Behavioral/Systems/Cognitive

Synergy, Redundancy, and Independence in Population Codes, Revisited

Peter E. Latham[1] **and Sheila Nirenberg**[2]

[1]Gatsby Computational Neuroscience Unit, University College London, London WC1N 3AR, United Kingdom, and [2]Department of Neurobiology, University of California at Los Angeles, Los Angeles, California 90095-1763

Decoding the activity of a population of neurons is a fundamental problem in neuroscience. A key aspect of this problem is determining whether correlations in the activity, i.e., noise correlations, are important. If they are important, then the decoding problem is high dimensional: decoding algorithms must take the correlational structure in the activity into account. If they are not important, or if they play a minor role, then the decoding problem can be reduced to lower dimension and thus made more tractable. The issue of whether correlations are important has been a subject of heated debate. The debate centers around the validity of the measures used to address it. Here, we evaluate three of the most commonly used ones: synergy, $\Delta I_{\text{shuffled}}$, and ΔI. We show that synergy and $\Delta I_{\text{shuffled}}$ are confounded measures: they can be zero when correlations are clearly important for decoding and positive when they are not. In contrast, ΔI is not confounded. It is zero only when correlations are not important for decoding and positive only when they are; that is, it is zero only when one can decode exactly as well using a decoder that ignores correlations as one can using a decoder that does not, and it is positive only when one cannot decode as well. Finally, we show that ΔI has an information theoretic interpretation; it is an upper bound on the information lost when correlations are ignored.

Key words: **retina; encoding; decoding; neural code; information theory; population coding; signal correlations; noise correlations**

Introduction

One of the main challenges we face in neuroscience is understanding the neural code; that is, understanding how information about the outside world is carried in neuronal spike trains. Several possibilities exist: information could be carried in spike rate, spike timing, spike correlations within single neurons, spike correlations across neurons, or any combination of these.

Recently, a great deal of attention has been focused on the last possibility, on spike correlations cross neurons (e.g., synchronous spikes). The reason for the strong emphasis on this issue is that its resolution has significant impact on downstream research: whether or not correlations are important greatly affects the strategies one can take for population decoding. If correlations are important, then direct, brute force approaches are ruled out: one simply cannot find the mapping from stimulus to response, as such a mapping would require measuring response distributions in high dimensions, a minimum of N dimensions for N neurons. For more than three or four neurons, the amount of data needed to do this becomes impossibly large, and the direct approach becomes intractable (Fig. 1*a*). Instead, indirect methods for estimating response distributions, such as modeling the correlations parametrically, must be used.

Received Dec. 31, 2004; revised March 11, 2005; accepted March 29, 2005.

P.E.L. was supported by the Gatsby Charitable Foundation and National Institute of Mental Health Grant R01 MH62447. S.N. was supported by National Eye Institute Grant R01 EY012978. We acknowledge Peter Dayan and Liam Paninski for insightful discussion and comments on this manuscript.

Correspondence should be addressed to Sheila Nirenberg, Department of Neurobiology, University of California at Los Angeles, 10833 Le Conte Avenue, Los Angeles, CA 90095. E-mail: sheilan@ucla.edu.

DOI:10.1523/JNEUROSCI.5319-04.2005

 0270-6474/05/255195-12$15.00/0

If, on the other hand, correlations turn out not to be important, then direct approaches can be used, even for large populations. This is because one can build the mapping from stimulus to response for a population of neurons from the individual, single neuron mappings (Fig. 1*b*). Such an approach would allow rapid movement on the question of how neuronal activity is decoded.

The issue of whether correlated firing is important has been fraught with debate. Several authors (Eckhorn et al., 1988; Gray and Singer 1989; Gray et al., 1989; Meister et al., 1995; Vaadia et al., 1995; deCharms and Merzenich, 1996; Dan et al., 1998; Steinmetz et al., 2000) have suggested that they are, whereas others (Nirenberg et al., 2001; Oram et al., 2001; Petersen et al., 2001; Levine et al., 2002; Panzeri et al., 2002a,b; Averbeck and Lee, 2003; Averbeck et al., 2003; Golledge et al., 2003) have argued that they are not or that they play a minor role. The debate has arisen in large part because different methods have been used to assess the role of correlations, and different methods yield different answers.

One early method was to look for stimulus-dependent changes in cross-correlograms (Eckhorn et al., 1988; Gray and Singer, 1989; Gray et al., 1989; Vaadia et al., 1995; deCharms and Merzenich, 1996). This method, however, has two problems. One is that firing rates can significantly alter the shape of cross-correlograms, making it difficult to separate information carried in firing rates from information carried in correlations. The other is that cross-correlograms only tell us about one type of correlation, synchronous or near-synchronous spikes. Correlations that occur on a longer timescale, or correlations among patterns of spikes, are missed by this method.

More recently, information-theoretic approaches have been

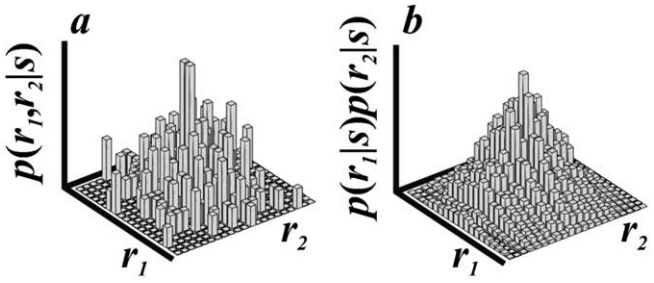

Figure 1. Estimating conditional response distributions from data. ***a***, Estimate of the correlated response distribution, $p(r_1, r_2|s)$, for a single stimulus, s. The responses, r_1 and r_2, are taken to be spike count in a 300 ms window. They range from 0 to 19, so there are 400 ($= 20 \times 20$) bins. A total of 250 trials were used, which leads to a very noisy estimate. ***b***, Estimate of the independent response distribution, $p(r_1|s)p(r_2|s)$. The single neuron distributions, $p(r_1|s)$ and $p(r_2|s)$, can be estimated individually, using all 250 trials for each one, and the joint distribution can then be constructed by multiplying them together. This leads to a much smoother (and more accurate) estimate of the probability distribution.

applied to the problem, because they are more quantitative and are sensitive to correlations other than just synchrony. These methods, however, also turned out to have problems. In particular, two measures that have appeared in the literature, $\Delta I_{\text{shuffled}}$ (Nirenberg and Latham 1998; Panzeri et al., 2001; Golledge et al., 2003; Osborne et al., 2004) and synergy/redundancy (Brenner et al., 2000; Liu et al., 2001; Machens et al., 2001; Schneidman et al., 2003), seem intuitive but, in fact, turn out to be confounded. Here, we describe a measure that is not confounded and, in fact, provides an upper bound on the importance of correlations for decoding. In addition, we show why the other two methods, $\Delta I_{\text{shuffled}}$ and synergy/redundancy, can lead one astray.

Results

Definition of correlations

Correlations in neuronal responses arise from two sources. One is the stimulus: if multiple neurons see the same stimulus, then their responses will be related. For example, if a flash of light is presented to the retina, all of the ON retinal ganglion cells will tend to fire at flash onset and all of the OFF cells at flash offset. On average, ON cells will be correlated with ON cells, OFF cells will be correlated with OFF cells, and ON and OFF cells will be anticorrelated with each other.

These stimulus-induced correlations are typically referred to as "signal correlations" (Gawne and Richmond, 1993) and are defined as follows: responses from N neurons, denoted, r_i, $i = 1, \ldots, N$, are signal correlated if and only if

$$p(r_1, r_2, \ldots, r_N) \neq \prod_i p(r_i),$$

where $p(r_1, r_2, \ldots, r_N)$ and $p(r_i)$ are the joint and single neuron response distributions averaged over stimuli. These distributions are given by the standard relationships $p(r_1, r_2, \ldots, r_N) \equiv \Sigma_s p(r_1, r_2, \ldots, r_N|s)p(s)$ and $p(r_i) \equiv \Sigma_s p(r_i|s)p(s)$, where s is the stimulus. Here and in what follows, the response from neuron i, r_i, is essentially arbitrary; it could be firing rate, spike count, or a binary string indicating when a neuron did and did not fire.

The second source of correlations is common input, which can arise from either a common presynaptic source (e.g., two ON ganglion cells that receive input from the same amacrine cell) or direct or indirect interaction between neurons (e.g., gap junction coupling). Correlations of this type are called "noise correlations" (Gawne and Richmond, 1993), and they differ from signal correlations in that they are a measure of the response correlations on a stimulus-by-stimulus basis. Specifically, responses are noise correlated if and only if

$$p(r_1, r_2, \ldots, r_N|s) \neq \prod_i p(r_i|s).$$

A population of neurons will almost always exhibit a mix of signal and noise correlations. For example, two ON ganglion cells far apart on the retina (two cells that share no circuitry) will exhibit no noise correlations, but they will exhibit signal correlations, so long as the stimulus has sufficiently long-range spatial correlations to make them fire together. In contrast, two ON cells with overlapping receptive fields (two cells that do share circuitry) will exhibit both signal and noise correlation.

It is undisputed that signal correlations are important for decoding, that is, for inferring stimuli from responses. Our brains are built to take correlations in the outside world and reflect them in correlations in neuronal responses. What is not clear, however, is whether noise correlations are important for decoding. It is these that have been the subject of debate, and it is these that we focus on in this paper.

Testing whether correlations are important

The most straightforward way to test whether noise correlations are important for decoding is to build a decoder that does not take them into account and compare its performance with one that does (Dan et al., 1998; Wu et al., 2000, 2001; Nirenberg et al., 2001; Nirenberg and Latham, 2003; Averbeck and Lee, 2003, 2004). If the decoder that does not take correlations into account performs as well as the one that does, then correlations are not important for decoding. If it does not perform as well, then they are.

To construct a decoder that takes correlations into account (and, because we do not know the algorithm the brain uses, we assume optimal decoding), we first record neuronal responses to many stimuli and build the response distribution, $p(\mathbf{r}|s)$. [Here, $\mathbf{r} \equiv (r_1, r_1, \ldots, r_N)$ is shorthand for the responses from all N neurons.] We then use Bayes' theorem to construct the probability distribution of stimuli given responses, yielding

$$p(s|\mathbf{r}) = \frac{p(\mathbf{r}|s)p(s)}{p(\mathbf{r})}.$$

We will take the approach that $p(s|\mathbf{r})$ is our decoder, although in practice one often takes the additional step of choosing a particular stimulus from this distribution, such as the one that maximizes $p(s|\mathbf{r})$.

To construct a decoder that does not take correlations into account, we perform essentially the same steps we used to construct $p(s|\mathbf{r})$. The only difference is that our starting point is the independent response distribution rather than the true one—the response distribution one would build with knowledge of the single neuron distributions but no knowledge of the correlations. This distribution, denoted $p_{\text{ind}}(\mathbf{r}|s)$, is

$$p_{\text{ind}}(\mathbf{r}|s) = \prod_i p(r_i|s). \quad (1)$$

Given $p_{\text{ind}}(\mathbf{r}|s)$, we can then construct the "independent" stimulus distribution, $p_{\text{ind}}(s|\mathbf{r})$, from Bayes' theorem,

$$p_{\text{ind}}(s|\mathbf{r}) = \frac{p_{\text{ind}}(\mathbf{r}|s)p(s)}{p_{\text{ind}}(\mathbf{r})},$$

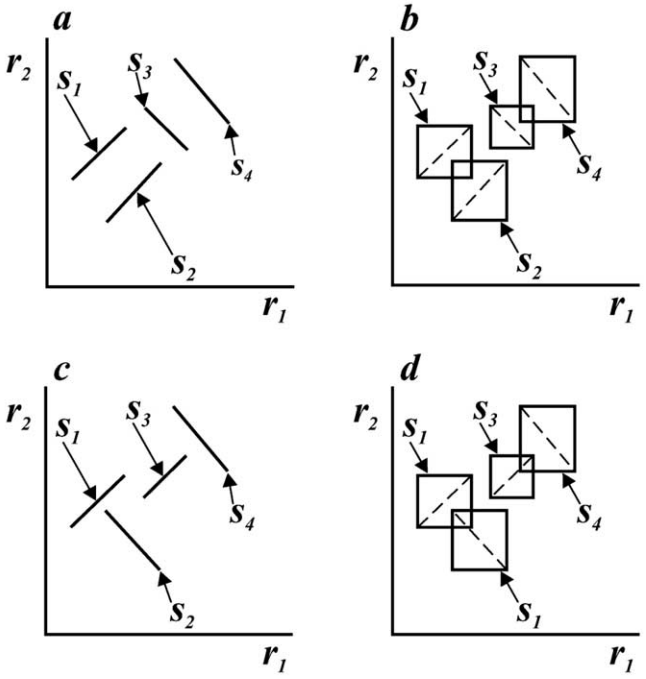

Figure 2. Correlations can exist without being important for decoding. ***a***, Correlated response distributions for four stimuli, shown as solid lines. For each stimulus, the responses lie along the line segments indicated by the arrows. [Formally, $p(r_1, r_2|s_i) \propto \delta(s_i - (r_2 - a_i r_1))$, where a_i is the slope of the line segment, and there is an implicit cutoff when r_1 is below some minimum or above some maximum.] If the stimulus is known, r_1 predicts r_2 and vice versa, making the responses perfectly correlated. Because the responses form disjoint sets, all responses are uniquely decodable. ***b***, Independent response distributions for the same four stimuli, shown as open boxes (the correlated distributions are shown also, as dashed lines). For each stimulus, the responses lie inside the boxes indicated by the arrows. The boxes overlap, and, if a response were to occur in the overlap region, it would not be uniquely decodable, because it could have been produced by more than one stimulus. However, the responses never land in this region (because they always land on the dashed lines). Thus, a decoder built with no knowledge of the correlational structure would be able to decode the true responses perfectly. ***c***, A very similar set of correlated distributions. A decoder with knowledge of the correlations would be able to decode all responses perfectly. ***d***, The independent response distributions derived from ***c***. The true responses can lie in the overlap region, and a decoder that had no knowledge of the correlational structure would not be able to decode such responses perfectly. Thus, the correlations here are clearly important for decoding.

where $p_{\text{ind}}(\mathbf{r}) \equiv \Sigma_s p_{\text{ind}}(\mathbf{r}|s)p(s)$ is the total independent response distribution. By construction, $p_{\text{ind}}(s|\mathbf{r})$ does not use any knowledge of the noise correlations.

To assess the role of correlations, we simply ask the decoders to decode the true responses (the responses that were actually recorded from the animal) and assess how well they do. Specifically, we take responses from simultaneously recorded neurons and compute both $p(s|\mathbf{r})$ and $p_{\text{ind}}(s|\mathbf{r})$. If they are the same for all stimuli and responses that could occur, then we know that we do not have to take correlations into account when we decode; if they are different for at least one stimulus–response pair, then we know that we do.

In Figure 2*a*, we perform this procedure for a pair of neurons with correlated responses. Although the responses are, in fact, highly correlated, the correlations do not matter: if one goes through each of the true responses, one can see that they can be decoded exactly as well using the decoder built from the independent response distribution as they can using the decoder built from the true response distribution. Or, expressed in terms of probabilities, $p_{\text{ind}}(s|\mathbf{r}_1, \mathbf{r}_2) = p(s|\mathbf{r}_1, \mathbf{r}_2)$ for all responses that can occur. This demonstrates a key point: cells can be highly correlated without those correlations being important for decoding.

Of course, cells can also be correlated with the correlations being important. An example of this is illustrated in Figure 2*c*, which shows a pair of neurons whose correlational structure is very similar to that shown in Figure 2*a*, but different enough so that $p_{\text{ind}}(s|\mathbf{r}_1, \mathbf{r}_2) \neq p(s|\mathbf{r}_1, \mathbf{r}_2)$. In this case, knowledge of correlations is necessary to decode correctly, so correlations are important for decoding.

Dan et al. (1998) were the first ones we know of to assess the role of correlations by building decoders that do and do not take correlations into account: they asked whether a decoder with no knowledge of synchronous spikes would do worse than one with such knowledge. Wu et al. (2000, 2001) later extended the idea so that it could be applied to essentially arbitrary correlational structures, not just synchronous spikes, and, recently, we extended it further and developed an information-theoretic cost function that measured the importance of correlations (Nirenberg et al., 2001; Nirenberg and Latham, 2003). Below we show that this cost function provides an upper bound on the information one loses by ignoring correlations. First, however, we show that the other information-theoretic measures that have been proposed do not do this.

Other approaches

Other approaches for assessing the importance of correlations have been proposed. In this section, we consider two of the most common ones, $\Delta I_{\text{shuffled}}$ and $\Delta I_{\text{synergy}}$.

Shuffled information

The first measure we consider is $\Delta I_{\text{shuffled}}$ (see Eq. 2 below). This measure emerged from an approach similar to the one described in the previous section, in the sense that the overall idea is to assess the importance of correlations by removing them and looking for an effect. The difference, however, is in how we look. In the previous section, we looked for an effect by building two decoders, one using the true responses and one using the independent ones. The two decoders are then asked to decode the true responses, and their performance is compared. In the $\Delta I_{\text{shuffled}}$ approach, the same two decoders are built. What is different, however, is what is decoded: the true decoder is asked to decode the true responses, and the independent one is asked to decode the independent responses. As we will see, this seemingly small difference in what is decoded has a big effect on the outcome.

The quantitative measure associated with this approach, $\Delta I_{\text{shuffled}}$, is the difference between the information one obtains from the true responses and the information one obtains from the independent responses (i.e., the "shuffled" responses, whose name comes from the fact that, in experiments, the independent responses are produced by shuffling trials). This difference is given by

$$\Delta I_{\text{shuffled}} = I(s; \mathbf{r}) - I_{\text{shuffled}}(s; \mathbf{r}). \quad (2)$$

Here, $I(s; \mathbf{r})$ is the mutual information between stimuli and responses (Shannon and Weaver, 1949),

$$I \equiv -\sum_{\mathbf{r}} p(\mathbf{r}) \log_2 p(\mathbf{r}) + \sum_s p(s) \sum_{\mathbf{r}} p(\mathbf{r}|s) \log_2 p(\mathbf{r}|s), \quad (3)$$

and $I_{\text{shuffled}}(s; \mathbf{r})$ is defined analogously, except that $p(\mathbf{r}|s)$ is replaced by $p_{\text{ind}}(\mathbf{r}|s)$ (Eq. 1) and $p(\mathbf{r})$ by $p_{\text{ind}}(\mathbf{r})$. Specifically,

$$I_{\text{shuffled}}(s; \mathbf{r}) \equiv -\sum_{\mathbf{r}} p_{\text{ind}}(\mathbf{r}) \log_2 p_{\text{ind}}(\mathbf{r}) + \sum_{s} p(s) \sum_{\mathbf{r}} p_{\text{ind}}(\mathbf{r}|s) \log_2 p_{\text{ind}}(\mathbf{r}|s).$$

Because I is computed with knowledge of correlations and I_{shuffled} is computed without this knowledge, one might expect that when $\Delta I_{\text{shuffled}} \neq 0$, correlations are important, and when $\Delta I_{\text{shuffled}} = 0$, they are not. This expectation, however, is not correct. The reason is that when one computes $\Delta I_{\text{shuffled}}$, one is computing information from a response distribution that never occurred. As a result, one can end up measuring information about a stimulus from responses that the brain never sees.

A corollary of this is that I_{shuffled} can actually be larger than I. This is troublesome in light of the numerous proposals that correlations can act as an extra channel of information (Eckhorn et al., 1988; Gray and Singer, 1989; Gray et al., 1989; Vaadia et al., 1995; Meister et al., 1995; deCharms and Merzenich, 1996; Steinmetz et al., 2000), because, if they do, removing them should lead to a loss of information rather than a gain. Also troublesome are the observations that $\Delta I_{\text{shuffled}}$ can be nonzero even when a decoder does not need to have any knowledge of the correlations to decode the true responses, and it can be zero when a decoder does need to have this knowledge (see below).

To gain deeper insight into what $\Delta I_{\text{shuffled}}$ does and does not measure, let us consider two simple examples, shown in Figure 3. In both examples, correlations are not important; that is, $p_{\text{ind}}(s|r_1, r_2) = p(s|r_1, r_2)$ for all responses that can occur, but, in one case $\Delta I_{\text{shuffled}}$ is positive and in the other $\Delta I_{\text{shuffled}}$ is negative.

In the first example, the true (correlated) response distribution is shown in Figure 3*a*, and its independent (shuffled) counterpart is shown in Figure 3*b*. It is not hard to see why $\Delta I_{\text{shuffled}}$ is positive in this case: when the responses are correlated, they are disjoint and thus perfectly decodable (Fig. 3*a*). When they are made independent, however, they are no longer disjoint; an overlap region is produced, and responses that land in that region could have been caused by either stimulus (Fig. 3*b*). Thus, the responses in the independent case provide less information about the stimuli than the responses in the correlated case. A straightforward calculation shows that $\Delta I_{\text{shuffled}} = \frac{1}{4}$ (see Appendix A).

This example emphasizes why $\Delta I_{\text{shuffled}}$ is a misleading measure for decoding. Because the amount of information in the shuffled responses is less than that in the true responses, one gets the impression that the decoder built from the shuffled responses would not perform as well as the one built from the true responses. In reality, however, it does perform as well: every one of the true responses is decoded exactly the same using the two decoders. This is reflected in the fact that $p_{\text{ind}}(s|r_1, r_2)$ is equal to $p(s|r_1, r_2)$ for all responses that actually occur.

This is an important point. When one takes a set of correlated responses and makes them independent, one creates a new response distribution. However, the fact that this new response distribution, $p_{\text{ind}}(r_1, r_2|s)$, is not equal to the true one, $p(r_1, r_2|s)$, does not imply the reverse, that $p_{\text{ind}}(s|r_1, r_2) \neq p(s|r_1, r_2)$. In fact, as the above example indicates and as we showed in Figure 2*a*, it is easily possible to have $p_{\text{ind}}(s|r_1, r_2) = p(s|r_1, r_2)$ when $p_{\text{ind}}(r_1, r_2|s) \neq p(r_1, r_2|s)$. This is a surprising finding, and something that would not have been (could not have been) revealed by $\Delta I_{\text{shuffled}}$.

In the second example, the true response distribution is shown in Figure 3*c*, and its independent counterpart is shown in Figure 3*d*. Here, $\Delta I_{\text{shuffled}} < 0$: when the responses are correlated, they land in the overlap region (the region in which the responses could have been caused by either stimulus) on one-half the trials (Fig. 3*c*), whereas when they are independent, they land in the overlap region on one-quarter of the trials (Fig. 3*d*). Consequently, when the responses are independent, they provide more information on average (because they are ambiguous less often), and a straightforward calculation yields $\Delta I_{\text{shuffled}} = -\frac{1}{4}$ (see Appendix A). It is also easy to show that a decoder does not need to know about these correlations: regardless of whether a decoder assumes the neurons are uncorrelated, responses in the overlap region provide no information about the stimulus and responses not in the overlap region are decoded perfectly (see Appendix A). Thus, as with the previous example, the fact that $\Delta I_{\text{shuffled}}$ is negative is easy to misinterpret: it gives the impression that the correlations are important when they are not.

It is not hard, by extending these examples, to find cases in which $\Delta I_{\text{shuffled}} = 0$ when correlations actually are important [Nirenberg and Latham (2003), their supporting information].

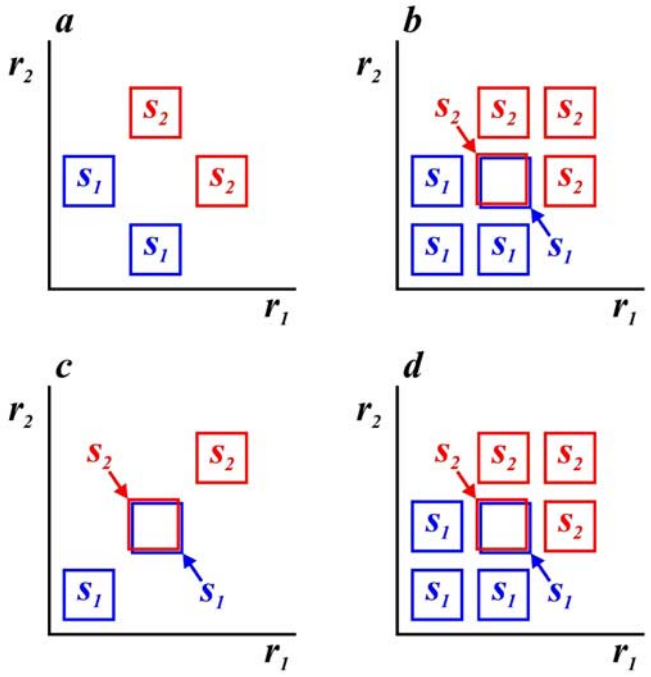

Figure 3. $\Delta I_{\text{shuffled}}$ can be both positive and negative when correlations are not important for decoding. ***a***, Correlated response distributions for two stimuli, s_1 and s_2, which occur with equal probability. For each stimulus, the responses fall inside the boxes labeled by that stimulus. Because the responses are disjoint, all are uniquely decodable. ***b***, Independent response distributions for the same stimuli. (The center boxes are offset so both can be seen.) Responses in the center box, which occur on one-quarter of the trials, provide no information about the stimulus. Thus, the independent responses provide less information than the true responses (because they are sometimes ambiguous), so $\Delta I_{\text{shuffled}} > 0$ (see Appendix A). However, as with Figure 2*a*, the true responses never land in the ambiguous region, so a decoder that has no knowledge of the correlations will decode exactly as well as one that has full knowledge of them. ***c***, A different set of correlated response distributions, also for two stimuli. In this case, the responses land in the center box on one-half of the trials, and thus there is ambiguity about what the stimulus is on one-half the trials. ***d***, The corresponding independent distribution (which is the same as in ***b***). Here, the independent responses are ambiguous on only one-quarter of the trials, so they provide more information about the stimulus than the true responses. Thus, for this example, $\Delta I_{\text{shuffled}} < 0$ (see Appendix A). However, regardless of whether a decoder knows about the correlations, if a response lands in the overlap region, the stimulus probabilities are the same (one-half for each), so, as in ***a***, knowledge of the correlations is not necessary for decoding.

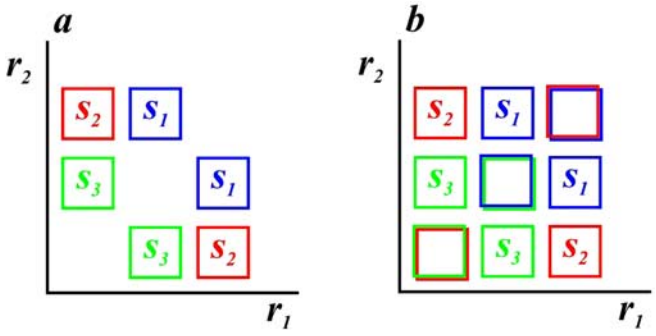

Figure 4. A synergistic code in which correlations are not important for decoding. ***a***, Correlated response distributions for three stimuli, s_1, s_2, and s_3, which occur with equal probability. For each stimulus, the responses fall inside the boxes. Because the responses form disjoint sets, all responses are uniquely decodable. For this distribution, it is not hard to show that $\Delta I_{\text{synergy}} = \log_2(4/3)$ (see Appendix A). ***b***, Independent response distributions for the same stimuli. (The boxes along the diagonal are offset so both can be seen.) If a response were to land in a box along the diagonal, it would not be uniquely decodable; it could have been produced by two stimuli. However, as with Figures 2*a* and 3*a*, the responses never occur along the diagonal. Thus, even if a decoder knew nothing at all about the correlational structure, it would be able to decode perfectly all responses that actually occur (which are the only ones that matter). The probability distributions for this figure were derived from Schneidman et al. (2003); see Discussion.

This is because the shuffled information can be positive for some parts of the code and negative for others, producing cancellations that make $\Delta I_{\text{shuffled}} = 0$. In fact, all combinations are possible: $\Delta I_{\text{shuffled}}$ can be positive, negative, or zero both when correlations are important and when they are not [Nirenberg and Latham (2003), their supporting information], making this quantity a bad one for assessing the role of correlations in the neural code.

This is not to say that $\Delta I_{\text{shuffled}}$ is never useful; it can be used to answer the question: given a correlational structure, would more information be transmitted using that structure or using independent responses (Abbott and Dayan, 1999; Sompolinsky et al., 2001; Wu et al., 2002)? This is interesting from a theoretical point of view, because it sheds light on issues of optimal coding, but it is a question about what could be rather than what is.

Synergy and redundancy

Another common, but less direct, measure that has been proposed to assess the role of correlations is the synergy/redundancy measure, denoted $\Delta I_{\text{synergy}}$ (Brenner et al., 2000; Machens et al., 2001; Schneidman et al., 2003). It is defined to be

$$\Delta I_{\text{synergy}} = I(s; \mathbf{r}) - \sum_i I(s; r_i), \qquad (4)$$

where I, the mutual information, is given in Equation 3.

Positive values of $\Delta I_{\text{synergy}}$ are commonly assumed to imply that correlations are important, a claim made explicitly by Schneidman et al. (2003). This claim, however, breaks down with close examination, for essentially the same reason as in the previous section: $\Delta I_{\text{synergy}}$ fails to take into account that $p_{\text{ind}}(\mathbf{r}|s)$ can assign nonzero probability to responses that do not occur. To see this explicitly, we will consider an example in which $\Delta I_{\text{synergy}}$ is positive but correlations are not important; that is, $p_{\text{ind}}(s|r_1, r_2) = p(s|r_1, r_2)$ for all true responses.

The example is illustrated in Figure 4. Figure 4*a* shows the correlated distribution. As one can see, the responses form a disjoint set, so every pair of responses corresponds to exactly one stimulus. Figure 4*b* shows the independent distribution. Here, the responses do not form a disjoint set. That is because responses along the diagonal, in which both neurons produce approximately the same output, can be caused by more than one stimulus. On the surface, then, it appears that, without knowledge of the correlated distribution, one cannot decode all of the responses perfectly. However, this is not the case: as in Figures 2*a* and 3*a*, the responses that could have been caused by more than one stimulus (the ones along the diagonal) never happen. Thus, all responses that do occur can be decoded perfectly. This means that we can decode exactly as well with no knowledge of the correlational structure as we can with full knowledge, even though $\Delta I_{\text{synergy}} > 0$.

If $\Delta I_{\text{synergy}}$ can be positive when one can decode optimally with no knowledge of the correlations [that is, when $p_{\text{ind}}(s|\mathbf{r}) = p(s|\mathbf{r})$], then what does synergy really tell us? To answer this, note that it is a general feature of population codes that observing more neurons provides more information. How much more, however, spans a large range and depends in detail on the neural code. At one end are completely redundant codes, for which observing more neurons adds no information (for example, Fig. 3*c*). At the other end are synergistic codes, for which observing more neurons results in a large increase in information. Thus, the degree of synergy (the value of $\Delta I_{\text{synergy}}$) tells us where along this range a neural code lies. It does not, however, tell us about the importance of correlations.

Although we focused here on showing that $\Delta I_{\text{synergy}}$ can be positive when correlations are not important for decoding, it can also be shown, and has been shown in previous work, that $\Delta I_{\text{synergy}}$ can be negative or zero when correlations are not important [when $p_{\text{ind}}(s|\mathbf{r}) = p(s|\mathbf{r})$]. Likewise, it can be positive, negative, or zero when correlations are important [when $p_{\text{ind}}(s|\mathbf{r}) \neq p(s|\mathbf{r})$] [Nirenberg and Latham (2003), their supporting information]. Thus, $\Delta I_{\text{synergy}}$ is not a very useful measure for assessing the importance of correlations for decoding.

Redundancy reduction

A long-standing proposal about early sensory processing is that one of its primary purposes is to reduce redundancy (Attneave, 1954; Barlow, 1961; Srinivasan et al., 1982; Atick and Redlich, 1990; Atick, 1992) (but see Barlow, 2001). Given that codes with $\Delta I_{\text{synergy}} < 0$ are referred to as "redundant," one might interpret this proposal to mean that $\Delta I_{\text{synergy}}$ should be maximized, so that the code exhibits as little redundancy as possible. Unfortunately, redundant has two meanings. One is "not synergistic" ($\Delta I_{\text{synergy}} < 0$). The other, as originally defined by Shannon and Weaver (1949), is "not making full use of a noisy channel." Under the latter definition, redundancy, denoted $\mathcal{R}$, is given by

$$\mathcal{R} \equiv 1 - \frac{I}{C},$$

where I is, as above, mutual information, and C is channel capacity: the maximum value of the mutual information with respect to $p(s)$ for fixed $p(\mathbf{r}|s)$.

The redundancy-reduction hypothesis refers to minimizing $\mathcal{R}$, not maximizing $\Delta I_{\text{synergy}}$. This is sensible: minimizing $\mathcal{R}$ corresponds to making the most efficient use of a channel. Maximizing $\Delta I_{\text{synergy}}$ also seems sensible on the surface, because maximum synergy codes can be highly efficient [in fact, they can transmit an infinite amount of information (Fig. 5)]. However, biological constraints prevent their implementation, and they are almost always effectively impossible to decode (Fig. 5*b*). Thus, maximization principles involving $\Delta I_{\text{synergy}}$ are unlikely to yield insight into the neural code.

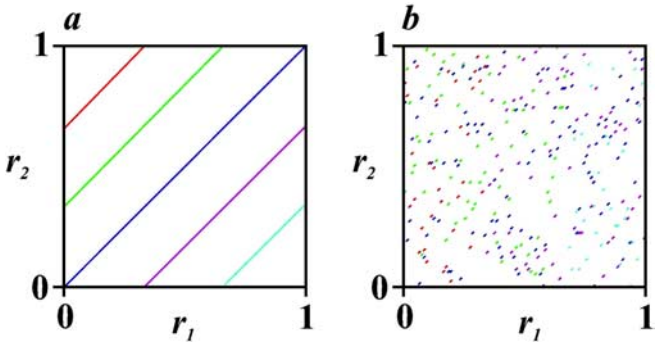

Figure 5. Highly synergistic codes are efficient but typically very difficult to decode. ***a***, Two-neuron response distribution for five stimuli, color coded for clarity. The distribution is of the form $p(s|r_1, r_2) \propto \delta(s - (r_1 - r_2))$, with both r_1 and r_2 restricted to lie between 0 and 1. The stimulus, s, is a continuous variable that is uniformly distributed between -1 and 1. Observing both responses tells us exactly what the stimulus is, so the responses provide an infinite amount of information (the stimulus is specified with infinite precision). Observing any one response, however, provides only a finite amount of information about the stimulus. Consequently, $\Delta I_{\text{synergy}} = \infty$. This coding scheme is advantageous because it can transmit an infinite amount of information and is easy to decode ($s = r_1 - r_2$). Note, however, that it requires perfect correlation, which is not biologically plausible. ***b***, A distribution in which the $r_1 - r_2$ plane was scrambled: it was divided into a 100×100 grid, and the squares in each column were randomly permuted. If we knew the scrambling algorithm, we could decode responses perfectly: $p(s|r_1, r_2) \propto \delta(s - \mathcal{U}(r_1, r_2))$, where $\mathcal{U}$ is an operator that transforms r_1 and r_2 in scrambled coordinates to $r_1 - r_2$ in unscrambled ones. However, the decoder would have to store $100 \log_2(100!)$ bits just to unscramble the responses ($\log_2(100!)$ bits per column), and, in general, for an $n \times n$ grid, it would have to store $n \log_2(n!) \sim n^2 \log_2(n/e)$ bits. Moreover, adding even a small amount of noise would destroy the ability of the responses to tell us anything about the stimulus. In the space of response distributions, non-smooth ones such as this are overwhelmingly more likely than the smooth one shown in ***a***. Thus, minimizing redundancy (which leads to maximum synergy) would almost always produce an encoding scheme that is virtually impossible to decode.

Quantifying the importance of correlations

Our analysis so far has led us to the following statement: correlations are important for decoding if $p_{\text{ind}}(s|\mathbf{r}) \neq p(s|\mathbf{r})$. However, what if we want to assess how important they are? Intuitively, we should be able to do this by computing the distance between $p_{\text{ind}}(s|\mathbf{r})$ and $p(s|\mathbf{r})$. If these two distributions are close, then correlations should be relatively unimportant; if they are far apart, then correlations should be important. The question we address now is: what do we mean by "close"?

In previous work (Nirenberg et al., 2001; Nirenberg and Latham, 2003), we argued that the appropriate measure of close is the Kullback-Leibler distance averaged over responses. This distance, which we refer to as ΔI, is given by

$$\Delta I = \sum_{\mathbf{r}} p(\mathbf{r}) \sum_{s} p(s|\mathbf{r}) \log_2 \frac{p(s|\mathbf{r})}{p_{\text{ind}}(s|\mathbf{r})}. \qquad (5)$$

[See also Panzeri et al. (1999) and Pola et al. (2003), who defined the same quantity, but used the notation $I_{\text{cor-dep}}$ instead of ΔI].

This measure has a number of desirable properties. First, because ΔI is weighted by $p(\mathbf{r})$, it does the sensible thing and weights responses by how likely they are to occur, with zero weight for responses that never occur (the ones the brain never sees). [This feature also takes care of the problem that $p(s|\mathbf{r})$ is undefined when $p(\mathbf{r}) = 0$.] Second, ΔI is bounded from below by zero, which makes it a nice candidate for a cost function. Third, it is zero if and only if $p_{\text{ind}}(s|\mathbf{r}) = p(s|\mathbf{r})$ for every response that actually occurs (Nirenberg and Latham, 2003), so $\Delta I = 0$ implies that correlations are completely unimportant for decoding. (It is not hard to show that $\Delta I = 0$ for the examples given in Figs. 2*a*, 3, 4; see Appendix A.) Fourth, it is a cost function that can be thought of in terms of yes/no questions, a common and intuitive way of expressing information theoretic quantities. Specifically, ΔI is the cost in yes/no questions for not knowing about correlations: if one were guessing the stimulus based on the neuronal responses, $\mathbf{r}$, then it would take, on average, ΔI more questions to guess the stimulus if one knew nothing about the correlations than if one knew everything about them (Nirenberg et al., 2001; Nirenberg and Latham, 2003).

The fourth property makes it possible to compare ΔI with the mutual information, I, between stimuli and responses, because I is the reduction in the average number of yes/no questions associated with observing neuronal responses (Cover and Thomas, 1991). The observation that ΔI can be expressed in terms of yes/no questions thus led us to identify the ratio $\Delta I/I$ as a measure of the relative importance of correlations. Here we solidify this identification by showing that ΔI is a rigorous upper bound on the information loss. This result is useful because it allows us to interpret the spate of recent experiments in which $\Delta I/I$ was found to be on the order of 10% or less (Nirenberg et al., 2001; Petersen et al., 2001, 2002; Panzeri et al., 2002b; Pola et al., 2003): in those experiments, one could ignore correlations when decoding and lose at most ~10% of the information.

ΔI is an upper bound on information loss

Showing that ΔI is an upper bound on information loss is not straightforward because classical information theory deals with true probability distributions. We, however, want to compute information when a decoder is based (via Bayes' theorem) on the wrong probability distribution: $p_{\text{ind}}(\mathbf{r}|s)$ rather than $p(\mathbf{r}|s)$. To do this, we use what is really a very simple idea, one that is closely related to discriminability. The idea is that if a decoder has knowledge of the full correlational structure, it will (typically) be able to make finer discriminations than if it does not, and so will be able to provide more information about the stimulus. For example, a decoder based on $p(\mathbf{r}|s)$ might have a discrimination threshold of, say, 1°, whereas one based on $p_{\text{ind}}(\mathbf{r}|s)$ might have a threshold that is twice as large, say 2°. The link between discrimination thresholds and information theory is that the factor of two decrease in the ability to discriminate implies a 1 bit decrease in information, because one-half as many orientations are distinguishable. For this example, then, the information loss associated with ignoring correlations, ΔI, is 1 bit.

To generalize this idea so that it can be applied to any stimulus set, not just simple ones, such as orientated bars, we adopt an information-theoretic construct known as a "codebook." In information theory, a codebook consists of a set of codewords, each of which is a list of symbols. These codewords are sent over a noisy channel to a "receiver," who also has access to the codebook. The job of the receiver is to examine the corrupted symbols and determine which codeword was sent. In our application, each symbol is a different stimulus, so a codeword consists of a set of stimuli sent in a particular order. For example, in the case of oriented bars, a codebook might consist of two length-3 codewords, one corresponding to bars that are presented sequentially at 2°, 1°, and 3°, and another to bars presented at 3°, 2°, and 1°. The job of the receiver is to examine the neuronal responses and determine which of the two orders was presented.

The reason we take this approach is that the number of codewords that can be put in the codebook before the receiver starts making mistakes is directly related to the mutual information between stimuli and responses. This follows from what is probably the central, and most profound, result in information theory, which is: if each codeword in the codebook consists of n stimuli,

then the upper bound on the number of codewords that can be sent almost error-free is 2^{nI}, where I is the mutual information between stimuli and responses (Shannon and Weaver, 1949; Cover and Thomas, 1991). Importantly, and this is the critical insight that allows us to relate ΔI to information loss, the upper bound depends on the probability distribution used by the receiver. If the receiver uses the true response distribution, $p(\mathbf{r}|s)$, to build a decoder, then the upper bound really is 2^{nI}. If, on the other hand, the receiver uses an incorrect response distribution, then the upper bound is typically smaller: $2^{nI^\star}$, where $I^\star \leq I$ (Merhav et al., 1994). The information loss associated with using the wrong distribution, which in our case is $p_{\text{ind}}(\mathbf{r}|s)$ rather than $p(\mathbf{r}|s)$, is $I - I^\star$.

Although this approach seems abstract, in fact, it is closely related to discriminability. For example, if orientations separated by 1° can almost always be distinguished, then we can put ~180 codewords of length 1 in our codebook (1°, 2°, . . . , 180°), 180^2 of length 2 [(1°, 1°), (1°, 2°), . . . , (180°, 180°)], and so on. If, on the other hand, orientations separated by ½° can almost always be distinguished, then the number of codewords would be 360 and 360^2, respectively. Thus, the number of codewords in our codebook tells us directly how far apart two orientations must be before they can be easily discriminated.

The advantage of the codebook/codeword approach over discrimination thresholds is that we can compute information loss simply by counting codewords. Here, we outline the main steps needed to do this; the details are provided in Appendix B. Because this is a very general approach (it can be used to evaluate information loss associated with any wrong distribution, not just one based on the independent responses), in what follows, we use $q(\mathbf{r}|s)$ to denote the wrong conditional response distribution rather than $p_{\text{ind}}(r|s)$. At the end, we can return to our particular problem by replacing $q(\mathbf{r}|s)$ with $p_{\text{ind}}(\mathbf{r}|s)$.

As discussed above, we will construct codewords that consist of n stimuli presented in a prespecified order that is known to the receiver, with each codeword having equal probability. We will consider discrete stimuli that come from an underlying distribution $p(s)$, and the stimuli that make up each codeword will be drawn independently from this distribution. We will use c to denote codewords and m to denote the number of codewords, so our codebook has the form

$$\begin{array}{lcccccc} c(1) & = & s_1(1) & s_2(1) & s_3(1) & \ldots & s_n(1) \\ c(2) & = & s_1(2) & s_2(2) & s_3(2) & \ldots & s_n(2) \\ \ldots & & & & & & \\ c(m) & = & s_1(m) & s_2(m) & s_3(m) & \ldots & s_n(m). \end{array}$$

For example, if the stimuli were orientations at even multiples of 1°, then a particular six-symbol codeword might be (2°, 10°, 7°, 1°, 14°, 2°).

The question we address here is: if the receiver uses $q(\mathbf{r}|s)$ to build a decoder, how large can we make m before the error rate becomes appreciable? The natural way to answer this is via Bayes' theorem: given a uniform prior on the codewords (a result of the fact that the codewords are sent with equal probability) and a set of observations, $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$, then the probability the receiver uses for a particular codeword, denoted $q(w|\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$, is

$$q(w|\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) \propto q(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n|s_1(w), s_2(w), \ldots, s_n(w))p(w) \propto \prod_i q(\mathbf{r}_i|s_i(w)), \quad (6)$$

where the last $\propto$ follows from our uniform prior, which is that $p(w)$ is independent of w. Given Equation 6, the optimal estimate of which message was sent, denoted $\hat{w}$, is

$$\hat{w} = \arg\max_w \left(\prod_{i=1}^{n} q(\mathbf{r}_i|s_i(w)) \right).$$

We can now compute the probability of making a decoding error. If message $w^\star$ is sent, then the probability, $P_e^{(n)}$, of an error for a particular codeword $w \neq w^\star$ is

$$P_e^{(n)} = \Pr\left(\prod_{i=1}^{n} q(\mathbf{r}_i|s_i(w)) > \prod_{i=1}^{n} q(\mathbf{r}_i|s_i(w^\star)) \right). \quad (7)$$

As we show in Appendix B, $P_e^{(n)}$ is independent of which codeword, $w^\star$, is sent. Thus, for any $w^\star$, the probability of making at least one error when we have m codewords in our codebook, denoted $P_{e,m}^{(n)}$, is

$$P_{e,m}^{(n)} = 1 - [1 - P_e^{(n)}]^m. \quad (8)$$

(Equation 8 should really have $m - 1$ in the exponent. However, because m is always large, here and in what follows, we will make no distinction between m and $m - 1$.)

To see how this equation gives us the upper bound on the transmitted information, let us define $I^\star$ via the relationship

$$P_e^{(n)} = 2^{-nI^\star}, \quad (9)$$

and let

$$m = 2^{n(I^\star - \epsilon)}, \quad (10)$$

where ϵ is a small, positive constant. Then, inserting Equations 9 and 10 into Equation 8, we find, after straightforward algebra, that

$$P_{e,m}^{(n)} = 2^{-n\epsilon} + \mathcal{O}(2^{-2n\epsilon}). \quad (11)$$

What Equation 11 tells us is that as $n \to \infty$ the probability of an error vanishes, no matter how small ϵ is. Consequently, $2^{nI^\star}$ is an upper bound on m and thus an upper bond on the number of codewords that can be sent with vanishingly small probability of error. This in turn implies that $I^\star$ is the information associated with using the wrong probability distribution. Using Equation 9 to express $I^\star$ in terms of $P_e^{(n)}$, we have

$$I^\star = \lim_{n\to\infty} \left[-\frac{1}{n} \log_2 P_e^{(n)} \right], \quad (12)$$

where $P_e^{(n)}$ is given by Equation 7.

Calculating $I^\star$ thus amounts to solving a math problem: computing $P_e^{(n)}$. This we do in Appendix B. Although there is no closed form expression for $I^\star$, we are able to show that

$$\Delta I \geq I - I^\star.$$

In other words, ΔI is an upper bound on the information loss associated with using the wrong distribution. When the wrong distribution is one that has no information about the correlations, then ΔI is an upper bound on the information loss associated with ignoring correlations.

Discussion

Determining whether correlations are important for decoding, both across spike trains from multiple neurons and within spike trains from single neurons, is a critical problem in neural coding. Addressing this problem requires a quantitative measure, one that takes as input neural data and produces as output a number that tells us how important correlations are. Efforts to develop such a measure, however, have led to a great deal of controversy, in large part because those efforts typically rely on intuitive notions. Two measures in particular, $\Delta I_{\text{shuffled}}$ and $\Delta I_{\text{synergy}}$, fall into this category. The first, $\Delta I_{\text{shuffled}}$, is the difference between the information carried by an ensemble of cells and the information that would be carried by the same cells if their responses were conditionally independent (Eq. 2). The second, $\Delta I_{\text{synergy}}$, is the difference between the information carried by an ensemble of cells and the sum of the information carried by the individual cells (Eq. 4). Both sound reasonable at first glance, but deeper analysis raises serious concerns. In particular, beyond the intuitive meaning that has been attached to these quantities, no rigorous link has been made to the importance of correlations for decoding. In addition, as we showed in Results, $\Delta I_{\text{shuffled}}$ and $\Delta I_{\text{synergy}}$ can take on essentially any values (from positive to zero to negative), both when one needs to know about correlations to decode and when one does not.

In this paper, we took a different strategy. Following the work of others (Dan et al., 1998; Panzeri et al., 1999; Wu et al., 2001, 2002), we approached the problem in the context of what is arguably the central question in neural coding, which is: how do we decode the activity of populations of neurons? Importantly, the answer depends almost exclusively on whether knowledge of correlations is necessary for constructing optimal decoding algorithms. If it is, then we will need to develop parametric models of the correlational structure in neuronal responses; if it is not, then we can use methods that ignore correlations and so are much simpler. We thus developed a measure based on the simple question: do we need to know about correlations to construct optimal decoding algorithms? We showed that the answer is determined by the value of ΔI (Eq. 5). Specifically, if $\Delta I = 0$, then knowledge of correlations is not necessary for optimal decoding, and if $\Delta I > 0$, then it is necessary, and its value places an upper bound on the amount of information lost by ignoring correlations. Thus, ΔI tells us whether correlations are important in a sense that has real and immediate impact on our strategies for decoding populations of neurons.

Recently, Schneidman et al. (2003) criticized this approach. Their criticism was, however, based on a single, very strong premise, which was that $\Delta I_{\text{synergy}}$ is the correct measure of the importance of correlations (see Schneidman et al., 2003, their Fig. 6 and associated text). Given this premise, they concluded that any measure that does not give the same answer as $\Delta I_{\text{synergy}}$ must be flawed. What was missing, however, was an explanation of why their premise is correct; as far as we know, neither they nor anyone else has demonstrated that $\Delta I_{\text{synergy}}$ is the correct measure of the importance of correlations.

Finally, we should point out that identifying a relevant measure (ΔI) is the first step in determining the role of correlations. Methods must be developed to apply this measure to populations of neurons. Here the underlying idea, which is that one can assess the role of correlations by building decoders that ignore them, will be just as useful as the measure itself. This is because one does not actually have to calculate ΔI [a difficult estimation problem, especially for population codes (Paninski, 2003, 2004; Nemenman et al., 2004)], but instead one can build decoders that do and do not take some aspect of correlations into account. If taking correlations into account improves decoding accuracy, then correlations are important for decoding; otherwise, they are not. This approach has already been used successfully for population decoding in motor cortex (Averbeck and Lee, 2003), and we expect it to be the method of choice in the future.

Appendix A: Calculation of ΔI, $\Delta I_{\text{shuffled}}$ and $\Delta I_{\text{synergy}}$

In this appendix, we compute $\Delta I_{\text{shuffled}}$ for the probability distributions shown in Figure 3, *a* and *c*, $\Delta I_{\text{synergy}}$ for the probability distribution shown in Figure 4*a*, and ΔI for both.

Our first step is to turn the continuous response distributions illustrated in these figures into discrete ones, which we can do because the relevant aspect of a response is which box it lands in. Thus, we will let both r_1 and r_2 take on integer values that range from 1 to 3, with the former labeling column and the later labeling row. With this scheme, the response $(r_1, r_2) = (2, 3)$, for example, refers to the box that is in the second column and the top row.

For simplicity, we will assume that the stimuli occur with equal probability. Thus, $p(s_1) = p(s_2) = 1/2$ in Figure 3 and $p(s_1) = p(s_2) = p(s_3) = 1/3$ in Figure 4. We will also assume that, given a stimulus, the responses that can occur, occur with equal probability. Loosely speaking, this means that all boxes of the same color are equally likely. For example, in Figure 3*a*, $p(2, 3|s_2) = p(3, 2|s_2) = 1/2$, and, in Figure 4*b*, $p(2, 2|s_1) = p(2, 3|s_1) = p(3, 2|s_1) = p(3, 3|s_1) = 1/4$.

Of the three information-theoretic quantities, ΔI (Eq. 5) is the easiest to compute, so we will start with it. Consider first the distributions in Figures 3, *a* and *b*, and 4, *a* and *b*. For these, all responses that actually occur (the ones in Figs. 3*a* and 4*a*, respectively) are uniquely decodable regardless of whether the decoder assumes independence, which is clear by examining Figures 3, *a* and *b*, and 4, *a* and *b*. Thus, $p_{\text{ind}}(s|r_1, r_2) = p(s|r_1, r_2)$, which implies, via Equation 5, that $\Delta I = 0$.

For the distributions in Figure 3, *c* and *d*, the situation is marginally more complex. The upper right and lower left responses, $(r_1, r_2) = (3, 3)$ and $(1, 1)$, respectively, are uniquely decodable whether are not the decoder assumes independence. The center response, $(r_1, r_2) = (2, 2)$, is not. Instead, that response gives no information about the stimulus, meaning that both s_1 and s_2 are equally likely, and this is true regardless of whether the decoder assumes independence (one can compute this directly, or use symmetry between s_1 and s_2). Thus, for all responses that actually occur (those in Fig. 3*c*), $p_{\text{ind}}(s|r_1, r_2) = p(s|r_1, r_2)$, and again $\Delta I = 0$.

Our next task is to compute $\Delta I_{\text{shuffled}}$ (Eq. 2) for Figure 3, *a*,*b* and *c*,*d*. Consider first Figure 3*a*. The responses in this figure are disjoint, so they are uniquely decodable. Thus, the mutual information is equal to the entropy of the stimulus, which is 1 bit. For the corresponding shuffled distribution, Figure 3*b*, recall that, for each stimulus, all squares of the same color occur with equal probability. Using this fact and examining Figure 3*b*, we see that, on 3/4 of the trials, the responses are uniquely decodable (they provide 1 bit of information), whereas on 1/4 of the trials, the responses provide no information. Thus, the responses convey, on average, 3/4 bits. This gives $\Delta I_{\text{shuffled}} = 1 - 3/4 = 1/4$.

Turning now to Figure 3*c*, we see that, on one-half of the trials, the responses provide 1 bit of information (because they are uniquely decodable), and, on the other half, they provide no information. Thus, the mutual information is 1/2 bits. Because

Figure 3*d* is the same as 3*b*, the mutual information of the shuffled distribution is again ¾ bits, and $\Delta I_{\text{shuffled}} = ½ - ¾ = -¼$.

Our last task is to compute $\Delta I_{\text{synergy}}$ (Eq. 4) for the distribution shown in Figure 4*a*. As in Figure 3*a*, the responses are uniquely decodable. Consequently, the mutual information, $I(s; r_1, r_2)$, is equal to the entropy of the stimulus, which is $\log_2 3$ bits. To compute the mutual information between one of the responses and the stimuli, note that receiving any one response reduces the number of possible stimuli from three to two. For example, if we observe that $r_1 = 2$, then we know that either s_1 or s_3 was presented, both with equal probability. Thus, the mutual information is $(\log_2 3 - \log_2 2)$ bits. Because $I(s; r_1) = I(s; r_2)$, it follows that $\Delta I_{\text{shuffled}} = \log_2 3 - 2(\log_2 3 - \log_2 2) = \log_2 (⅔)$.

Appendix B: Information-theoretic cost of ignoring correlations

The goal in this appendix is to compute $P_e^{(n)}$, from which we can calculate $I^\star$, the information associated with a decoder that uses the wrong distribution. This computation is divided into two parts. In the first, we derive a set of equations whose solution gives us $I^\star$. This is a rederivation of a result shown originally by Merhav et al. (1994) and is a straightforward application of large-deviation theory. We include the rederivation here both for completeness, and because we use a simpler (although slightly less rigorous) method than that of Merheve and colleagues. Unfortunately, the equations for $I^\star$ have no closed-form solution. However, they can be used to show that $I - I^\star$, the information loss associated with ignoring correlations, is bounded by ΔI. This is the focus of the second part.

Derivation of equations for $I^\star$

According to Equation 12, to compute $I^\star$ we first need to compute $P_e^{(n)}$, the latter given by Equation 7. Our first step, performed for convenience only, is to express $P_e^{(n)}$ in terms of sums of logs of probabilities rather than products of probabilities. This leads to the relationship

$$P_e^{(n)} = \Pr\left(\frac{1}{n}\sum_{i=1}^{n} \log_2 q(\mathbf{r}_i|s_i(w)) > \frac{1}{n}\sum_{i=1}^{n} \log_2 q(\mathbf{r}_i|s_i(w^\star))\right). \quad \text{(B1)}$$

The main idea behind the computation of $P_e^{(n)}$ is that the first term inside the parentheses is a random variable, and the probability that it is greater than the second can be computed using large-deviation theory. This is somewhat tricky because both terms in Equation B1 are random variables. Fortunately, it turns out that we can replace the second one by its average. Intuitively, this is because we are interested in the probability of an error for a single sent codeword, so outliers are not important. Placing this intuition on a rigorous footing, however, requires a fair amount of effort and is the subject of the next section. Those who are satisfied with the intuitive argument should skip directly to the section *Large-deviation theory applied to Equation B11*, which follows Equation B11.

Justification for averaging the second term in Equation B1

The assertion that the second term in Equation B1 should be replaced by its average while the first should not seems, at first glance, oddly asymmetric: after all, both terms are random variables, so why should we replace one by its average and not the other? The intuitive answer, as stated above, is that only one codeword ($w^\star$) is sent at a time, but, for each one we send, there are a large number of possible incorrect ones (all the rest of the w). The goal of this section is to make that intuitive answer mathematically precise.

Our starting point is Equation 8 for the probability, $P_{e,m}^{(n)}$, of making at least one error on m codewords. That equation made the implicit assumption that $P_e^{(n)}$ does not depend on which codeword, $w^\star$, was sent. To verify that this assumption is correct, we momentarily drop it and instead average over all possible codewords $w^\star$. When we do that, we will see that $P_e^{(n)}$ is effectively independent of $w^\star$.

From the point of view of decoding error, the only relevant aspect of $w^\star$ is its effect on the second term in Equation B1. We thus define

$$\mathcal{L} \equiv \frac{1}{n}\sum_{i=1}^{n} \log_2 q(\mathbf{r}_i|s_i(w^\star)). \quad \text{(B2)}$$

Because both s and r_i are discrete variables, it follows that $\mathcal{L}$ is also a discrete variable. Letting $P(\mathcal{L})$ denote its probability distribution, the expression for the probability of making at least one error is

$$P_{e,m}^{(n)} = \sum_{\mathcal{L}} P(\mathcal{L})\{1 - (1 - 2^{-nJ^\star(\mathcal{L})})^m\}, \quad \text{(B3)}$$

where $J^\star$ is defined via the relationship

$$2^{-nJ^\star(\mathcal{L})} \equiv \Pr\left(\frac{1}{n}\sum_{i=1}^{n} \log_2 q(\mathbf{r}_i|s_i(w)) > \mathcal{L}\right). \quad \text{(B4)}$$

Note that $2^{-nJ^\star(\mathcal{L})}$ is just the probability of an error, $P_e^{(n)}$, for a particular codeword, $w^\star$. The particular codeword in this case is the one that corresponds to $\mathcal{L}$ via Equation B2.

Unfortunately, we cannot compute analytically the sum in Equation B3. What we can do, however, is find the maximum number of codewords, m, for which $P_{e,m}^{(n)}$ vanishes exponentially fast with n. The information, $I^\star$, is then equal to the log of this number divided by n (see Eq. 10).

We begin by letting $m = 2^{n(I^\star - \epsilon)}$, as in Equation 10. Then, applying the formula $(1 - x)^y = \exp[y \ln(1 - x)]$, $P_{e,m}^{(n)}$ may be written

$$P_{e,m}^{(n)} = \sum_{\mathcal{L}} P(\mathcal{L})\{1 - \exp[2^{n[I^\star - \epsilon]}\ln(1 - 2^{-nJ^\star(\mathcal{L})})]\}. \quad \text{(B5)}$$

Our next step is to make the ansatz

$$I^\star = J^\star(\overline{\mathcal{L}}), \quad \text{(B6)}$$

where $\overline{\mathcal{L}}$ is the mean value of $\mathcal{L}$. What we do now is show that, with this ansatz, $P_{e,m}^{(n)}$ is exponentially small in n so long as $\epsilon > 0$. Inserting Equation B6 into B5, we have

$$P_{e,m}^{(n)} = \sum_{\mathcal{L}} P(\mathcal{L})\{1 - \exp[2^{n[J^\star(\overline{\mathcal{L}}) - \epsilon]}\ln(1 - 2^{-nJ^\star(\mathcal{L})})]\}. \quad \text{(B7)}$$

Because the expression in curly brackets is, for large n, essentially a step function in $J^\star$, we need to treat the sum over $\mathcal{L}$ differently on either side of the step. We thus divide it into two terms: one

with $\mathcal{L} < \mathcal{L}_0$ and one with $\mathcal{L} \geq \mathcal{L}_0$, where $\mathcal{L}_0$ is defined via the relationship

$$J^*(\mathcal{L}_0) = J^*(\overline{\mathcal{I}}) - \epsilon/2. \tag{B8}$$

Equation B2 then becomes

$$P_{e,m}^{(n)} = \sum_{\mathcal{L} \geq \mathcal{L}_0} P(\mathcal{L})\{1 - \exp[2^{n[J^*(\overline{\mathcal{I}})-\epsilon]}\ln(1 - 2^{-nJ^*(\mathcal{L})})]\} + \sum_{\mathcal{L} < \mathcal{L}_0} P(\mathcal{L})\{1 - \exp[2^{n[J^*(\overline{\mathcal{I}})-\epsilon]}\ln(1 - 2^{-nJ^*(\mathcal{L})})]\}.$$

The reason for this particular division into two terms is that we will be able to find exponentially small upper bounds for both. To find the bound for the first term, we use the fact that the expression in curly brackets is a decreasing function of $J^*(\mathcal{L})$. Then, because $J^*(\mathcal{L})$ is an increasing function of $\mathcal{L}$ (see Eq. B4), we can upper bound this term by replacing $J^*(\mathcal{L})$ by $J^*(\mathcal{L}_0)$ and then summing over all $\mathcal{L}$. For the second term, an upper bound is easily provided by simply dropping the exponential piece, which can only make this term smaller. Performing these manipulations and using Equation B8 for $J^*(\mathcal{L}_0)$, we find that

$$P_{e,m}^{(n)} \leq 1 - \exp[-2^{-n\epsilon/2}G(2^{-n[J^*(\overline{\mathcal{I}})-\epsilon/2]})] + \sum_{\mathcal{L} < \mathcal{L}_0} P(\mathcal{L}), \tag{B9}$$

where $G(x) \equiv \ln(1 - x)/(-x)$. Note that $\lim_{x \to 0} G(x) = 1$.

To find the large n behavior of the second term in Equation B9, we use the fact that $\mathcal{L}_0 < \overline{\mathcal{I}}$, which follows from the definition of $\mathcal{L}_0$ (Eq. B8) and the fact that $J^*(\mathcal{L})$ is an increasing function of $\mathcal{L}$. Moreover, for small ϵ, $\overline{\mathcal{I}} - \mathcal{L}_0 \propto \epsilon$, independent of n. Thus, Hoeffding's theorem (Hoeffding, 1963) tells us that the second term in Equation B9 is bounded by $2^{-n\kappa\epsilon^2}$, where κ is a positive, n-independent constant. We thus have

$$P_{e,m}^{(n)} \leq 1 - \exp[-2^{-n\epsilon/2}G(2^{-n|J^*(\overline{\mathcal{I}})-\epsilon/2|})] + 2^{-n\kappa\epsilon^2}.$$

Finally, because $G(x)$ approaches 1 in the small x limit, it follows that the first two terms in this expression reduce to $2^{-n\epsilon/2}$. Thus, in the limit of small ϵ (where $\kappa\epsilon^2 << \epsilon/2$), we have

$$\lim_{n \to \infty} \frac{\log_2 P_{e,m}^{(n)}}{n} \leq -\kappa\epsilon^2. \tag{B10}$$

Equation B10 is the main result of this section. It tells us that if $I^* = J^*(\overline{\mathcal{I}})$, then, for any fixed ϵ, the probability of the receiver making an error is exponentially small in n. Thus, $2^{nJ^*(\overline{\mathcal{I}})}$ is an upper bound on the number of codewords that can be sent with vanishingly small probability of error, and we identify $J^*(\overline{\mathcal{I}})$ as the information, I^*, associated with the wrong distribution. What this means is that, given the definition of $J^*(\mathcal{L})$ (Eq. B4), we can compute I^* by replacing the second term on the right-hand side of Equation B1 with $\overline{\mathcal{I}}$. As promised, then, $P_e^{(n)}$ is independent of w^*. Because $\overline{\mathcal{I}} = \langle \log_2 q(\mathbf{r}|s) \rangle_{p(\mathbf{r},s)}$, the equation for I^* is

$$2^{-nI^*} = \Pr\left(\frac{1}{n} \sum_{i=1}^{n} \log_2 q(\mathbf{r}_i|s_i(w)) > \langle \log_2 q(\mathbf{r}|s) \rangle_{p(\mathbf{r},s)} \right). \tag{B11}$$

Here and in what follows, the notation $\langle \ldots \rangle_p$ means average the terms inside the angle bracket with respect to the probability distribution p.

Large-deviation theory applied to Equation B11

Our task now is to compute the probability on the right-hand side of Equation B11. The key observation we need to do this is that, when $w \neq w^*$, $\mathbf{r}_i$ and $s_i(w)$ are independent, where independent in this context means $p(\mathbf{r}_i, s_i(w)) = p(\mathbf{r}_i)p(s_i(w))$. What we compute, then, is the probability that samples of $\mathbf{r}$ and s drawn from the distribution $p(\mathbf{r})p(s)$ will yield enough of an outlier to satisfy the inequality in Equation B11. This can be done using Sanov's theorum (Sanov, 1957; Cover and Thomas, 1991; Dembo and Zeitouni, 1993), which tells us that this probability, 2^{-nI^*}, is given by

$$2^{-nI^*} = 2^{-nD(p^*(\mathbf{r},s) \| p(\mathbf{r})p(s))}, \tag{B12}$$

where $D(\cdot\|\cdot)$ is the Kullback-Leibler divergence (in the above expression, it is equal to $\langle \log_2[p^*(\mathbf{r}, s)/p(\mathbf{r})p(s)] \rangle_{p^*(\mathbf{r},s)}$), and $p^*(\mathbf{r}, s)$ is chosen to minimize $D(p^*(\mathbf{r}, s)\|p(\mathbf{r})p(s))$ subject to the constraints

$$\langle \log_2 q(\mathbf{r}|s) \rangle_{p^*(\mathbf{r},s)} = \langle \log_2 q(\mathbf{r}|s) \rangle_{p(\mathbf{r},s)} \tag{B13a}$$

$$\sum_s p^*(\mathbf{r}, s) = p(\mathbf{r}). \tag{B13b}$$

The first constraint, Equation B13a, tells us that $p^*(\mathbf{r}, s)$ produces enough of an outlier to just barely satisfy the inequality in Equation B11. The second, Equation B13b, tells us that the responses are typical and is derived using the following reasoning. The probability of an error, 2^{-nI^*}, should be thought of as a probability over codewords; that is, 2^{-nI^*} is the probability that a randomly chosen codeword will satisfy the inequality in Equation B11. This probability depends, of course, on the $\mathbf{r}_i$. For large n, it is overwhelmingly likely that the $\mathbf{r}_i$ will be typical, meaning that the fraction of times a particular $\mathbf{r}$ appears is equal to its probability, $p(\mathbf{r})$ (Cover and Thomas, 1991). Moreover, we do not have to worry about outliers because, as mentioned above, we are interested in the probability of error per codeword sent. The same cannot be said about the stimuli: there are an exponentially large number of codewords in the codebook, and the ones most likely to produce an error are those that deviate from the distribution $p(s)$, which is why we do not have the constraint $\Sigma_{\mathbf{r}} p^*(\mathbf{r}, s) = p(s)$.

Equation B12 tells us that the mutual information associated with the wrong distribution, $q(\mathbf{r}|s)$, is given by

$$I^* = D(p^*(\mathbf{r}, s)\|p(\mathbf{r})p(s)). \tag{B14}$$

To compute I^*, we need to find $p^*(\mathbf{r}, s)$, the distribution that minimizes $D(p^*(\mathbf{r}, s)\|p(\mathbf{r})p(s))$ subject to the constraints given in Equation B13. This is a straightforward problem in constrained minimization: $p^*(\mathbf{r}, s)$ is found by solving the equation

$$\frac{d}{dp^*(\mathbf{r}, s)} \left[D(p^*(\mathbf{r}, s)\|p(\mathbf{r})p(s)) - \beta \langle \log_2 q(\mathbf{r}|s) \rangle_{p^*(\mathbf{r}, s)} - \sum_{\mathbf{r}} \lambda(\mathbf{r}) \sum_s p^*(\mathbf{r}, s) \right] = 0, \tag{B15}$$

and then choosing β and $\lambda(\mathbf{r})$, the Lagrange multipliers, to satisfy the two constraints in Equation B13.

Equations B13–B15 are, with minor changes in notation, the same as those found by Merhav et al. (1994).

Bounding $I^\star$

Although these equations cannot be solved analytically, they can be reduced to a form that allows easy derivation of a bound on $I^\star$. To do that, we proceed in steps: first we find the solution for arbitrary β and $\lambda(\mathbf{r})$, then we eliminate $\lambda(\mathbf{r})$ by enforcing the constraint in Equation B13b, and finally we cast the remaining equation for β in terms of a minimization problem.

Denoting the solution to Equation B15 $\tilde{p}(\mathbf{r}, s; \beta, \lambda)$, we have, after straightforward algebra,

$$\tilde{p}(\mathbf{r}, s; \beta, \lambda) \propto p(\mathbf{r})p(s)2^{\beta \log_2 q(\mathbf{r}|s) + \lambda(\mathbf{r})}.$$

We can solve for $\lambda(\mathbf{r})$, and thus eliminate it, by enforcing the constraint given in Equation B13b, and we arrive at

$$\tilde{p}(\mathbf{r}, s; \beta) = \frac{p(\mathbf{r})p(s)2^{\beta \log_2 q(\mathbf{r}|s)}}{Z(\mathbf{r}, \beta)},$$

where the normalization, $Z(\mathbf{r}, \beta)$, is given by

$$Z(\mathbf{r}, \beta) \equiv \sum_s p(s)2^{\beta \log_2 q(\mathbf{r}|s)}. \quad \text{(B16)}$$

It is easy to show that $\tilde{p}(\mathbf{r}, s; \beta)$ satisfies Equation B13b; that is, $\Sigma_s \tilde{p}(\mathbf{r}, s; \beta) = p(\mathbf{r})$.

Our next step is to find the value of β, denoted $\beta^\star$, that satisfies Equation B13a. Although we cannot do this analytically, we can show that $\beta^\star$ satisfies a convex optimization problem. This is clearly convenient for numerical work, and it is also convenient for deriving bounds on $I^\star$. We start by defining

$$\Delta\tilde{I}(\beta) \equiv D(p(\mathbf{r}, s) \| \tilde{p}(\mathbf{r}, s; \beta)). \quad \text{(B17)}$$

The quantity $\Delta\tilde{I}(\beta)$ is significant for two reasons, both of which we demonstrate below: it has a single minimum at $\beta = \beta^\star$, and its value at that minimum is $I - I^\star$, the information loss associated with ignoring correlations.

To show that $\Delta\tilde{I}(\beta^\star)$ is a minimum, we first differentiate both sides of Equation B17 with respect to β, which yields

$$\frac{\partial \Delta\tilde{I}(\beta)}{\partial \beta} = \langle \log_2 q(\mathbf{r}|s) \rangle_{\tilde{p}(\mathbf{r}, s; \beta)} - \langle \log_2 q(\mathbf{r}|s) \rangle_{p(\mathbf{r}, s)}. \quad \text{(B18)}$$

The right-hand side of Equation B18 vanishes when $\beta = \beta^\star$: by definition $\tilde{p}(\mathbf{r}, s; \beta^\star) = p^\star(\mathbf{r}, s)$, and Equation B13a tells us that $\langle \log_2 q(\mathbf{r}|s) \rangle_{p^\star(\mathbf{r}, s)} = \langle \log_2 q(\mathbf{r}|s) \rangle_{p(\mathbf{r}, s)}$. Thus, $\beta^\star$ is an extremum. To show that this extremum is a minimum, we compute the second derivative of $\Delta\tilde{I}(\beta)$. Straightforward algebra yields

$$\frac{\partial^2 \Delta\tilde{I}(\beta)}{\partial \beta^2} = \ln 2 \langle \mathrm{Var}[\log_2 q(\mathbf{r}|s)]_{\tilde{p}(s|\mathbf{r}; \beta)} \rangle_{p(\mathbf{r})}, \quad \text{(B19)}$$

where $\tilde{p}(s|\mathbf{r}; \beta) \equiv \tilde{p}(\mathbf{r}, s; \beta)/p(\mathbf{r})$. Excluding the trivial case of a deterministic mapping from stimulus to response, the variance on the right-hand side of Equation B19 is positive. Thus, $\Delta\tilde{I}(\beta)$ is convex and so has a single minimum at $\beta^\star$.

To show that $\Delta\tilde{I}(\beta^\star) = I - I^\star$, we use the definition of $\Delta\tilde{I}(\beta)$ and a small amount of algebra to derive the relationship

$$\Delta\tilde{I}(\beta^\star) = I - [\beta^\star \langle \log_2 q(\mathbf{r}|s) \rangle_{p(\mathbf{r}, s)} - \langle \log_2 Z(\mathbf{r}, \beta^\star) \rangle_{p(\mathbf{r}, s)}].$$

Then, using Equations B13a and B13b to replace the averages with respect to $p(\mathbf{r}, s)$ by averages with respect to $p^\star(\mathbf{r}, s)$ and comparing the resulting expression with Equation B14, it is easy to see that the second term in brackets is equal to $I^\star$.

This analysis tells us that the information loss associated with the wrong distribution is the minimum value of $\Delta\tilde{I}(\beta)$. That allows us to perform three quick sanity checks. First, because $\Delta\tilde{I}$ is non-negative (it is a Kullback-Leibler divergence; see Eq. B17), the information loss can never be less than zero. Second, when $q(\mathbf{r}|s) = p(\mathbf{r}|s)$, $\Delta\tilde{I}(1) = 0$, indicating that there is no information loss when we use the true distribution. And third, $\tilde{p}(\mathbf{r}, s; 0) = p(\mathbf{r})p(s)$, which means that $\Delta\tilde{I}(0) = I$; this indicates that the information loss can never exceed the actual information, I.

Because there is no closed-form expression for $\Delta\tilde{I}$, it is useful to find an upper bound on it. We can do this by evaluating $\Delta\tilde{I}(\beta)$ at any β. A convenient choice is $\beta = 1$, at which point we have, using Equation B17 for $\Delta\tilde{I}(\beta)$, Equation B16 for $Z(\mathbf{r}, \beta)$, and a small amount of algebra,

$$\Delta\tilde{I}(\beta^\star) \le \Delta\tilde{I}(1) = \left\langle \log_2 \left[\frac{p(s|\mathbf{r}) \sum_{s'} q(\mathbf{r}|s')p(s')}{q(\mathbf{r}|s)p(s)} \right] \right\rangle_{p(\mathbf{r}, s)}. \quad \text{(B20)}$$

Finally, to simplify the right-hand side of Equation B20, we define $q(s|\mathbf{r})$ via

$$q(s|\mathbf{r}) = \frac{q(\mathbf{r}|s)p(s)}{\sum_{s'} q(\mathbf{r}|s')p(s')},$$

and we find that

$$I - I^\star = \Delta\tilde{I}(\beta^\star) \le \Delta\tilde{I}(1) = \langle D(p(s|\mathbf{r}) \| q(s|\mathbf{r})) \rangle_{p(\mathbf{r})}. \quad \text{(B21)}$$

When $q(s|\mathbf{r}) = p_{\mathrm{ind}}(s|\mathbf{r})$, the right-hand side of Equation B21 is ΔI. Thus, Equation B21 tells us that ΔI is an upper bound on the information loss.

References

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

Atick JJ (1992) Could information theory provide an ecological theory of sensory processing? Network 3:213–251.

Atick JJ, Redlich AN (1990) Towards a theory of early visual processing. Neural Comput 2:308–320.

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

Averbeck BB, Lee D (2003) Neural noise and movement-related codes in the macaque supplementary motor area. J Neurosci 23:7630–7641.

Averbeck BB, Lee D (2004) Coding and transmission of information by neural ensembles. Trends Neurosci 27:225–230.

Averbeck BB, Crowe DA, Chafee MV, Georgopoulos AP (2003) Neural activity in prefrontal cortex during copying geometrical shapes. II. Decoding shape segments from neural ensembles. Exp Brain Res 150:142–153.

Barlow H (1961) The coding of sensory messages. In: Current problems in animal behavior (Thorpe WH, Zangwill OL, eds), pp 331–361. Cambridge, MA: Cambridge UP.

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

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

Cover TM, Thomas JA (1991) Elements of information theory. New York: Wiley.

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

deCharms RC, Merzenich MM (1996) Primary cortical representation of sounds by the coordination of action-potential timing. Nature 381:610–613.

Dembo A, Zeitouni O (1993) Large deviation techniques and applications. New York: Springer.

Eckhorn R, Bauer R, Jordan W, Brosch M, Kruse W, Munk M, Reitboeck HJ (1988) Coherent oscillations: a mechanism of feature linking in the visual cortex? Multiple electrode and correlation analyses in the cat. Biol Cybern 60:121–130.

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

Golledge HD, Panzeri S, Zheng F, Pola G, Scannell J, Giannikopoulos DV, Mason RJ, Tovee MJ, Young MP (2003) Correlations, feature-binding and population coding in primary visual cortex. NeuroReport 14:1045–1050.

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

Gray CM, Konig P, Engel AK, Singer W (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature 338:334–337.

Hoeffding W (1963) Probability inequalities for sums of bounded random variables. J Am Stat Assoc 58:13–30.

Levine MW, Castaldo K, Kasapoglu MB (2002) Firing coincidences between neighboring retinal ganglion cells: inside information or epiphenomenon? Biosystems 67:139–146.

Liu RC, Tzonev S, Rebrik S, Miller KD (2001) Variability and information in a neural code of the cat lateral geniculate nucleus. J Neurophysiol 86:2789–2806.

Machens CK, Stemmler MB, Prinz P, Krahe R, Ronacher B, Herz AV (2001) Representation of acoustic communication signals by insect auditory receptor neurons. J Neurosci 21:3215–3227.

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

Merhav N, Kaplan G, Lapidoth A, Shamai Shitz S (1994) On information rates for mismatched decoders. IEEE Trans Inform Theory 40:1953–1967.

Nemenman I, Bialek W, de Ruyter van Steveninck R (2004) Entropy and information in neural spike trains: progress on the sampling problem. Phys Rev E Stat Nonlin Soft Matter Phys 69:056111.

Nirenberg S, Latham PE (1998) Population coding in the retina. Curr Opin Neurobiol 8:488–493.

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

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

Oram MW, Hatsopoulos NG, Richmond BJ, Donoghue JP (2001) Excess synchrony in motor cortical neurons provides redundant direction information with that from coarse temporal measures. J Neurophysiol 86:1700–1716.

Osborne LC, Bialek W, Lisberger SG (2004) Time course of information about motion direction in visual area MT of macaque monkeys. J Neurosci 24:3210–3222.

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

Paninski L (2004) Estimating entropy on *m* bins given fewer than *m* samples. IEEE Trans Inform Theory 50:2200–2203.

Panzeri S, Treves A, Schultz S, Rolls ET (1999) On decoding the responses of a population of neurons from short time windows. Neural Comput 11:1553–1577.

Panzeri S, Golledge HDR, Zheng F, Tovée MJ, Young MP (2001) Objective assessment of the functional role of spike train correlations using information measures. Vis Cogn 8:531–547.

Panzeri S, Golledge HDR, Zheng F, Pola G, Blanche TJ, Tovée MJ, Young MP (2002a) The role of correlated firing and synchrony in coding information about single and separate objects in cat V1. Neurocomputing 44–46:579–584.

Panzeri S, Pola G, Petroni F, Young MP, Petersen RS (2002b) A critical assessment of different measures of the information carried by correlated neuronal firing. Biosystems 67:177–185.

Petersen RS, Panzeri S, Diamond ME (2001) Population coding of stimulus location in rat somatosensory cortex. Neuron 32:503–514.

Petersen RS, Panzeri S, Diamond ME (2002) Population coding in somatosensory cortex. Curr Opin Neurobiol 12:441–447.

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

Sanov IN (1957) On the probability of large deviations of random variables. Mat Sbornik 42:11–44.

Schneidman E, Bialek W, Berry MJ (2003) Synergy, redundancy, and independence in population codes. J Neurosci 23:11539–11553.

Shannon CE, 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 Stat Nonlin Soft Matter Phys [Erratum 2002 65(4 Pt 2B)] 64:051904.

Srinivasan MV, Laughlin SB, Dubs A (1982) Predictive coding: a fresh view of inhibition in the retina. Proc R Soc Lond B Biol Sci 216:427–459.

Steinmetz PN, Roy A, Fitzgerald PJ, Hsiao SS, Johnson KO, Niebur E (2000) Attention modulates synchronized neuronal firing in primate somatosensory cortex. Nature 404:187–190.

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.

Wu S, Nakahara H, Murata N, Amari S (2000) Population decoding based on an unfaithful model. Advances in neural information processing systems, pp 167–173. Cambridge, MA: MIT.

Wu S, Nakahara H, Amari S (2001) Population coding with correlation and an unfaithful model. Neural Comput 13:775–797.

Wu S, Amari S, Nakahara H (2002) Population coding and decoding in a neural field: a computational study. Neural Comput 14:999–1026.