

Reich, D. S., Mechler, F., Purpura, K. P., & Victor, J. D. (2000). Interspike intervals, receptive fields, and information encoding in primary visual cortex. *J. Neurosci.*, *20*, 1964–1974.

Reich, D. S., Mechler, F., & Victor, J. D. (2001). Independent and redundant information in nearby cortical neurons. *Science*, *294*, 2566–2568.

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., Franco, L., Aggelopoulos, N. C., & Reece, S. (2003). An information theoretic approach to the contributions of the firing rates and the correlations between the firing of neurons. *J. Neurophysiol.*, *89*, 2810–2822.

Romo, R., Hernandez, A., Zainos, A., & Salinas, E. (2003). Correlated neuronal discharges that increase coding efficiency during perceptual discrimination. *Neuron*, *38*, 649–657.

Schneidman, E., Bialek, W., & Berry, M. J. (2003). Synergy, redundancy, and independence in population codes. *J. Neurosci.*, *23*(37), 11539–11553.

Schultz, S., & Panzeri, S. (2001). Temporal correlations and neural spike train entropy. *Phys. Rev. Lett.*, *86*, 5823–5826.

Shadlen, M. N., & Newsome, W. T. (1998). The variable discharge of cortical neurons: Implications for connectivity, computation and coding. *J. Neurosci.*, *18*(10), 3870–3896.

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

Thiele, A., Distler, C., & Hoffmann, K. P. (1999). Decision-related activity in the macaque dorsal visual pathway. *European J. Neurosci.*, *11*, 2044–2058.

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

Vaadia, E., Haalman, I., Abeles, M., Bergman, H., Prut, Y., Slovin, H., & Aertsen, A. (1995). Dynamics of neuronal interactions in monkey cortex in relation to behavioural events. *Nature*, *373*, 515–518.

Victor, J. D. (2002). Binless strategies for estimation of information from neuronal data. *Physical Review, E 66*, 51903–51918.

Received September 7, 2004; accepted March 3, 2005.