Behavioral/Systems/Cognitive

Synergy, Redundancy, and Independence in Population Codes

Elad Schneidman,[1,2] William Bialek,[2] and Michael J. Berry II[1]
Departments of [1]Molecular Biology and [2]Physics, Princeton University, Princeton, New Jersey 08544

A key issue in understanding the neural code for an ensemble of neurons is the nature and strength of correlations between neurons and how these correlations are related to the stimulus. The issue is complicated by the fact that there is not a single notion of independence or lack of correlation. We distinguish three kinds: (1) activity independence; (2) conditional independence; and (3) information independence. Each notion is related to an information measure: the information between cells, the information between cells given the stimulus, and the synergy of cells about the stimulus, respectively. We show that these measures form an interrelated framework for evaluating contributions of signal and noise correlations to the joint information conveyed about the stimulus and that at least two of the three measures must be calculated to characterize a population code. This framework is compared with others recently proposed in the literature. In addition, we distinguish questions about how information is encoded by a population of neurons from how that information can be decoded. Although information theory is natural and powerful for questions of encoding, it is not sufficient for characterizing the process of decoding. Decoding fundamentally requires an error measure that quantifies the importance of the deviations of estimated stimuli from actual stimuli. Because there is no *a priori* choice of error measure, questions about decoding cannot be put on the same level of generality as for encoding.

Key words: **encoding; decoding; neural code; information theory; signal correlation; noise correlation**

Introduction

One of the fundamental insights of neuroscience is that single neurons make a small, but understandable, contribution to an animal's overall behavior. However, most behaviors involve large numbers of neurons, thousands or even millions. In addition, these neurons often are organized into layers or regions, such that nearby neurons have similar response properties. Thus, it is natural to ask under what conditions groups of neurons represent stimuli and direct behavior in either a synergistic, redundant, or independent manner. With the increasing availability of multielectrode recordings, it now is possible to investigate how sensory data or motor intentions are encoded by groups of neurons and whether that population activity differs from what can be inferred from recordings of single neurons. Complementary to this question is how population activity can be decoded and used by subsequent neurons.

The code by which single neurons represent and transmit information has been studied intensively (Perkel and Bullock, 1968; Rieke et al., 1997; Dayan and Abbott, 2001). Many of the conceptual approaches and analytic tools used for the single neuron case can be extended to the multiple neuron case. The key additional issue is the nature and strength of correlations between neurons. Such correlations have been measured using simultaneous recording, and their influence on population encoding has been assessed with a variety of methods (Perkel et al., 1967; Mastronarde, 1983; Aertsen et al., 1989; Gray and Singer, 1989; Abeles et al., 1993; Laurent and Davidowitz, 1994; Meister et al., 1995; Vaadia et al., 1995; Krahe et al., 2002). The intuitive notion of synergy has been quantified in various systems using information theory (Gawne and Richmond, 1993; Gat and Tishby, 1999; Brenner et al., 2000; Petersen et al., 2001). Studies of population decoding have examined how animals might extract information from multiple spike trains (Georgopoulos et al., 1986; Abeles et al., 1993; Zohary et al., 1994; Warland et al., 1997; Brown et al., 1998; Hatsopoulos et al., 1998), as well as the limits of possible decoding algorithms (Palm et al., 1988; Seung and Sompolinsky, 1993; Salinas and Abbott, 1994; Brunel and Nadal, 1998; Zemel et al., 1998).

Here, we describe a quantitative framework for characterizing population encoding using information theoretic measures of correlation. We distinguish the sources of correlation that lead to synergy and redundancy and define bounds on those quantities. We also discuss the consequences of assuming independence for neurons that are actually correlated. Many of the quantities we define have been published previously (Gawne and Richmond, 1993; Gat and Tishby, 1999; Panzeri et al., 1999; Brenner et al., 2000; Chechik et al., 2002). Here, we bring them together, show their interrelations, and compare to alternative definitions. In particular, Nirenberg et al. (2001, 2003) have proposed a measure of the amount of information lost when a decoder ignores noise correlations. We show that their interpretation of this quantity is incorrect, because it leads to contradictions, including that in some circumstances, the amount of information loss may be

Received March 12, 2003; revised Sept. 15, 2003; accepted Sept. 17, 2003.

This work was supported by a Pew Scholar Award and a grant from the E. Mathilda Ziegler Foundation to M.J.B. and by a grant from the Rothschild Foundation to E.S. We thank Adrienne Fairhall for many helpful discussions.

Correspondence should be addressed to Michael J. Berry II, Department of Molecular Biology, Princeton University, Princeton, NJ 08544. E-mail: berry@princeton.edu.

 0270-6474/03/2311539-15$15.00/0

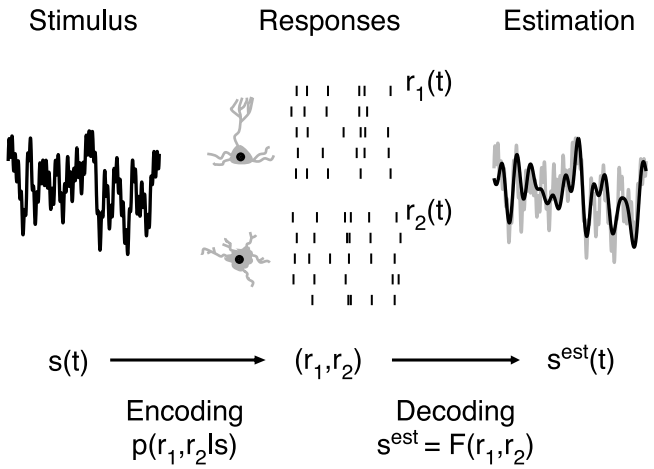

Figure 1. A diagram of neural encoding and decoding. A pair of neurons, 1 and 2, encode information about a stimulus, $s(t)$, with spike trains, $r_1(t)$ and $r_2(t)$. This may be described by the conditional probability distribution of the responses given the stimulus $p(r_1, r_2|s)$. Decoding is the process of trying to extract this information explicitly, which may be done by other neurons or by the experimentalist. This process is described by a function, F, that acts on r_1 and r_2 and gives an estimated version of the stimulus.

greater than the amount of information that is present. We argue that their measure is related more closely to questions of decoding than encoding, and we discuss its interpretation.

Results

To understand the manner in which neurons represent information about the external world, it is important to distinguish the concepts of encoding and decoding. Figure 1 shows a schematic of encoding and decoding for a pair of neurons. Encoding is the conversion of stimuli into neural responses; this process is what we observe experimentally. Decoding is a procedure that uses the neural spike trains to estimate features of the original stimulus or make a behavioral decision. The experimentalist uses a chosen algorithm to either reconstruct stimulus features or to predict a motor or behavioral outcome. The goal is to understand how information encoded by neurons can be explicitly recovered by downstream neurons and what decisions the animal might make based on these neural responses.

Neural encoding

In general, neural responses are noisy, meaning that repeated presentations of the same stimulus give rise to different responses (Verveen and Derksen, 1968; Mainen and Sejnowski, 1995; Bair and Koch, 1996). Although the observed noise often has a component caused by incomplete control of experimental variables, all neural systems exhibit sources of noise that operate even under ideal experimental conditions. Thus, the relationship between a stimulus and the resulting neural response must be described by a probabilistic dictionary (for review, see Rieke et al., 1997). In particular, for every possible stimulus s, there is a probability distribution over the possible responses r given that stimulus, namely $p(r|s)$.

Questions of neural encoding involve what response variables represent information about the stimulus, what features of the stimulus are represented, and specifically how much one can learn about the stimulus from the neural response. Given the distribution of stimuli in the environment, $p(s)$, the encoding dictionary $p(r|s)$ contains the answers to these questions.

Because the encoding dictionary is a complex object, it has often been useful to summarize its properties with a small number of functions, such as the spike-triggered average stimulus or the firing rate as a function of stimulus parameters. An especially appealing measure is the mutual information between the stimuli and the responses (Shannon and Weaver, 1949; Cover and Thomas, 1991):

$$I(S; R) = \sum_{s \in S} \sum_{r \in R} p(s, r) \log_2 \left[\frac{p(s, r)}{p(s)p(r)} \right] \text{bits}, \qquad (1)$$

where S denotes the set of stimuli $\{s\}$ and R denotes the set of responses $\{r\}$. The mutual information measures how tightly neural responses correspond to stimuli and gives an upper bound on the number of stimulus patterns that can be discriminated by observing the neural responses. Its values range from zero to either the entropy of the stimuli or the entropy of the responses, whichever is smaller. The mutual information is zero when there is no correlation between stimuli and responses. The information equals the entropy of the stimulus when each possible stimulus generates a uniquely identifiable response, and it equals the entropy of the responses when there is no noise (Shannon and Weaver, 1949). Many authors have studied single neuron encoding using information theory (Mackay and McCulloch, 1952; Fitzhugh, 1957; Eckhorn and Popel, 1974; Abeles and Lass, 1975; Optican and Richmond, 1987; Bialek et al., 1991; Strong et al., 1998).

Mutual information is appealing for several reasons. First, it is a very general measure of correlation between stimulus and response and can be thought of as including contributions from all other measures of correlation. Second, it does not make assumptions about what features of the stimuli or responses are relevant, which makes information theory uniquely well suited to the analysis of neural responses to complex, naturalistic stimuli (Lewen et al., 2001). Third, as signals flow through the nervous system, information can be lost but never gained, a property known as the data processing inequality (Cover and Thomas, 1991). Finally, mutual information is the unique functional of the encoding dictionary that obeys simple plausible constraints, such as additivity of information for truly independent signals (Shannon and Weaver, 1949). For these reasons, we focus here on an information theoretic characterization of population encoding.

Spike train entropies and mutual information are notoriously difficult to estimate from limited experimental data. Although this is an important technical difficulty, there are many cases in which the mutual information has been estimated for real neurons responding to complex, dynamic inputs, with detailed corrections for sampling bias (Strong et al., 1998; Berry and Meister, 1998; Buracas et al., 1998; Reich et al., 2000; Reinagel and Reid, 2000). Many authors have explored strategies for estimating spike train entropies (Treves and Panzeri, 1995; Strong et al., 1998; Victor, 2002; Nemenman et al., 2003; Paninski, 2003), and there is continuing interest in finding improved strategies. We emphasize that these technical difficulties can and should be separated from the conceptual questions involving which information theoretic quantities are interesting to calculate and what they mean.

Encoding versus decoding

While the concept of encoding is relatively straightforward for neurons, decoding is more subtle. Many authors think implicitly or explicitly about an intermediate step in decoding, namely the

formation of the conditional stimulus distribution, $p(s|r)$, using Bayes' rule:

$$p(s|r) = \frac{p(r|s)\,p(s)}{p(r)}. \quad (2)$$

This probability distribution describes how one's knowledge of the stimulus changes when a particular neural response is observed; this distribution contains all of the encoded information (de Ruyter van Stevenick and Bialek, 1988). Some even call this intermediate step decoding (Dayan and Abbott, 2001). Although this distinction might be viewed as semantic, we note that the action of a stimulus–response pathway in an organism results in an actual motor output, not a distribution of possible outputs. Thus, the decision-making process that produces a single output is different from forming $p(s|r)$ and is necessary to use the information encoded by neural spike trains. Furthermore, there are some methods of stimulus estimation, such as linear decoding, that do not make explicit reference to $p(s|r)$ (Bialek et al., 1991), so this intermediate step is not always required. For these reasons, we prefer to think of decoding as the process that actually estimates the stimulus and the formation of the conditional stimulus distribution, where relevant, as the raw material on which many decoding algorithms act. As such, we refer to this distribution as a decoding dictionary.

In the case of encoding, there is a single response distribution to be measured, $p(r|s)$, and the mutual information between stimulus and response implied by this distribution provides a powerful characterization of the encoding properties of these neural responses. However, in the case of decoding, there are many possible algorithms that can be used on the same neural responses. Often, one talks about an "optimal" decoder, meaning that one chooses a class of possible decoding algorithms and adjusts the specific parameters of that algorithm for the best results. This raises the question of what makes one decoder better than another. One obvious figure of merit is the information that the estimated stimulus conveys about the original stimulus, $I(S; S^{\text{est}})$. Intuitively, the best decoder is the one that captures the most of the encoded information. Furthermore, the data processing inequality implies that $I(S; S^{\text{est}}) \leq I(S; R)$, so that there is an absolute standard against which to make this comparison.

Unfortunately, mutual information alone is an insufficient measure with which to evaluate the success of a decoder. Mutual information only measures the correspondence between the original and estimated stimulus, not whether the estimated stimulus equals or approximates the original stimulus. This fact is shown in Figure 2 by an example of a perfectly scrambled decoder. This decoder achieves a one-to-one mapping between the estimated and original stimuli but always makes the wrong estimate. Such a decoder retains all of the information about the stimulus but is obviously doing a bad job. For an organism to appropriately act on the information encoded by neural spike trains, it must actually make the correct estimate. Thus, decoders fundamentally must be evaluated with respect to an error measure, $E(s, s^{\text{est}})$, that describes the penalty for differences between the estimated and original stimuli.

Importantly, there is no universal measure of whether an error is large or small. For instance, a particular error in estimating the location of a tree branch may be fatal if you are a monkey trying to jump from one branch to the next but acceptable when trying to reach for a piece of fruit. Errors may also be strongly asymmetric: failing to notice the presence of a predator may result in death, whereas unnecessarily executing an escape response only wastes

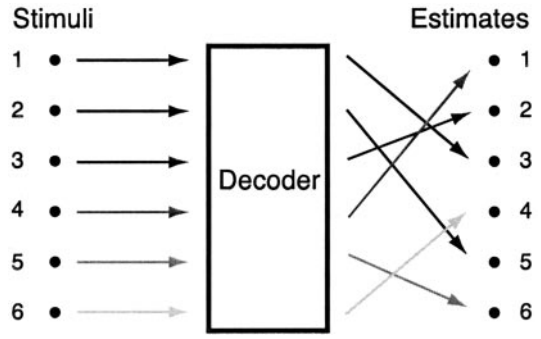

Figure 2. Schematic of a scrambled decoding process. Six stimuli, $\{s\}$, are encoded by neural responses and mapped by a decoder onto six estimated stimuli, $\{s^{\text{est}}\}$. This mapping is one-to-one, so it preserves all the information in the stimulus. However, the estimates are scrambled, so that this decoder never gives the correct answer.

finite resources. Thus, any notion of a natural measure of the error stems from the objective that the decoder is trying to achieve. Because there is no "correct" error measure against which to judge the success of a decoder, statements about decoding cannot be put on the same level of generality as statements about encoding. Information theory can still play a role in characterizing decoding, but only in conjunction with an error measure.

Population encoding

Many questions about the nature of encoding by a population of neurons are extensions of the questions dealing with a single neuron. Instead of studying the single-cell response distribution, we need to use the set of responses of N neurons, given by $p(\vec{r}|s)$, where $\vec{r} = \{r_1, r_2, \ldots, r_N\}$. Similarly, using the joint probability distribution, $p(s, \vec{r})$, we can calculate the mutual information between the set of responses and the stimulus. For two cells:

$$I(S; R_1, R_2) = \sum_s \sum_{r_1, r_2} p(s, r_1, r_2) \log_2 \left[\frac{p(s, r_1, r_2)}{p(s)\,p(r_1, r_2)} \right]. \quad (3)$$

The main additional issue for neural encoding by a population of cells is the correlation among these cells and how these correlations relate to the stimulus. To understand how a population code differs from the codes of its constituent neurons, we must identify appropriate measures of correlation and independence and quantify their relation to the stimulus. In many ways, the question of how responses of multiple neurons can be combined to provide information about the stimulus is related to the question of how successive responses (spikes, bursts, etc.) of a single neuron can be combined to provide information about a stimulus that varies in time (see, for example, Brenner et al., 2000).

Three kinds of independence

Independence and correlation are complementary concepts: independence is the lack of correlation. The statistics community has long noted the distinction between independence and conditional independence and its implications (Dawid, 1979). This distinction has been applied to neuroscience in the classic work of Perkel et al. (1967). Following their example, it has been common to use cross-correlation as a measure of these dependencies (Palm et al., 1988). In the case of the neural code, we are interested primarily in the relation between stimuli and responses, which is itself another form of correlation. Thus, for neural codes, there are three kinds of independence. This diversity is the result of the fact that different sources of correlation have different impacts on

the manner in which neural activity encodes information about a stimulus (Gawne and Richmond, 1993; Gat and Tishby, 1999; Panzeri et al., 1999; Brenner et al., 2000; Chechik et al., 2002). These notions are distinct in the sense that if a pair of neurons possesses one form of independence, it does not necessarily possess the others. Here, we present definitions of the three kinds of independence, along with corresponding information theoretic measures of correlation, which quantify how close the neurons are to being independent.

Activity independence. The most basic notion of correlation is that the activity of one cell, r_1, depends on the activity of another cell, r_2, when averaged over the ensemble of stimuli. This notion of correlated activity is assessed by looking at the joint distribution of the responses of a cell pair, $p(r_1, r_2)$. This joint distribution can be found from the simultaneously recorded responses by summing over stimuli:

$$p(r_1, r_2) = \sum_s p(r_1, r_2|s)p(s). \quad (4)$$

If there is no correlated activity between the pair of cells, then this distribution factors:

$$p(r_1, r_2) = p(r_1)\,p(r_2). \quad (5)$$

The natural measure of the degree of correlation between the activity of two neurons is the information that the activity of one cell conveys about the other:

$$I(R_1; R_2) = \sum_{r_1,r_2} p(r_1, r_2)\log_2\left[\frac{p(r_1, r_2)}{p(r_1)\,p(r_2)}\right] \text{bits}. \quad (6)$$

If the activity of the cells is independent, then $I(R_1; R_2) = 0$. Because the information is bounded from above by the entropy of the responses of each cell, it is possible to use a normalized measure, $I(R_1;R_2)/\min[H(R_1), H(R_2)]$, where $H(R_i)$ is the entropy of the responses of cell i. This normalized measure ranges between 0 and 1. The value of $I(R_1;R_2)$ implicitly depends on the stimulus ensemble S, as can be seen from Equation 4. For simplicity, we leave this dependence out of our notation, but one should keep in mind that activity independence is a property of both a population of neurons and an ensemble of stimuli.

One could ask, perhaps more abstractly, for a measure of similarity between the distributions $p(r_1, r_2)$ and $p(r_1)\,p(r_2)$ and then interpret this measure as a degree of (non)independence. There are even other, common information theoretic measures, such as the Kullback–Leibler (KL) divergence (Cover and Thomas, 1991) or the Jensen–Shannon divergence (Lin, 1991). It is important to note that all such similarity measures are answers to specific questions and, as such, cannot necessarily be used interchangeably. For instance, the Jensen–Shannon divergence measures how reliably one can decide if a given response comes from the joint distribution, $p(r_1, r_2)$, or the product distribution, $p(r_1)\,p(r_2)$, given that these are the only alternatives. It has a maximal value of 1 bit, when the two distributions are perfectly distinguishable. In contrast, the mutual information has a maximal value equal to the spike train entropy, when the two responses are identical.

In this case, the KL divergence between $p(r_1, r_2)$ and $p(r_1)\,p(r_2)$ is, in fact, identical to the mutual information between R_1 and R_2. This holds because the mutual information is a special type of KL divergence, one that is taken between two particular probability distributions. However, the converse is not true: the KL divergence between two arbitrary distributions is not necessarily a mutual information. Therefore, the specific questions answered by the KL divergence are, in general, different from those answered by the mutual information (see below for a discussion of the interpretation of the KL divergence).

The mutual information $I(R_1; R_2)$ measures directly how much (in bits) the response of one cell predicts about the response of the other. We will see that this mutual predictability contributes to redundancy in what the cells can tell us about their stimulus. In addition to being an appealing and general measure of correlation, we will see below that this choice of information measure results in an interrelated framework for the three different kinds of independence.

Conditional independence. Correlated activity between two neurons can arise either from shared stimulation, such as from correlations in their stimuli or overlap in their receptive fields, or from shared sources of noise, such as a presynaptic neuron that projects to both neurons or a common source of neuromodulation. In the former case, the correlations between neurons can be explained from knowledge of how each neuron alone responds to the stimulus, whereas in the latter case they cannot. Therefore, an important distinction is whether the correlations are solely attributable to the stimulus ("signal" correlations) or not ("noise" correlations). Although this nomenclature is widely used, one should keep in mind that "noise" correlations are not always detrimental to the neural code.

The strength of noise correlations can be assessed by looking at the joint distribution of neural activity conditioned on the stimulus $p(r_1, r_2|s)$. If two neurons respond independently to the stimulus, they are called "conditionally independent," and the distribution of responses factors for all s:

$$p(r_1, r_2|s) = p(r_1|s)\,p(r_2|s). \quad (7)$$

As in the case of activity independence, a natural measure of conditional independence is the mutual information between cells given the stimulus

$$I(R_1; R_2|s) = \sum_{r_1,r_2} p(r_1, r_2|s)\log_2\left[\frac{p(r_1, r_2|s)}{p(r_1|s)\,p(r_2|s)}\right]. \quad (8)$$

By measuring the dependence between neurons for each stimulus s, this quantity ignores all correlations that arise from shared stimulation and, thus, equals zero only if there are no noise-induced correlations. A normalized measure is $I(R_1;R_2|s)/\min(H(R_1|s),H(R_2|s))$, which ranges between 0 and 1. For many purposes, it is useful to compute the average over stimuli, $\langle I(R_1; R_2|s)\rangle_s$.

The distinction between signal and noise correlations relates directly to an important distinction in experimental technique: noise correlations can only be measured by recording simultaneously from a pair of neurons. A simple technique of demonstrating the existence of noise correlations is the shuffle test or "shift predictor" (Perkel et al., 1967; Palm et al., 1988), where the cross-correlation between simultaneously recorded pairs of neurons are compared on the same stimulus trial versus different stimulus trials. Of course, as a practical matter, it is preferable to measure even signal correlations simultaneously and from the same preparation, because of nonstationarities in neural responses.

Although the shuffle-corrected cross-correlation function may seem intuitive and straightforward, it actually suffers from ambiguities in how to normalize and interpret its values. The apparent strength of cross-correlation between two neurons de-

pends on the auto-correlation function of each neuron, so that observed changes in cross-correlation contain this potential confound (Brody, 1999). Also, the cross-correlation function can be expressed in different units: firing rate of one cell relative to the other, fraction of total spikes within a time window, etc. There are subtle differences between these choices of units (such as whether the measure is symmetric) that make their interpretation tricky. In contrast, the quantity $\langle I(R_1; R_2|s)\rangle_s$ provides a characterization of noise correlations that resolves these ambiguities, has a clear-cut interpretation, and is sensitive to forms of correlation not captured by the shuffle-corrected correlogram (e.g., if the response of one neuron is more precise when the other neuron is active).

Pairs of neurons that are conditionally independent are not necessarily activity independent, because shared stimulation may still induce correlations in their responses when averaged over the entire stimulus ensemble. For a simple example, consider two binary neurons that produce either a spike or no spike in response to two, equally likely stimuli. They each respond to the first stimulus with a 50% probability of spiking, but neither fires in response to the second stimulus. These neurons possess conditional independence, because their joint response distribution factors for each stimulus, but not activity independence, because if one cell stays silent, the other is more likely to stay silent.

Conversely, pairs of neurons that are activity independent are not necessarily conditionally independent, because noise correlations may increase the probability that neurons fire together for some stimuli and decrease it for others, such that those contributions roughly cancel when averaged over the stimulus ensemble. For an example of this case, consider an extreme instance of stimulus-dependent correlations: binary neurons such that for the first stimulus either both fire or both remain silent with equal probability, but for the second stimulus, either one fires a spike and the other remains silent, or vice versa, with equal probability. Here, the neurons are positively correlated for the first stimulus and negatively correlated for the second. They are clearly not conditionally independent, but because the positive and negative correlations occur with equal strength, they are activity independent. Notice that if the two stimuli occur with unequal probability, then the cell pair is no longer activity independent. As these examples demonstrate, activity independence and conditional independence are distinct measures of correlation between neurons.

Information independence. A final notion of correlation relates to the information encoded by a cell pair. Intuitively, if the cells are sensitive to completely different features of the stimulus, then the information they convey together should just be the sum of what they convey separately:

$$I(S; R_1, R_2) = I(S; R_1) + I(S; R_2). \quad (9)$$

Cell pairs that do not encode information independently can be either synergistic, meaning that they convey more information in their joint responses than the sum of their individual information, or redundant, meaning that they jointly convey less. Thus, the obvious measure of information independence is the synergy (Gawne and Richmond, 1993; Gat and Tishby, 1999; Panzeri et al., 1999; Brenner et al., 2000):

$$\mathrm{Syn}(R_1, R_2) = I(S; R_1, R_2) - I(S; R_1) - I(S; R_2). \quad (10)$$

Negative values of this quantity indicate redundancy. A normalized version of the synergy is given by $\mathrm{Syn}(R_1, R_2)/I(S; R_1, R_2)$, which ranges between -1, when the responses of the two neurons are related by a one-to-one mapping, and 1, when the cell pair only conveys information by its joint response and there is zero information contained in the responses of each individual cell.

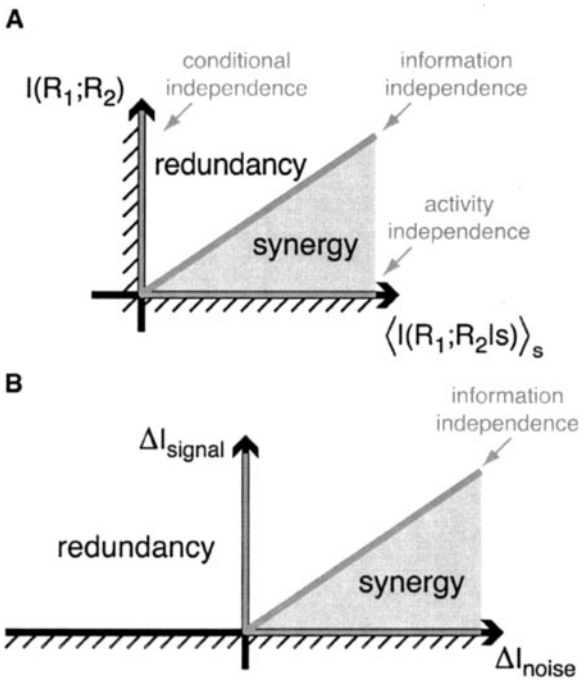

Figure 3. Graphical presentations of synergy as a combination of other measures of independence. *A*, Following Equation 11, we can represent the synergy or redundancy of a pair of cells as a point in a plane with the axes $\langle I(R_1; R_2|s)\rangle_s$ and $I(R_1; R_2)$. Because both of these measures are non-negative, only the top right quadrangle is allowed. Neurons that possess activity independence lie on points along the abscissa. Neurons that possess conditional independence lie on points along the ordinate. Information independence corresponds to the diagonal that separates the synergistic values from the redundant ones. *B*, Similarly, following Equation 16, we can also express the synergy as a point in a plane with the axes $\Delta I_{\mathrm{noise}}$ and $\Delta I_{\mathrm{signal}}$. Because $\Delta I_{\mathrm{signal}}$ is non-negative, only the top half plane is allowed.

It is important to note that synergy, as defined here, is a property that is averaged over the stimulus ensemble. Cell pairs can be synergistic for some subset of the stimuli, redundant during others, and independent for yet other stimuli. Hence, when cells are found to be information independent, this may result from averaging over synergistic and redundant periods rather than from independence at all times.

An alternative way to write the synergy is as the difference between the mutual information between the cells given the stimulus and the information that they share that is not explicitly related to the stimulus (Brenner et al., 2000):

$$\mathrm{Syn}(R_1, R_2) = \langle I(R_1; R_2|s)\rangle_s - I(R_1; R_2), \quad (11)$$

which is a combination of the measures of conditional and activity independence (see Eq. 6 and 8). If a pair of neurons possesses both activity and conditional independence, then there is no synergy or redundancy. However, information independence may hold without activity independence and conditional independence, when these two terms cancel. Thus, the three measures of independence and correlation are interconnected, giving a structured framework for the quantification of correlation and independence. Figure 3*A* shows a graphic presentation of synergy as a combination of the two other independence measures, reflecting that two dimensions are needed to describe the nature of neural (in)dependence.

Because each term in Equation 11 is non-negative, the first

term contributes only synergy and the second only redundancy. By writing the synergy in this form, one can readily see that $\langle I(R_1;R_2|s)\rangle_s$ is an upper bound on the synergy. Because this term is non-negative for all stimuli, there can be no cancellation in its value when the cell pair is synergistic for some stimuli and redundant for others. Similarly, $-I(R_1; R_2)$ is a bound on the redundancy of a pair of neurons.

Assuming conditional independence

Sampling the distribution of joint responses of pairs or groups of cells requires, in general, exponentially more data than the single cell case. Hence, the characterization of neural population activity is often severely constrained by experimental limitations. Because it is easier to sample the responses of individual cells, even when neurons can be recorded simultaneously, one may try to approximate the joint distribution by assuming that the cell pair is conditionally independent. Furthermore, when using recordings from different trials (Georgopoulos et al., 1986), or even different animals (Chechik et al., 2002), one must make this assumption.

When ignoring the fact that the pair of cells were recorded simultaneously or when combining the nonsimultaneous recordings of cells presented with the exact same stimulus, a customary guess for the joint response distribution is given by:

$$p_{\text{shuffle}}(r_1, r_2|s) = p(r_1|s)\,p(r_2|s). \tag{12}$$

We use the notation "shuffle," because this is the joint response distribution that would result from compiling the responses of simultaneously recorded cells from different, or shuffled, stimulus trials (similar to the "shift predictor") (Perkel et al., 1967; Palm et al., 1988). Notice also that this assumption implies that the strength of noise correlations measured by Equation 8 is zero. The information that the shuffled cell responses convey about the stimulus is given by:

$$I_{\text{shuffle}}(S; R_1, R_2) = \sum_s p(s) \sum_{r_1,r_2} p(r_1|s)\,p(r_2|s) \log_2 \left[\frac{p(r_1|s)\,p(r_2|s)}{\sum_{s'} p(r_1|s')\,p(r_2|s')\,p(s')} \right]. \tag{13}$$

The difference between the information conveyed by a cell pair in the real case and I_{shuffle},

$$\Delta I_{\text{noise}} = I(S; R_1, R_2) - I_{\text{shuffle}}(S; R_1, R_2), \tag{14}$$

measures the contribution of noise-induced correlations to the encoded information. This value may be either positive or negative, depending on whether those correlations lead to synergy or redundancy (for specific example, see Fig. 5). Furthermore, the difference between the sum of the information that each of the cells individually conveys about the stimulus and I_{shuffle}:

$$\Delta I_{\text{signal}} = I(S; R_1) + I(S; R_2) - I_{\text{shuffle}}(S; R_1, R_2), \tag{15}$$

measures the effect of signal-induced correlations on the encoded information. This value is non-negative (see Appendix A), because signal correlations indicate that the two cells are, in part, encoding identical information and, thus, implies redundancy. The difference between these two terms gives the synergy of the two cells:

$$\text{Syn}(R_1, R_2) = \Delta I_{\text{noise}} - \Delta I_{\text{signal}}. \tag{16}$$

When neurons are not recorded simultaneously, one typically assumes that $\Delta I_{\text{noise}} = 0$. With this assumption and the fact that ΔI_{signal} is non-negative, the only possible result is apparent net redundancy. This is reflected in Figure 3*B*, which gives a graphic presentation of the signal and noise components as the two dimensions that span the synergy. We emphasize that although the ΔI_{signal} and ΔI_{noise} quantify the influence of signal and noise correlations, unlike the quantities defined previously, these are not mutual information measures.

Population encoding for three or more neurons

In the preceding sections, we focused on the case of two neurons. The basic distinctions we made between activity and conditional independence as well as their connections to the distinction between signal and noise correlations will hold for the case of three or more neurons. One should note, however, that correlations among n neurons can be assessed in more than one way. For instance, one can compare the correlations among n neurons to the correlations only observable among $n - 1$ neurons (Martignon et al., 2000) or one can compare n neuron correlations to n independent single cells (Chechik et al., 2002). For the case of two cells, these two comparisons are the same, but for three or more cells they differ (Schneidman et al., 2003).

Comparison to other measures

Approximate conditional stimulus distributions

In a recent study, Nirenberg et al. (2001) studied the importance of noise correlations for how information is encoded by pairs of ganglion cells in the retina. Noise correlations can be ignored explicitly by assuming that the joint response distribution for two neurons is given by Equation 7. Bayes' rule can be used to find the stimulus distribution conditioned on the neural response for that case:

$$p_{\text{shuffle}}(s|r_1, r_2) = \frac{p(r_1|s)\,p(r_2|s)\,p(s)}{\sum_{s'} p(r_1|s')\,p(r_2|s')\,p(s')}. \tag{17}$$

Nirenberg et al. (2001) denoted this quantity by $p_{\text{ind}}(s|r_1, r_2)$, but we use p_{shuffle} to avoid confusion between different kinds of independence. They suggested using the KL divergence between the true decoding dictionary $p(s|r_1, r_2)$ and the approximate dictionary $p_{\text{shuffle}}(s|r_1, r_2)$ to quantify the amount of information that is lost by using a decoder that assumes conditional independence. Averaged over the real, correlated responses, r_1 and r_2, one obtains:

$$\hat{D} = \sum_{r_1,r_2} p(r_1, r_2)\, D_{\text{KL}}[\,p(s|r_1, r_2) \| p_{\text{shuffle}}(s|r_1, r_2)\,], \tag{18}$$

This measure does not refer to any specific algorithm for estimating the stimulus or errors made by that algorithm but, instead, is meant to be a general characterization of the ability of any decoder to make discriminations about the stimulus, if knowledge of the noise correlations is ignored. Nirenberg et al. (2001) argued that it is appropriate to consider an approximate decoding dictionary combined with the real spike trains, because the brain always automatically has access to the real, correlated spike trains but may make simplifying assumptions about how to decode the

information that those spike trains contain. They state that $\hat{D}$ measures the loss in information that results from ignoring correlations in the process of decoding and, thus, refer to this measure as ΔI.

Nirenberg and Latham (2003) make a connection between the KL divergence and the encoded information by using an argument about the number of yes/no questions one must ask to specify the stimulus (see below). Although this argument may initially seem reasonable, closer consideration reveals that it is flawed. This can be demonstrated by the direct contradiction that results from assuming that the KL divergence measures an information loss, as well as the contradictory implications of this argument. In particular, there are situations in which this putative information loss can be greater than the amount of information present. Furthermore, interpreting the measure $\hat{D}$ as a general test of the importance of noise correlations for encoding information about a stimulus is problematic, because of the highly counterintuitive results that one finds when applying the measure to toy models.

Contradiction. The central claim made by Nirenberg et al. (2001) is that $\hat{D}$ measures the amount of information about the stimulus that is lost when one ignores noise correlations. If this were true, then the information that such a decoder can capture would be given by:

$$\text{“}I_{\text{no-noise}}(S; R_1, R_2)\text{”} = I(S; R_1, R_2) - \hat{D}$$

$$= \sum_{r_1,r_2} p(r_1, r_2) \sum_{s} p(s|r_1, r_2) \log_2 \frac{p_{\text{shuffle}}(s|r_1, r_2)}{p(s)} \quad (19)$$

This expression for “$I_{\text{no-noise}}$” is unusual. It does not obviously have the form of a mutual information, as is evident from the fact that the probability distribution inside the logarithm is not the same as that multiplying the logarithm. The fact that Equation 19 is not a mutual information term can be demonstrated by specific examples. Figure 4 shows one such case for a pair of model neurons that can generate three different responses (0, 1, or 2 spikes) to each of three, equally likely stimuli. The joint response distribution, $p(r_1, r_2|s)$ is shown in Figure 4*A*. For this toy model, $\hat{D}$ exceeds the total information encoded by both neurons, and, consequently, Equation 19 is negative. This example demonstrates that if one assumes that $\hat{D}$ is an information loss, then one would sometimes lose more information than was present by ignoring noise correlations. Because the mutual information between the output of a decoder and the input stimulus cannot be negative, this is a clear contradiction. Therefore, $\hat{D}$ is not an information loss.

Counter-intuitive properties of $\hat{D}$. Because $\hat{D}$ is always positive, one might wonder whether it sets a useful upper bound on the importance of noise correlations. Again rewriting:

$$\hat{D} = \sum_{s} p(s) \sum_{r_1,r_2} p(r_1, r_2|s) \log \frac{p(r_1, r_2|s)}{p(r_1|s)\, p(r_2|s)}$$

$$- \sum_{r_1,r_2} p(r_1, r_2) \log \frac{p(r_1, r_2)}{p_{\text{shuffle}}(r_1, r_2)}, \quad (20)$$

where:

$$p_{\text{shuffle}}(r_1, r_2) = \sum_{s'} p(s')\, p(r_1|s')\, p(r_2|s'). \quad (21)$$

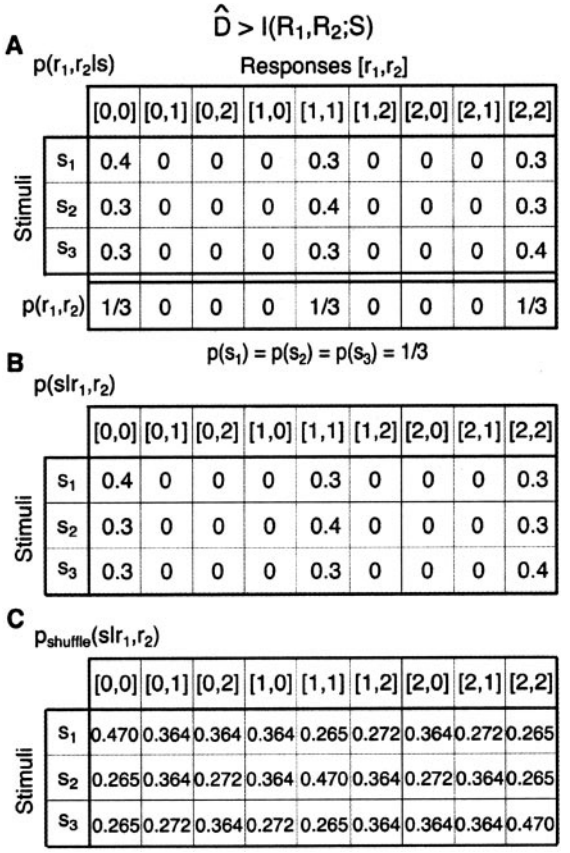

Figure 4. $\hat{D}$ can be larger than the information that the cells encode about the stimulus. *A*, The conditional joint response distribution $p(r_1, r_2|s)$, of two neurons responding to three stimuli. Each of the neurons responds with either zero, one, or two spikes. $p(r_1, r_2)$ is the average of $p(r_1, r_2|s)$ over the stimuli. The *a priori* probability of each of the stimuli equals 1/3. *B*, The conditional stimulus distribution for the cell pair, $p(s|r_1, r_2)$, obtained using Bayes' rule. *C*, The conditional stimulus distribution that assumes no noise correlation, $p_{\text{shuffle}}(s|r_1, r_2)$, obtained by inverting $p(r_1|s)p(r_2|s)$ using Bayes' rule; see text for details. For this case, the information that both cells carry about the stimulus, $I(R_1, R_2; S)$ equals 0.0140 bits, whereas $\hat{D}$ equals 0.0145 bits.

We see that both terms are non-negative, because they both have the form of a KL divergence. The first term, in fact, is $\langle I(R_1;R_2|s)\rangle_s$, which is a measure of the strength of noise correlations and an upper bound on the synergy. Because the second term is non-negative, $\hat{D} \leq \langle I(R_1; R_2|s)\rangle_s$. Therefore, $\hat{D}$ does not constitute an upper bound on the importance of noise correlations as is also demonstrated by specific examples in Figure 5. Even so, perhaps $\hat{D}$ constitutes a tighter upper bound on the synergy than $\langle I(R_1; R_2|s)\rangle_s$? This turns out not to be the case, as shown below.

In Figure 5, we imagine a simple situation in which a pair of neurons can only generate two responses, spike or no spike, and they are only exposed to two different, equally likely stimuli. In both of these examples, the neurons fire a spike with $p = 0.5$ for the first stimulus, but neither fires a spike for the second stimulus. As such, they are sparse in a manner similar to many real neurons. In example A, the response to the first stimulus is perfectly anticorrelated, meaning that if one cell fires a spike, the other stays silent, and vice versa. Knowledge of this noise correlation resolves any ambiguity about the stimulus, such that the joint mutual information is one bit. Because each cell mostly remains silent in this stimulus ensemble, the individual mutual information of each cell is considerably lower, and the synergy of the cell pair is

A $\hat{D} < Syn(R_1,R_2)$

Stimuli	$p(r_1,r_2\|s)$	[0,0]	[0,1]	[1,0]	[1,1]
	s_1	0	0.5	0.5	0
	s_2	1	0	0	0

B $\hat{D} > 0$ but $\Delta I_{noise} < 0$

Stimuli	$p(r_1,r_2\|s)$	[0,0]	[0,1]	[1,0]	[1,1]
	s_1	0.5	0	0	0.5
	s_2	1	0	0	0

Figure 5. Examples of counter-intuitive values of $\hat{D}$. For both examples, there are two stimuli and two neural responses. The probability of each stimulus is 1/2. *A*, One conditional joint response distribution, $p(r_1, r_2|s)$, which results in the synergy of the cells being larger than $\hat{D}$; $I(R_1, R_2, S) = 1$ bit; $Syn(R_1, R_2) = 0.377$ bits; $\hat{D} = 0.161$ bits. *B*, Another conditional joint response distribution, $p(r_1, r_2|s)$, which results in $\hat{D}$ being larger than zero when the noise correlations contribute net redundancy; $I(R_1, R_2; S) = 0.311$ bits; $Syn(R_1, R_2) = -0.311$ bits; $\hat{D} = 0.053$ bits.

$+0.377$ bits. Using Bayes' rule to find the real conditional stimulus distribution and the one that ignores noise correlations, one finds that $\hat{D} = 0.161$ bits, or about 2.3 times smaller than the synergy. This is a strange result, because synergy can only arise from noise correlations. Thus, one naively expects that all synergy is lost when one ignores the noise correlations. Consistent with this expectation, the upper bound on the synergy, $\langle I(R_1; R_2|s)\rangle_s$ is 0.5 bits, and information lost by using shuffled spike trains is $\Delta I_{noise} = 0.451$ bits.

In example B, the two cells have a complete positive correlation for the first stimulus, meaning that they either both fire a spike or both remain silent, each with $p = 0.5$. As before, they both remain silent for the second stimulus. In this case, the two neurons always have exactly the same response. As a result, the synergy equals -0.311 bits, which is a redundancy of 100%. However, they still have quite strong noise correlations, and $\hat{D} = 0.053$ bits or 16.9% of the joint mutual information, which is virtually the same fraction as in example A. This comparison indicates that $\hat{D}$ cannot distinguish between noise correlations that lead to redundancy and those that lead to synergy. In this example, shuffling the spike trains breaks the complete redundancy of the two cells and actually increases the encoded information. Correspondingly, $\Delta I_{noise} = -0.238$ bits or -76% of the joint mutual information (negative values imply that a shuffled set of responses would have more information than the original spike trains).

Figure 6 shows an example with three stimuli and neurons capable of three responses. Here, the neurons have anticorrelations that allow all three stimuli to be perfectly resolved, and the synergy equals $+0.415$ bits or 26.2% of the joint mutual information. However, the correlations between these cells are such that $\hat{D} = 0$. Interestingly, $p_{shuffle}(s|r_1, r_2)$ is not equal to $p(s|r_1, r_2)$ for all joint responses, but in all cases in which they are unequal, the joint response probability $p(r_1, r_2) = 0$ (for related examples and discussion, see Meister and Hosoya, 2001). This example is an extreme illustration, in which the measure $\hat{D}$ implies that there is no cost to ignoring noise correlations, when, in fact, observing the responses of the two cells together provides substantially more information about the stimulus than expected from observations on the individual neurons in isolation. Clearly, the measure $\hat{D}$ cannot be relied on to detect the impact of interesting and important noise correlations on the neural code.

$Syn(R_1,R_2) > 0$ but $\hat{D} = 0$

Stimuli	$p(r_1,r_2\|s)$	[0,0]	[0,1]	[0,2]	[1,0]	[1,1]	[1,2]	[2,0]	[2,1]	[2,2]
	s_1	0	0.5	0	0.5	0	0	0	0	0
	s_2	0	0	0	0	0	0.5	0	0.5	0
	s_3	0	0	0.5	0	0	0	0.5	0	0

$p(s_1) = p(s_2) = p(s_3) = 1/3$

Figure 6. Cells may be synergistic but $\hat{D} = 0$. The conditional joint response distribution $p(r_1, r_2|s)$ of two neurons responding to three stimuli, each with probability 1/3. In this case, the cells are synergistic but $\hat{D}$ is zero: $I(R_1, R_2; S) = 1.585$ bits; $Syn(R_1, R_2) = 0.415$ bits; $\hat{D} = 0$ bits.

Problematic implications of $\hat{D}$. Although Nirenberg and Latham (2003) do not attempt to explore all of the consequences of interpreting their KL divergence as a general measure of information loss, we show here that this argument leads to further contradiction. One corollary of their claim comes from its extension to cases other than assessing the impact of ignoring noise correlations (Nirenberg and Latham, 2003). Hence, one can also ask how much information is lost by a decoder built from any approximate version $\tilde{p}(s|r_1, r_2)$ of the conditional stimulus distribution, and the answer, if we follow the arguments of Nirenberg and Latham (2003), must be $D_{KL}[\, p(s|r_1, r_2) \| \tilde{p}(s|r_1, r_2)\,]$. However, in general, the KL divergence between $p(s|r_1, r_2)$ and $\tilde{p}(s|r_1, r_2)$ can be infinite, if for some value of s, r_1 and r_2, $\tilde{p} = 0$ and $p \neq 0$. This result is clearly impossible to interpret.

Another corollary is that the information loss resulting from ignoring noise correlations is defined for every joint response (r_1, r_2). This means that we can also use the formalism to determine how much information the decoder loses when acting on the shuffled spike trains. This expression is:

$$\hat{D} = \sum_{r_1,r_2} p_{shuffle}(r_1, r_2)\, D_{KL}[\, p(s|r_1, r_2) \| p_{shuffle}(s|r_1, r_2)\,], \qquad (22)$$

However, we have shown above that the mutual information that a pair of neurons conveys about the stimulus under the assumption of conditional independence is $I_{shuffle}$ and the consequent difference in mutual information is ΔI_{noise}. Equation 22 is not identical to ΔI_{noise}. In particular, ΔI_{noise} can be either positive or negative, because the assumption of conditional independence sometimes implies a gain of information rather than a loss (Abbott and Dayan, 1999). This typically occurs when the neurons have positive correlations (Fig. 5, example B) (Petersen et al., 2001). In this case, shuffling the spike trains actually reduces their joint noise and, therefore, can increase the information conveyed about the stimulus. In contrast, Equation 22 is never negative, implying that there is always a loss of information. Thus, another contradiction results.

What does $\hat{D}$ measure? Nirenberg et al. (2001) argue that their average KL divergence measures the number of additional yes/no questions that must be asked to determine the stimulus when a decoder uses the dictionary $p_{shuffle}(s|r_1, r_2)$ instead of $p(s|r_1, r_2)$. They identify this number of yes/no questions with a

loss in mutual information about the stimulus. However, this identification is mistaken. The KL divergence is not equivalent to entropy or an entropy difference (Cover and Thomas, 1991). Any information theoretic quantity that has units of bits can, intuitively, be thought of as representing the number of yes/no questions needed to specify its random variable. However, this does not imply that all such quantities are equivalent. For instance, both the entropy and the mutual information have units of bits. However, they are conceptually different: a neuron firing randomly at a high rate has lots of entropy but no information, whereas a neuron firing at a low rate, but locked precisely to a stimulus, has less entropy but more information.

The precise information theoretic interpretation of the KL divergence comes in the context of coding theory (Cover and Thomas, 1991). If signals x are chosen from a probability distribution $p(x)$, then there exists a way of representing these signals in binary form such that the average "code word" has length equal to the entropy of the distribution. Each binary digit of this code corresponds to a yes/no question that must be asked about the value of x, and, hence, the code length can be thought of as representing the total number of yes/no questions that must be answered, on average, to determine the value of x. Achieving this optimal code requires a strategy matched to the distribution $p(x)$ itself; in particular, the code length for each value x should be chosen to be $-\log_2 p(x)$. The KL divergence between two distributions, $p(x)$ and $q(x)$ and $D_{\mathrm{KL}}[p(x)\|q(x)]$, measures the average extra length of code words for signals x drawn from $p(x)$ using a code that was optimized for $q(x)$. It is not an information loss in any sense. Instead, one might think of D_{KL} as measuring a form of coding inefficiency. In the present context, however, this loss of coding efficiency does not refer to the code of the neuron but, rather, to a nonoptimal code that would be constructed by a hypothetical observer for the conditional stimulus distribution $p(s|r_1, r_2)$.

The KL divergence is commonly used in the literature simply as a measure that quantifies the difference between two probability distributions, without reference to its precise information theoretic interpretations. In this sense, $\hat{D}$ is a sensible measure of the (dis)similarity between $p_{\mathrm{shuffle}}(s|r_1, r_2)$ and $p(s|r_1, r_2)$, but it does not assess how much information about the stimulus can be obtained by using one distribution or the other. Moreover, as a general measure of the dissimilarity of probability distributions, the KL divergence is one of several common choices. Other sensible measures include the L_2 norm and the Jensen–Shannon divergence. Each of these measures is the answer to a specific question about the dissimilarity of two distributions.

Because $\hat{D}$ is a KL divergence between approximate and real decoding dictionaries and because it cannot be interpreted as a loss of encoded information, this quantity should be thought of as a measure related to the problem of decoding and not to the problem of encoding. One important consequence of this distinction is that one cannot reach very general conclusions using any decoding-related measure. As noted above, there are many possible decoding algorithms, and the success of any algorithm is dependent on the choice of an error measure. Thus, the conclusions one reaches about the problem of decoding must always be specific to a given decoding algorithm and a particular error measure.

In the case of $\hat{D}$, one is implicitly assuming that the decoding dictionary is represented by a code book that is optimized for $p_{\mathrm{shuffle}}(s|r_1, r_2)$. This is not the only possible code book that ignores noise correlations. Another possibility is to use one optimized for $p(s)$, which completely ignores the neural response altogether. This counter-intuitive choice does explicitly ignore the noise correlations and, in some circumstances (e.g., the example in Fig. 4), it actually is more efficient than the one optimized for $p_{\mathrm{shuffle}}(s|r_1, r_2)$.

Another source of confusion is that $\hat{D}$ is expressed in units of bits, rather than reconstruction error. This is highly misleading, because the encoded information is also expressed in bits. Although the encoded information provides a completely general bound concerning the performance of any possible decoder, it is important to keep in mind that $\hat{D}$ does not have this level of generality, despite its suggestive units.

What does it mean to ignore noise correlations? The most obvious sense in which one can ignore noise correlations is to combine spike trains from two different stimulus trials. As described above, shuffling the spike trains changes the joint response distribution $p(r_1, r_2|s)$ into $p_{\mathrm{shuffle}}(r_1, r_2|s)$ (Eq. 12) and consequently changes the probability of finding any joint response to $p_{\mathrm{shuffle}}(r_1, r_2)$ (see Eq. 21). Finally, the information that the shuffled spike trains encode about the stimulus is $I_{\mathrm{shuffle}}(S; R_1, R_2)$ (Eq. 13). However, the measure $\hat{D}$ refers to a different circumstance: it assumes a decoding dictionary that ignores noise correlations but combines this with the real, correlated spike trains. For some purposes, this may be an interesting scenario. If $\hat{D}$ does not assess the impact of this assumption on the information encoded about the stimulus, then what is the answer to this question?

In general, this question is ill-defined. The obvious approach is to construct the new joint probability distribution $q(s, r_1, r_2) = p_{\mathrm{shuffle}}(s|r_1, r_2)p(r_1, r_2)$ which combines a decoding dictionary that ignores noise correlations with the real, correlated spike trains. Then, the mutual information between stimulus and responses under the joint distribution q is given by:

$$I_{\mathrm{q}}(S; R_1, R_2) = \sum_{r_1, r_2, s} q(s, r_1, r_2) \log_2 \frac{q(s, r_1, r_2)}{q(s) q(r_1, r_2)}, \qquad (23)$$

where $q(r_1, r_2) = \Sigma_s q(s, r_1, r_2)$ and $q(s) = \Sigma_{r_1, r_2} q(s, r_1, r_2)$. However, this scenario is strange, because simultaneously assuming $p_{\mathrm{shuffle}}(s|r_1, r_2)$ and $p(r_1, r_2)$ implies (through Bayes' rule) that the distribution over the stimuli $q(s)$ is different from the original $p(s)$.

It is also worth noting that this formalism can be extended to the case of assuming any approximate decoding dictionary, $\tilde{p}(s|r_1, r_2)$, by again forming the joint distribution, $\tilde{q}(s, r_1, r_2) = \tilde{p}(s|r_1, r_2)p(r_1, r_2)$. Similarly, a different distribution over the joint responses, $\tilde{p}(r_1, r_2)$, can be inserted. However, the distribution over stimuli $\tilde{p}(s)$ will, in general, not be equal to the actual distribution, $p(s)$. This can lead to contradictory results; for instance, the apparent mutual information can exceed the original stimulus entropy, $I_{\mathrm{q}} > H(S)$, because the new distribution over stimuli $\tilde{p}(s)$ might have larger entropy than $p(s)$.

Nirenberg and Latham (2003) discuss a special case of comparing two neural codes, in which one code is a reduced code or subset of the first (Nirenberg and Latham, 2003). One example of a reduced neural code would be a code that counts spikes in a large time window versus one that keeps many details of spike timing by constructing "words" using spike counts in a smaller time bin (Strong et al., 1998). In this case, the joint response of the reduced code, r', can always be found from the joint response of the full code, r, by a deterministic function, $r' = F[r]$. Because R' is a reduced code, it always conveys less information about the stimulus than the full code: $I(S; R') \leq I(S; R)$. This difference in

information can be rewritten as an average of KL divergences in a form that is suggestive of the measure $\hat{D}$. However, it is important to keep in mind that a neural code that ignores noise correlations by combining spike trains from shuffle trials is not a reduced version of the real neural code. The shuffled responses, R_{shuffle}, include the entire set of responses found in the simultaneous responses, R, but they occur with different probabilities. There is no deterministic function that can act on r on every trial to produce r_{shuffle}. Therefore, the information lost by constructing a reduced neural code is not directly relevant to the case of ignoring noise correlations.

The subtlety of what it means to "ignore correlations" can be seen in a simple example. Suppose that we observe a set of signals $\{y_1, y_2, \ldots y_N\}$, all of which are linearly related to some interesting signal x. There are many simple situations in which our best estimate of x (e.g., the estimate that makes the smallest mean square error) is just a linear combination or weighted sum of the y_i, that is $x_{\mathrm{est}} = \Sigma_i w_i y_i$. Such a "decoder" obviously does not detect correlations among the y_i in any explicit way, because there is no term approximate to $y_i y_j$ that would be analogous to detecting synchronous spikes form different neurons. In contrast, the optimal values of the weights w_i depend in detail on the signal and noise correlations among the y_i (as is relevant for the linear decoding of spike trains discussed below). Is this implicit dependence sufficient to say that the linear decoder makes use of correlations? Or does it ignore correlations because it does not explicitly detect them? Even if we can resolve these ambiguities in simple linear models, would our definitions of what it means to ignore correlations be sufficiently general that they could be applied to arbitrary neural responses? These difficulties simply do not arise in the discussion of synergy and redundancy from an information theoretic point of view.

Series expansion of the mutual information

Panzeri et al. (1999) have presented an approximation of the information conveyed by a population of neurons based on a series expansion, in which successive terms correspond to different orders of correlation functions. The first-order term is equal to the information in the time varying firing rate of each cell, and the three second-order terms involve correlation functions among pairs of spikes. This expansion is in the same spirit as expansion series for the case of single neurons (DeWeese, 1995; Brenner et al., 2000). Within the series expansion of Panzeri et al., second-order terms that add and subtract to the synergy were identified and related to signal and noise correlations (Panzeri et al., 1999; Petersen et al., 2001). One second-order term, which depends only on signal correlations, gives rise only to redundancy; another second-order term, which depends only on noise correlations, gives rise only to synergy. The final second-order term, which mixes signal and noise correlations, can be either positive or negative.

This expansion relies on the assumption that the firing rate is low enough or that the sampled time bin is short enough that the probability of finding a spike in each time bin is much less than one. By truncating the expansion at second order, this method neglects correlations among more than two spikes, regardless of whether these spikes are from the same cell or two different cells (Panzeri et al., 1999). The authors show that, under some conditions, this second-order expansion is a good approximation to the fully sampled information (Petersen et al., 2001). One should keep in mind, though, that the adequacy of this expansion depends on the neural system under study as well as the ensemble of stimuli (Bezzi et al., 2002). To verify such adequacy, one must either show that contributions from higher-order terms are small or show that the second-order expansion gives results close to those of direct sampling (Strong et al., 1998; Reinagel and Reid, 2000). Of course, if direct sampling can be achieved, it is not clear what is gained by a second-order expansion. If instead, higher-order correlations cannot be adequately sampled, then the bias introduced by ignoring these terms may be smaller than the bias introduced by sampling them poorly.

Recently, Pola et al. (2003) generalized this approach beyond second order in correlation functions. They use analogous terms, in which full probability functions are substituted for correlation functions. Interestingly, the authors show that this breakdown of the mutual information into four terms is exact, meaning that no additional terms are necessary to sum to the joint mutual information. For the case of two neurons, the four terms can be written in the following form (see Appendix B for a derivation of these equations from the expressions used in Pola et al., 2003):

$$I(S; R_1, R_2) = I_{\mathrm{lin}} + \tilde{I}_{\mathrm{sig-sim}} + \tilde{I}_{\mathrm{cor-ind}} + \tilde{I}_{\mathrm{cor-dep}}, \quad (24)$$

where these terms are

$$\begin{aligned} I_{\mathrm{lin}} &= I(S; R_1) + I(S; R_2), \\ \tilde{I}_{\mathrm{sig-sim}} &= -\Delta I_{\mathrm{signal}}, \\ \tilde{I}_{\mathrm{cor-ind}} &= \Delta I_{\mathrm{noise}} - \hat{D}, \text{ and} \\ \tilde{I}_{\mathrm{cor-dep}} &= \hat{D}. \end{aligned} \quad (25)$$

Pola et al. (2003) give the following interpretation of the terms. I_{lin} is the information conveyed if the two neurons carry independent information. $\tilde{I}_{\mathrm{sig-sim}}$ expresses the loss of information because of similarity in the responses of the two cells averaged over the stimulus. It can only give rise to redundancy, as can be seen from its form as a KL divergence (see Appendix B). $\tilde{I}_{\mathrm{cor-ind}}$ gives a contribution to the joint information from the interaction between cross-cell correlation and signal similarity; its values can be either positive or negative. $\tilde{I}_{\mathrm{cor-dep}}$ gives a contribution attributable to stimulus-dependent correlations and can only be positive. $\tilde{I}_{\mathrm{cor-ind}}$ and $\tilde{I}_{\mathrm{cor-dep}}$ are both zero if the cells are conditionally independent. This decomposition is similar to the one we presented above. The only difference is that Pola et al. (2003) have resolved $\Delta I_{\mathrm{noise}}$ into $\tilde{I}_{\mathrm{cor-ind}}$ and $\tilde{I}_{\mathrm{cor-dep}}$.

However, it is important to note that the noise correlation terms used by Pola et al. (2003) are not themselves mutual informations. Obviously, because $\tilde{I}_{\mathrm{cor-dep}}$ equals $\hat{D}$, it suffers from all of the same problems of interpretation that we showed above. Namely, it is not a loss of mutual information and can be greater than the joint mutual information. Similarly, $\tilde{I}_{\mathrm{cor-ind}}$ is also not an information or information loss. This can be seen be examining its form in more detail (see Appendix B). As a consequence, $\tilde{I}_{\mathrm{cor-ind}}$ can contribute redundancy >100%; for the toy model in Figure 6, $\tilde{I}_{\mathrm{cor-ind}}/I(S; R_1, R_2) = -202\%$. As such, it is difficult to interpret this term.

One should note that there is no unique way to decompose the mutual information into a series of terms (DeWeese, 1995). The breakdown proposed by Pola et al. (2003) generalizes contributions to the mutual information that arose in the second-order expansion described previously by Panzeri et al. (1999). Rather than asking for the second-order approximation to the information, one could ask exactly for the information carried by spike pairs (Brenner et al., 2000). In the same spirit, our approach is based on decompositions of the synergy that either set bounds on synergy or redundancy (Eq. 11) or follow from the consequences

of ignoring noise correlations (Eq. 16). In each case, there is a well-posed question answered by the value of each term, giving each term a clear information-theoretic interpretation. These terms are not derived from an expansion to the joint mutual information and they do not individually exceed the joint mutual information as either a loss or a gain. In contrast, the noise correlation terms proposed by Pola et al. (2003) can individually have contradictory values, making their interpretation unclear.

Linear decoding

Linear decoding has been used to study how two or more cells jointly convey information. Warland et al. (1997) constructed linear decoding filters that were simultaneously optimized for two or more retinal ganglion cells stimulated with spatially uniform flicker. This method includes second-order correlations between pairs of spikes. They found that cells of different functional type roughly added their information, whereas cells of the same functional type were redundant. They also found that the filters for two or more cells optimized together were significantly different than the filters for each cell optimized alone. This suggests that signal correlations alone can alter the optimal decoding strategy, even if they do not give rise to synergy or redundancy.

Dan et al. (1998) studied the importance of a prominent kind of noise correlation between neurons in the LGN: the increased tendency of cells to fire spikes synchronously on the same stimulus trial (Alonso et al., 1996). The authors allowed synchronous spike pairs to have a different decoding filter than a spike from either cell occurring by itself. Thus, the spike trains of a pair of cells had three different neural symbols: synchronous spikes from cells A and B; a spike from cell A but not B; a spike from cell B but not A. They found that the linear decoder that simultaneously optimized these three filters extracted, on average, 20% more information than the decoder that assigned a filter for each cell, regardless of whether spikes were synchronous or not.

However as these authors note, the information estimated by linear decoding is a lower bound on the encoded information. Strictly speaking, one cannot conclude anything about the relationship between two quantities with knowledge of the lower bound on each. Thus, it is possible that a more sophisticated decoder could achieve even greater synergy for synchronous spikes in the LGN or remove the redundancy of retinal ganglion cells of the same functional type. Conversely, such a decoder might give more total information than a linear decoder but reveal that synchronous spikes in the LGN are redundant or that retinal ganglion cells of different functional types are also redundant. These possibilities indicate why it is useful to calculate the information theoretic quantities proposed here and establish the significance of correlations between pairs of neurons for information transmission in a manner that is not dependent on the choice of decoding algorithm.

Discussion

Here, we have presented an information theoretic framework for assessing the importance of correlated firing in the transmission of information by pairs of neurons. We have shown that there are three different notions of independence or lack of correlation: activity, conditional, and informational independence. For each notion of independence, there is a corresponding information theoretic quantity that measures the degree of correlation. These three kinds of independence are distinct in the sense that any one independence can hold without implying either of the other two. But these quantities are interrelated: the synergy of two neurons can be expressed as the difference between the measures of conditional and activity independence. In addition, the synergy can be written as the difference of contributions of signal and noise correlations. Although the synergy, thus, plays a central role in the characterization of population encoding, at least two measures of independence must be calculated to provide a complete description.

Although cross-correlation functions are more intuitive and more commonly used, we have relied on information theoretic tools here because they are more general measures of correlation and require only a minimal set of mathematical assumptions. Indeed, our approach overcomes known difficulties with the interpretations of cross-correlation values and gives a consistent set of measures for the description and interpretation of population encoding.

Our measures of redundancy and independence address fundamental issues in neural coding. Barlow (1961) and Attneave (1954) have proposed that redundancy reduction is a key attribute of efficient coding schemes, because populations of neurons with this property do not waste their representational capacity with repeated messages. In this sense, an efficient code is one in which all pairs of neurons are nonredundant, $Syn(R_1, R_2) \geq 0$. Similarly, one can think of an efficient code as one in which all pairs of neurons possess activity independence, $I(R_1; R_2) = 0$, because this condition also guarantees a lack of redundancy.

Such notions of efficient coding have been related to the structure of receptive fields in the retina, because the mechanism of center-surround antagonism can remove the dominant spatial correlations present in natural visual stimuli (Attneave, 1954; Barlow, 1961). Using the requirement of activity independence, researchers have made quantitative predictions of the linear filter characteristics of ganglion cells in the vertebrate retina (Atick, 1992) and second-order neurons in the fly retina (Srinivasan et al., 1982; van Hateren, 1992). Despite the success of these retinal theories, natural visual scenes possess correlations beyond those captured by their power spectrum (Ruderman, 1994). Similar ideas have been used to explore how redundancy reduction might be performed in the presence of such higher-order correlations. Independent component analysis can, in part, take these correlations into account, producing components that approximate activity independence and resemble V1 receptive fields (Bell and Sejnowski, 1997; Hyvarinen and Hoyer, 2001). Plausible forms of nonlinear gain control, perhaps in area V1, can also serve to remove these correlations (Schwartz and Simoncelli, 2001).

However, as a design principle for the neural systems, redundancy reduction is a problematic objective. Among other issues, efficient codes are very sensitive to noise and require careful processing when one wishes to extract the encoded information. It is not clear that compressed representations are valuable for the brain, because the existence of redundancy may actually signify the prevalence of the stimulus. Barlow revisited the possible role of redundancy in neural systems recently (Barlow, 2001) and suggested that redundancy may actually play an important role in the representation and analysis of probability distributions in neural systems.

We have made a distinction between encoding and decoding questions. Information theory is a powerful tool for the analysis of encoding questions, and it gives a bound on what may be decoded from neural responses. However, to evaluate the performance of a decoder, one must specify the cost of each of the possible errors that the decoder might make. Because mutual information is only sensitive to the correspondence between spike trains and stimuli, it cannot, by itself, characterize the quality of a decoder. As a result, questions about decoding cannot be put on the same level of generality as questions about encoding.

Because there typically is not an obvious choice for an error

measure, one would like to be able to draw conclusions about decoding that do not depend on a specific choice of error measure or decoding algorithm. We have discussed such an attempt by Nirenberg et al. (2001), who studied the impact of noise correlations on decoding. They tried to find a general answer to this question by comparing the difference between the true decoding dictionary, $p(s|r_1, r_2)$, and an approximate dictionary that explicitly ignored noise correlations. But, as we have shown, the measure proposed by Nirenberg et al. (2001) is not an information loss, does not assess the encoding properties of neurons, and is not the answer to the question what one may lose about the stimulus, if noise correlations are ignored by a decoder. For making statements about decoding that do not rely on a chosen error measure, one can compare different decoding dictionaries, but it should be kept in mind that there is no single measure with which to make this comparison and that such measures are not necessarily related to the encoded information. Consequently, the question of whether retinal ganglion cells are independent encoders is still open, as is the question about the effect of noise correlation on decoding their activity.

The general question of what can be achieved with a nonoptimal or approximated decoder is a very important one. In most cases, it is not clear that the nervous system or even the experimenter can learn or use the "right" decoder or even a good approximation. Moreover, biological constraints might limit the classes of possible decoders that might be learned or implemented. We emphasize again that the answer to this question relies principally on analysis of the possible decoding errors (Wu et al., 2003). Information theory is useful for decoding in conjunction with an error measure. Given a decoding algorithm that, for example, minimizes the L_2 norm between s and s^{est}, one can analyze the decoding errors and place bounds on $I(S; S^{\text{est}})$, as discussed for the reconstruction of time-dependent signals by Bialek et al. (1991). The mutual information between stimuli and estimated stimuli is a lower bound on the mutual information between stimuli and responses, and so one can proceed from error-based decoding to a statement about information transmission (Rieke et al., 1997).

For specific error functions or decoding schemes, one can place bounds on the extracted information (Lin, 1991; Samengo, 2002). For continuously valued stimuli, one can use the Fisher information (Cover and Thomas, 1991), which is not actually an information measure in the Shannon sense, to place a bound on the mean squared error between stimulus and estimated stimulus (Seung and Sompolinsky, 1993; Zhang et al., 1998; Abbott and Dayan, 1999; Sompolinsky et al., 2001). The Fisher information can be used to place a lower bound on the encoded information, but this bound is tight only under specialized conditions (Brunel and Nadal, 1998; Kang and Sompolinsky, 2001).

Rather than asking directly what neurons encode and what may be decoded, information theoretic measures have also been used to quantify the difference between neural responses. Comparing the responses of different cells to the same stimulus, one can ask how much information you gain about the identity of a cell by observing its response. The answer to this question is given by the Jensen–Shannon divergence (Lin, 1991), and this measure is useful for classifying cell types or comparing the encoding properties of individual cells (Schneidman et al., 2001, 2003). Alternatively, Johnson et al. (2001) have suggested using information theoretic measures of dissimilarity (the KL divergence, Chernoff "distance", or a "resistor-average" version of the KL divergence) to compare the distributions of neural responses to two different stimuli, $p(r|s_1)$ to $p(r|s_2)$. Relying on known results from classification theory (Hogg and Craig, 1995), these quantities can either bound or predict the asymptotic behavior of the error that one might make in assessing which of the two stimuli was presented by observing the neural responses. This approach would be especially useful for distinguishing stimulus features that result in different neural responses from those that do not. Note, however, that these measures correspond to the decoding error only for the simple case of two stimuli. One should also keep in mind that these measures do not relate directly to the encoded information.

As discussed here, there are many possible information theoretic measures with which to evaluate neural codes. However, they can differ importantly in their interpretation. In particular, each measure is the answer to one or more specific questions about the neural code. We hope that a clearer understanding of this relationship will help to resolve ongoing debates about the nature of neural codes in different circuits, animals, and species.

Appendix A: Proof that $\Delta I_{\text{signal}} \geq 0$

$$\begin{aligned}\Delta I_{\text{signal}} &= I(S; R_1) + I(S; R_2) - I_{\text{shuffle}}(S; R_1, R_2)\\ &= \sum_s p(s) \sum_{r_1} p(r_1|s) \log_2 \frac{p(r_1|s)}{p(r_1)}\\ &+ \sum_s p(s) \sum_{r_2} p(r_2|s) \log_2 \frac{p(r_2|s)}{p(r_2)}\\ &- \sum_s p(s) \sum_{r_1, r_2} p(r_1|s)\, p(r_2|s) \log_2 \frac{p(r_1|s)\, p(r_2|s)}{\sum_{s'} p(s')\, p(r_1|s')\, p(r_2|s')}.\end{aligned}$$

Inserting $\Sigma_{r_i} p(r_i|s) = 1$ and grouping terms gives:

$$\begin{aligned}\Delta I_{\text{signal}} &= \sum_s p(s) \sum_{r_1, r_2} p(r_1|s)\, p(r_2|s)\\ &\times \left[\log_2 \frac{p(r_1|s)}{p(r_1)} + \log_2 \frac{p(r_2|s)}{p(r_2)} - \log_2 \frac{p(r_1|s)\, p(r_2|s)}{\sum_{s'} p(s')\, p(r_1|s')\, p(r_2|s')} \right].\end{aligned}$$

Defining $p_{\text{shuffle}}(r_1, r_2) = \Sigma_{s'} p(s')\, p(r_1|s')\, p(r_2|s')$ gives:

$$\begin{aligned}\Delta I_{\text{signal}} &= \sum_{r_1, r_2} \sum_s p(s)\, p(r_1|s)\, p(r_2|s) \log_2 \frac{\sum_{s'} p(s')\, p(r_1|s')\, p(r_2|s')}{p(r_1)\, p(r_2)}\\ &= \sum_{r_1, r_2} p_{\text{shuffle}}(r_1, r_2) \log_2 \frac{p_{\text{shuffle}}(r_1, r_2)}{p(r_1)\, p(r_2)}.\end{aligned}$$

This quantity has the form of a KL divergence, which is non-negative:

$$\Delta I_{\text{signal}} = D_{\text{KL}}[p_{\text{shuffle}}(r_1, r_2) \| p(r_1)\, p(r_2)] \geq 0. \qquad (26)$$

Appensix B: Generalization of the Second-order Expansion of the mutual Information

The second-order expansion of the mutual information performed by Panzeri and co-workers (Panzeri et al., 1999; Panzeri and Schultz, 2001) makes an explicit connection between second-order correlation functions and contributions to the mutual in-

formation. They defined two types of correlation functions. The first, denoted by $\gamma(t_i, t_j; s)$, measures noise correlations:

$$\gamma(t_i, t_j; s) = \frac{\langle r_1(t_i; s) r_2(t_j; s)\rangle}{\langle r_1(t_i, s)\rangle \langle r_2(t_j; s)\rangle} - 1, \quad (27)$$

where $r(t; s)$ is the firing rate at time t in response to stimulus s, and the average $\langle\rangle$ is over repeated stimulus trials. The second correlation function, denoted $\nu(t_i, t_j)$, measures signal correlations:

$$\nu(t_i, t_j) = \frac{\langle \bar{r}_1(t_i; s)\bar{r}_2(t_j; s)\rangle_s}{\langle \bar{r}_1(t_i; s)\rangle_s \langle \bar{r}_2(t_j; s)\rangle_s} - 1, \quad (28)$$

where $\bar{r}(t; s)$ is the firing rate averaged over stimulus trials, and the average $\langle\rangle_s$ is over stimuli.

To generalize beyond second-order correlations, Pola et al. (2003) replaced their correlation functions between pairs of spikes with analogous ratios of probability distributions over all possible neural responses. The new measure of noise correlations was:

$$\gamma(r_1, r_2; s) = \frac{p(r_1, r_2|s)}{p(r_1|s)\, p(r_2|s)} - 1, \quad (29)$$

and the new measure of signal correlations was

$$\nu(r_1, r_2) = \frac{p(r_1, r_2)}{p(r_1)\, p(r_2)} - 1. \quad (30)$$

They inserted these functions into the expressions that they defined previously for the second-order contributions to the mutual information, also making the replacement of $p(r_1|s)$ for $r_1(t; s)$ and $p(r_2|s)$ for $r_2(t; s)$:

$$\tilde{I}_{\text{sig}-\text{sim}} = \frac{1}{\ln(2)} \sum_{r_1, r_2} p(r_1) p(r_2) \,[\nu(r_1, r_2) - (1 + \nu(r_1, r_2)) \ln(1 + \nu(r_1, r_2))]$$

$$\tilde{I}_{\text{cor}-\text{ind}} = -\sum_s p(s) \sum_{r_1, r_2} p(r_1|s) p(r_2|s) \gamma(r_1, r_2|s) \log_2(1 + \nu(r_1, r_2))$$

$$\tilde{I}_{\text{cor}-\text{dep}} = \sum_s p(s) \sum_{r_1, r_2} p(r_1|s) p(r_2|s) \,(1 + \gamma(r_1, r_2|s)) \times \log_2\left[\frac{p_{\text{shuffle}}(r_1, r_2)\,(1 + \gamma(r_1, r_2|s))}{\sum_{s'} p(s') p(r_1|s') p(r_2|s')\,(1 + \gamma(r_1, r_2|s'))}\right], \quad (31)$$

where $p_{\text{shuffle}}(r_1, r_2) = \Sigma_{s'} p(s')\, p(r_1|s')\, p(r_2|s')$.

Taking each contribution to the information separately, we substitute the expressions for $\gamma(r_1, r_2; s)$ and $\nu(r_1, r_2)$ (Eq. 29 and 30) and rearrange terms:

$$\tilde{I}_{\text{sig}-\text{sim}} = \frac{1}{\ln(2)} \sum_{r_1, r_2} [p_{\text{shuffle}}(r_1, r_2) - p(r_1) p(r_2)] - \sum_{r_1, r_2} p_{\text{shuffle}}(r_1, r_2) \log_2\left[\frac{p_{\text{shuffle}}(r_1, r_2)}{p(r_1) p(r_2)}\right], \quad (32)$$

the first term cancels to zero once summed over responses, and the second term can be seen to have the form of a KL divergence:

$$\tilde{I}_{\text{sig}-\text{sim}} = -D_{\text{KL}}[p_{\text{shuffle}}(r_1, r_2) \| p(r_1) p(r_2)], \quad (33)$$

using Equation 26 from Appendix A, we get

$$\tilde{I}_{\text{sig}-\text{sim}} = -\Delta I_{\text{signal}}. \quad (34)$$

Next, for the stimulus-independent noise correlation contribution:

$$\begin{aligned}\tilde{I}_{\text{cor}-\text{ind}} &= -\sum_s p(s) \sum_{r_1, r_2} [p(r_1, r_2|s) - p(r_1|s) p(r_2|s)] \log_2(1 + \nu(r_1, r_2)) \\ &= \sum_{r_1, r_2} [p_{\text{shuffle}}(r_1, r_2) - p(r_1, r_2)] \log_2\left[\frac{p_{\text{shuffle}}(r_1, r_2)}{p(r_1) p(r_2)}\right] \\ &= \sum_{r_1, r_2} p_{\text{shuffle}}(r_1, r_2) \log_2\left[\frac{p_{\text{shuffle}}(r_1, r_2)}{p(r_1) p(r_2)}\right] \\ &\quad - \sum_{r_1, r_2} p(r_1, r_2) \log_2\left[\frac{p(r_1, r_2)}{p(r_1) p(r_2)}\right] \\ &\quad + \sum_{r_1, r_2} p(r_1, r_2) \log_2\left[\frac{p(r_1, r_2)}{p_{\text{shuffle}}(r_1, r_2)}\right]. \end{aligned} \quad (35)$$

All three of these terms are KL divergences that have appeared here. Using equations 26, 6, and 20, respectively, we can substitute for all three:

$$\begin{aligned}\tilde{I}_{\text{cor}-\text{ind}} &= \Delta I_{\text{signal}} - I(R_1; R_2) + (\langle I(R_1; R_2|s)\rangle_s - \hat{D}) \\ &= \Delta I_{\text{signal}} + \text{Syn}(R_1, R_2) - \hat{D} \\ &= \Delta I_{\text{noise}} - \hat{D}. \end{aligned} \quad (36)$$

Finally, for the stimulus-dependent noise correlation contribution:

$$\begin{aligned}\tilde{I}_{\text{cor}-\text{dep}} &= \sum_s p(s) \sum_{r_1, r_2} p(r_1, r_2|s) \log_2\left[\frac{p_{\text{shuffle}}(r_1, r_2)}{p(r_1, r_2)} \frac{p(r_1, r_2|s)}{p(r_1|s) p(r_2|s)}\right] \\ &= \sum_{r_1, r_2} p(r_1, r_2) \sum_s p(s|r_1, r_2) \log_2\left[\frac{p(s|r_1, r_2)}{p_{\text{shuffle}}(s|r_1, r_2)}\right] \\ &= \hat{D}. \end{aligned} \quad (37)$$

References

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

Abeles M, Lass Y (1975) Transmission of information by the axon: II. The channel capacity. Biol Cybern 19:121–125.

Abeles M, Bergman H, Margalit E, Vaadia E (1993) Spatiotemporal firing patterns in the frontal cortex of behaving monkeys. J Neurophysiol 70:1629–1638.

Aertsen A, Gerstein G, Habib M, Palm G (1989) Dynamics of neuronal firing correlation: modulation of "effective connectivity." J Neurophysiol 61:900–917.

Alonso J, Usrey W, Reid R (1996) Precisely correlated firing in cells of the lateral geniculate nucleus. Nature 383:815–819.

Atick J (1992) Could information theory provide an ecological theory of sensory processing? Netw Comput Neural Syst 3:213–251.

Attneave F (1954) Some informational aspects of visual perception. Psychol Rev 61:183–193.

Bair W, Koch C (1996) Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey. Neural Comp 8:1185–1202.

Barlow H (1961) Possible principles underlying the transformation of sensory messages. In: Sensory communication (Rosenblith WA, ed). Cambridge, MA: MIT.

Barlow H (2001) Redundancy reduction revisited. Network: Comput Neural Syst 12:241–253.

Bell A, Sejnowski T (1997) The "independent components" of natural scenes are edge filters. Vision Res 37:3327–3338.

Berry M, Meister M (1998) Refractoriness and neural precision. J Neurosci 16:2381–2396.

Bezzi M, Diamond M, Treves A (2002) Redundancy and synergy arising from pairwise correlations in neuronal ensembles. J Comput Neurosci 12:165–174.

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

Brenner N, Strong S, Koberle R, Bialek W, de Ruyter van Steveninck R (2000) Synergy in a neural code. Neural Comput 12:1531–1552.

Brody C (1999) Correlations without synchrony. Neural Comput 11:1537–1551.

Brown E, Frank L, Tang D, Quirk M, Wilson M (1998) A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. J Neurosci 18:7411–7425.

Brunel N, Nadal J (1998) Mutual information, Fisher information, and population coding. Neural Comput 10:1731–1757.

Buracas G, Zador A, DeWeese M, Albright T (1998) Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. Neuron 20:959–969.

Chechik G, Globerson A, Anderson M, Young E, Nelken I, Tishby N (2002) Group redundancy measures reveal redundancy reduction along the auditory pathway. In: Advances in neural information processing systems 14 (Dietterich TG, Becker S, Ghahramani Z, eds), pp 173–180. Cambridge, MA: MIT.

Cover T, Thomas J (1991) Elements of information theory. New York: Wiley Interscience.

Dan Y, Alonso J, Usrey W, Reid R (1998) Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus. Nat Neurosci 1:501–507.

Dawid A (1979) Conditional independence in statistical theory. J R Stat Soc B 41:1–31.

Dayan P, Abbott L (2001) Theoretical neuroscience. Cambridge, MA: MIT.

de Ruyter van Stevenick R, Bialek W (1988) Real-time performance of a movement sensitive neuron in the blowfly visual system: coding and information transfer in short spike sequences. Proc R Soc Lond B Biol Sci 234:379–414.

DeWeese M (1995) Optimization principles for the neural code. PhD thesis, Princeton University.

Eckhorn R, Popel B (1974) Rigorous and extended application of information theory to the afferent visual system of the cat: I. Basic concepts. Kybernetik 16:191–200.

Fitzhugh R (1957) The statistical detection of threshold signals in the retina. J Gen Physiol 40:925–948.

Gat I, Tishby N (1999) Synergy and redundancy among brain cells of behaving monkeys. In: Advances in neural processing systems 11 (Kearns M, Solla S, Cohn D, eds), pp 465–471. Cambridge, MA: MIT.

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

Georgopoulos A, Schwartz A, Kettner R (1986) Neuronal population coding of movement direction. Science 233:1416–1419.

Gray C, Singer W (1989) Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. Proc Natl Acad Sci USA 86:1698–1702.

Hatsopoulos N, Ojakangas C, Paninski L, Donoghue J (1998) Information about movement direction obtained from synchronous activity of motor cortical neurons. Proc Natl Acad Sci USA 95:15706–15711.

Hogg R, Craig A (1995) Introduction to mathematical statistics Ed 5. Englewood Cliffs, NJ: Prentice-Hall.

Hyvarinen A, Hoyer P (2001) A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. Vision Res 41:2413–2423.

Johnson D, Gruner C, Baggrely K, Sesahgiri C (2001) Information-theoretic analysis of neural coding. J Comp Neurosci 16:2381–2396.

Kang K, Sompolinsky H (2001) Mutual information of population codes and distance measures in probability space. Phys Rev Lett 86:4958–4961.

Krahe R, Kreiman G, Gabbiani F, Koch C, Metzner W (2002) Stimulus encoding and feature extraction by multiple sensory neurons. J Neurosci 22:2374–2848.

Laurent G, Davidowitz H (1994) Encoding of olfactory information with oscillating neural assemblies. Science 265:1872–1875.

Lewen G, Bialek W, de Ruyter van Steveninck R (2001) Neural coding of naturalistic motion stimuli. Network: Comput Neural Syst 12:317–329.

Lin J (1991) Divergence measures based on the Shannon entropy. IEEE Trans Inf Theory 37:145–151.

Mackay D, McCulloch W (1952) The limiting information capacity of a neuronal link. Bull Math Biophys 14:127–135.

Mainen Z, Sejnowski T (1995) Reliability of spike timing in neocortical neurons. Science 268:1503–1508.

Martignon L, Deco G, Laskey K, Diamond M, Freiwald W, Vaadia E (2000) Neural coding: higher-order temporal patterns in the neurostatistics of cell assemblies. Neural Comput 16:2381–2396.

Mastronarde D (1983) Correlated firing of cat retinal ganglion cells. I. Spontaneously active inputs to X- and Y-cells. J Neurophysiol 49:303–324.

Meister M, Hosoya T (2001) Are retinal ganglion cells independent encoders? Preprint. See also http://rhino.harvard.edu/Publications/Meister_2001_Encoders.pdf.

Meister M, Legando L, Baylor D (1995) Concerted signaling by retinal ganglion cells. Science 270:1207–1210.

Nemenman I, Bialek W, de Ruyter van Steveninck R (2003) Entropy and information in neural spike trains: progress on the sampling problem. arXiv:physics/0306063.

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

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

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

Palm G, Aertsen A, Gerstein G (1988) On the significance of correlations among neuronal spike trains. Biol Cybern 59:1–11.

Paninski L (2003) Estimation of entropy and mutual information. Neural Comput 15:1191–1253.

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

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

Perkel D, Bullock T (1968) Neural coding. Neurosci Res Program Bull 6:221–343.

Perkel D, Gerstein G, Moore G (1967) Neuronal spike trains and stochastic point processes. II. Simultaneous spike trains. Biophys J 7:419–440.

Petersen R, 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: Comput Neural Syst 14:35–60.

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

Reinagel P, Reid R (2000) Temporal coding of visual information in the thalamus. J Neurosci 20:5392–5400.

Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W (1997) Spikes: exploring the neural code. Cambridge, MA: MIT.

Ruderman D (1994) The statistics of natural images. Network: Comput Neural Syst 5:517–548.

Salinas E, Abbott L (1994) Vector reconstruction from firing rates. J Comput Neurosci 1:89–107.

Samengo I (2002) Information loss in an optimal maximum likelihood decoding. Neural Comput 14:771–779.

Schneidman E, Brenner N, Tishby N, de Ruyter van Steveninck R, Bialek W (2001) Universality and individuality in a neural code. Adv Neural Inf Process Syst 13:159–165.

Schneidman E, Bialek W, Berry M (2003) An information theoretic ap-

proach to the functional classification of neurons. In: Advances in neural information processing systems 15 (Becker S, Thrun S, Obermayer K, eds), pp 197–204. Cambridge, MA: MIT.
Schneidman E, Still S, Berry M, Bialek W (2003) Network information and connected correlations. Phys Rev Lett, in press.
Schwartz O, Simoncelli E (2001) Natural signal statistics and sensory gain control. Nat Neurosci 4:819–825.
Seung H, Sompolinsky H (1993) Simple models for reading neuronal population codes. Proc Natl Acad Sci USA 90:10749–10753.
Shannon C, Weaver W (1949) The mathematical theory of communication. Urbana, IL: University of Illinois.
Sompolinsky H, Yoon H, Kang K, Shamir M (2001) Population coding in neuronal systems with correlated noise. Phys Rev E 64:8095–8100.
Srinivasan M, Lauglin S, Dubs A (1982) Predictive coding: a fresh view of inhibition in the retina. Proc R Soc Lond B 216:427–459.
Strong S, Koberle R, de Ruyter van Steveninck R, Bialek W (1998) Entropy and information in neural spike trains. Phys Rev Lett 80:197–200.
Treves A, Panzeri S (1995) The upward bias in measures of information derived from limited data samples. Neural Comput 7:399–407.
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.
van Hateren J (1992) Real and optimal neural images in early vision. Nature 360:68–70.
Verveen A, Derksen H (1968) Fluctuation phenomena in nerve membrane. Proc IEEE 56:906–916.
Victor J (2002) Binless strategies for estimation of information from neural data. Phys Rev E 66:051903.
Warland D, Reinagel P, Meister M (1997) Decoding visual information from a population of retinal ganglion. J Neurophysiol 78:2336–2350.
Wu S, Chen D, Nirajan M, Amari S (2003) Sequential bayesian decoding with a population of neurons. Neural Comput 15:993–1012.
Zemel R, Dayan P, Pouget A (1998) Probabilistic interpretation of population codes. Neural Comput 10:403–430.
Zhang K, Ginzburg I, McNaughton B, Sejnowski T (1998) Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. J Neurophysiol 79:1017–1044.
Zohary E, Shadlen M, Newsome W (1994) Correlated neuronal discharge rate and its implications for psychophysical performance. Nature 370:140–143.