Network: Comput. Neural Syst. **10** (1999) 325–340. Printed in the UK PII: S0954-898X(99)08490-0

How to measure the information gained from one symbol

Michael R DeWeese† and Markus Meister‡

† The Salk Institute, Sloan Center, La Jolla, CA 92037, USA
‡ Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA

E-mail: deweese@salk.edu

Received 31 March 1999, in final form 3 September 1999

Abstract. Information theory provides a powerful framework to analyse how neurons represent sensory stimuli or other behavioural variables. A recurring question regards the amount of information conveyed by a specific neuronal response. Here we show that the commonly used definition for this quantity has a serious flaw: the information accumulated during subsequent observations of neural activity fails to combine additively. Additivity is a highly desirable property, both on theoretical grounds and for the practical purpose of analysing population codes. We propose an alternative measure for the information per observation and prove that this is the only definition that satisfies additivity. The old and the new definitions measure very different aspects of the neural code, which is illustrated with visual responses from a motion-sensitive neuron in the primate cortex. Our analysis allows additional interpretation of several published results, which suggests that the neurons studied are operating far from their information capacity.

1. Introduction

Signals in the nervous system take many different physical forms, including membrane potentials, ionic currents, neurotransmitter concentrations and enzymatic activities. The essential substance that is transported through neural pathways by these ever-changing symbols is information. Thus, a student of the nervous system stands to benefit from understanding the characteristic properties of this substance called information. By analogy, a vascular physiologist would study fluid mechanics, because the physical properties of fluids place important bounds on how rapidly blood can be transported from one place to another through a given vessel. Similarly, there exists a theory of information that prescribes how to measure this quantity, and spells out important constraints on transmitting information through a communication channel such as a neural pathway. Ever since the formulation of this framework by Shannon (1948a, b), there has been considerable interest in how neural systems deal with these constraints, and how close they come to the performance limits specified by the theory (Rieke *et al* 1997).

In many studies of neural function we stimulate one end of the nervous system—for example, by a sensory input or by driving a specific neuron—and observe the response at another end—for example, the firing of a population of neurons, or motor output. Generally, different values of the input will lead to different values of the output, but clearly there are limits to this, and those limits define the operating range of the neural system. For example, as a result of noise or uncontrolled variables in the pathway, a given input value may lead to several different output values. Similarly, a given output may have been caused by different

0954-898X/99/040325+16$30.00

inputs. Studies of neural coding are essentially aimed at understanding these probabilistic relationships (Rieke *et al* 1997).

The fidelity of transmission between inputs and outputs can be assessed by Shannon's mutual information. This quantity measures the average information one obtains about the input value from observing an output value, or vice versa. However, in many cases it would be interesting to know the information gained from a specific observation, rather than the average. For example, some neural firing patterns may be more informative about the stimulus than others, and such a comparison could show which aspects of firing are reliable, and which others are affected by noise. Similarly, some stimuli may be more informative about the neural response than others, and such a comparison may reveal what the given neural circuit can sense. However, Shannon's formulation of the theory provides no prescription for measuring this information specific to a particular symbol. Subsequent theoretical work attempts to capture this specific information by two rather different definitions—which we will denote I_1 and I_2—and there exists an infinity of equally plausible alternatives.

Any acceptable measure of this specific information must have one important property: it should accumulate additively in the course of subsequent observations. Imagine making two observations that provide information about some variable of interest. Compute the specific information gained from the first, then the specific information obtained from the second; of course, in this latter step you must consider what knowledge you already gained from the first observation. It is then natural to expect that the specific information gained in each of the two steps should sum to the information gained if you simply made both observations at the same time. For example, the information obtained about a sensory stimulus from observing two neurons in a population should equal the information from the first neuron plus the information gained from the second neuron after one had already observed the first. Here we show that of the two definitions advanced so far, only I_2 has this property of additivity. In fact, we prove that it is the only possible definition that satisfies additivity, which identifies it uniquely as a proper measure of information. Unfortunately, this is not the expression used by neuroscientists: published studies of neural coding all calculate I_1. We show that I_1 is also unique, in that it is the only definition that never produces negative numbers, and argue that it has the properties of 'surprise' rather than 'information'. This paper begins with a brief introduction to the formalism of information theory. We then prove the unique properties of I_1 and I_2 and illustrate the difference between the two measures with a toy example from medical diagnostics. Both analyses are then applied to recordings of a motion-sensitive neuron in primate visual cortex. Finally, we make use of a classic theorem to reinterpret the measurements of the 'surprise' I_1 published in several previous studies.

1.1. Entropy

One can consider neural signalling as a communication process between a set of input symbols $X = \{x_i\}$—for example, different sensory stimuli—and a set of output symbols $Y = \{y_j\}$—for example, different neural firing rates. In the course of communication, different inputs occur with probability $p(x_i)$ and the outputs with probability $p(y_j)$. A given input x_i may lead to several different outputs y_j, with the conditional probability $p(y_j|x_i)$. Similarly, a given output y_j may have been caused by different inputs x_i with conditional probability $p(x_i|y_j) = p(y_j|x_i)p(x_i)/p(y_j)$. Thus, observation of the output generally leaves some uncertainty about the input. Shannon's information theory (Shannon and Weaver 1963, Cover and Thomas 1991) provides a way to measure this uncertainty: if events x_i in the ensemble X can happen with probability $p(x_i)$, then the uncertainty about X is defined as the entropy of

the ensemble

$$H(X) = -\sum_i p(x_i) \log p(x_i) \tag{1}$$

where 'log' denotes the logarithm to base 2. For example, the entropy of an event with two equally likely outcomes is $\log(2) = 1$, often expressed as '1 bit'.

This definition satisfies an important property that we associate intuitively with uncertainty: it is additive. If two different events X and Y happen independently of each other, then our uncertainty about both of their outcomes should equal the sum of the uncertainties about each event. More generally, if the two events are not statistically independent, so that observation of the first conveys some knowledge of the second, we expect that the uncertainty about the joint event should equal the uncertainty about the first event plus the uncertainty about the second given knowledge of the first. In fact, Shannon's entropy has this property:

$$\begin{aligned} H(X, Y) &= -\sum_{i,j} p(x_i, y_j) \log p(x_i, y_j) \\ &= -\sum_{i,j} p(y_j) p(x_i|y_j) \log[p(y_j) p(x_i|y_j)] \\ &= -\sum_j p(y_j) \log p(y_j) \sum_i p(x_i|y_j) - \sum_j p(y_j) \sum_i p(x_i|y_j) \log p(x_i|y_j) \\ &= H(Y) + H(X|Y) \end{aligned}$$

where $H(X|Y)$ is the average uncertainty remaining about X after one has observed Y, often called the 'equivocation' or 'conditional entropy',

$$\begin{aligned} H(X|Y) &= -\sum_j p(y_j) \sum_i p(x_i|y_j) \log p(x_i|y_j) \\ &= \sum_j p(y_j) H(X|y_j), \end{aligned}$$

and

$$H(X|y_j) = -\sum_i p(x_i|y_j) \log p(x_i|y_j)$$

is the entropy of X conditional on observation of a specific event y_j (Fano 1961, Hamming 1986, Cover and Thomas 1991).

This additivity with respect to compound events is a natural requirement we should impose on any quantitative measure, if we want it to meet our intuitive notions of uncertainty. It has been shown (Shannon 1948a, Khinchin 1957, Aczel and Daroczy 1975, Guiasu 1977) that this is the key requirement leading to Shannon's unique definition of entropy (1), and specifically its logarithmic form.

1.2. Mutual information

Having developed a measure of uncertainty, one can define the transmitted information: observing the output Y of the communication system reduces the uncertainty about the input X from $H(X)$ to $H(X|Y)$. Thus Shannon identified the information conveyed by Y about X as the average change in uncertainty from observing Y:

$$I(X; Y) = H(X) - H(X|Y) = \sum_{i,j} p(x_i, y_j) \log \left[\frac{p(x_i, y_j)}{p(x_i) p(y_j)} \right]. \tag{2}$$

Note that this measure is symmetric with respect to inputs and outputs,

$$I(X; Y) = I(Y; X).$$

So knowledge of the input conveys as much information about the output as observation of the output conveys about the input; for this reason $I(X; Y)$ is called the 'mutual information'.

Again, the mutual information satisfies an important intuitive additivity property. If we observe two events taken from the ensembles Y and Z, then the information $I(X; \{Y, Z\})$ that both events convey about X should be equal to the information gained from Y plus the information gained from Z given what we already knew from Y. This is easily verified (Fano 1961, Cover and Thomas 1991):

$$\begin{aligned} I(X; \{Y, Z\}) &= \sum_{ijk} p(x_i, y_j, z_k) \log \left[\frac{p(x_i, y_j, z_k)}{p(x_i) p(y_j, z_k)} \right] \\ &= I(X; Y) + \sum_j p(y_j) I(X; Z|y_j) \\ &= I(X; Y) + I(X; Z|Y) \end{aligned}$$

where

$$I(X; Z|Y) = \sum_j p(y_j) I(X; Z|y_j)$$

is the average information about X obtained from observing Z, given that one already has knowledge of some observation in Y, and

$$I(X; Z|y_j) = \sum_{ik} p(x_i, z_k|y_j) \log \left[\frac{p(x_i, z_k|y_j)}{p(x_i|y_j) p(z_k|y_j)} \right]$$

is the information about X from observing Z, given prior observation of the specific event y_j.

1.3. Specific information

The mutual information (2) specifies how much information is conveyed *on average* over all symbols. As discussed above, one is tempted to ask whether some symbols are more informative than others. So we seek a definition for the information $I(X; y_j)$ gained from observation of *a specific* output symbol y_j about the range of possible input symbols X; we will call this the 'specific information' obtained from y_j. Observation of y_j changes the probability distribution of the input X from $p(x)$ to $p(x|y_j)$. These two distributions completely define the knowledge we have about X before and after the observation. Thus the desired quantity $I(X; y_j)$ must be a functional of these two probability distributions:

$$I(X; y_j) = F[p(x), p(x|y_j)]. \tag{3}$$

Furthermore, we require that the average information gained over all possible observations y_j should equal the mutual information $I(X; Y)$:

$$\sum_j p(y_j) I(X; y_j) = I(X; Y). \tag{4}$$

An expression for $I(X; y_j)$ that satisfies these requirements is

$$I_1(X; y_j) = \sum_i p(x_i|y_j) \log \left[\frac{p(x_i|y_j)}{p(x_i)} \right]. \tag{5}$$

However, it is not unique in this regard. An alternative is

$$\begin{aligned} I_2(X; y_j) &= H(X) - H(X|y_j) \\ &= -\sum_i p(x_i) \log p(x_i) + \sum_i p(x_i|y_j) \log p(x_i|y_j). \end{aligned} \tag{6}$$

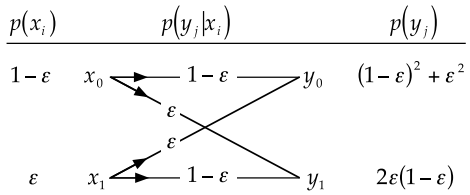

Figure 1. Schematic diagram for communication by a diagnostic test that produces false positives and false negatives, each with frequency ε. This channel is known as the 'binary symmetric channel with noise'. x_0: subject healthy; x_1: subject sick; y_0: test negative; y_1: test positive.

This second definition has a very simple interpretation: it amounts to the change in uncertainty about X that occurs when one observes y_j.

Among treatments of neural coding, virtually all of the literature uses definition $I_1(X; y_j)$ for the information conveyed by a specific y_j (Eckhorn and Pöpel 1974, 1975, Fuller and Looft 1984, Optican and Richmond 1987, de Ruyter van Steveninck and Bialek 1988, Bialek and Zee 1990, Richmond and Optican 1990, Theunissen and Miller 1991, Panzeri and Treves 1996, Rolls *et al* 1996, 1998, Buracas *et al* 1998). No compelling reason for this choice is provided, and we have found only one source (Rieke *et al* 1997, p 122) that mentions the existence of an alternative definition $I_2(X; y_j)$. In fact, there is an infinity of alternatives: for example, any weighted average of I_1 and I_2 will also satisfy the requirement (4). Thus, one needs to introduce an additional criterion in order to choose an appropriate definition for specific information. We will arrive at this criterion by inspecting how I_1 and I_2 behave in a simple example.

2. An example

Suppose a rare disease afflicts individuals in a population with relative frequency ε. You have available a diagnostic test for this disease. Unfortunately, the test is not perfect, but occasionally produces false positive results or false negative results, and both happen to occur with a probability ε (figure 1). You perform this test on a subject and obtain a positive result, y_1. This could have arisen from a correct test on a sick person (x_1) or from a false positive on a healthy person (x_0); in fact the two alternatives are equally likely. Thus, after observing the test result, it is now equally probable that the subject is healthy as that he is sick:

$$p(x_0|y_1) = p(x_1|y_1) = \tfrac{1}{2}.$$

What would the two measures I_1 and I_2 above tell us about the information gained from this specific test result?

$$\begin{aligned} I_1(X; y_1) &= p(x_0|y_1)\log\left[\frac{p(x_0|y_1)}{p(x_0)}\right] + p(x_1|y_1)\log\left[\frac{p(x_1|y_1)}{p(x_1)}\right] \\ &= \frac{1}{2}\log\frac{1}{2(1-\varepsilon)} + \frac{1}{2}\log\frac{1}{2\varepsilon} \\ &\approx \frac{1}{2}\log\frac{1}{4\varepsilon} \gg 1, \qquad \text{for} \quad \varepsilon \ll 1. \end{aligned} \tag{7}$$

So if we measure the specific information by I_1, we conclude that the test yielded a large amount of information, particularly if the disease is very rare. On the other hand,

$$\begin{aligned} I_2(X; y_1) &= p(x_0|y_1)\log p(x_0|y_1) + p(x_1|y_1)\log p(x_1|y_1) \\ &\quad - p(x_0)\log p(x_0) - p(x_1)\log p(x_1) \end{aligned}$$

$$\begin{aligned} &= \tfrac{1}{2}\log\tfrac{1}{2} + \tfrac{1}{2}\log\tfrac{1}{2} - \varepsilon\log\varepsilon - (1-\varepsilon)\log(1-\varepsilon) \\ &\approx -1, \qquad \text{for} \quad \varepsilon \ll 1. \end{aligned}$$

If we measure specific information by I_2, then we have lost information by this test result, approximately 1 bit. This is because I_2 simply measures how much our uncertainty about X has changed. Prior to the test, we were almost certain that the subject is healthy (assuming $\varepsilon \ll 1$) and thus the entropy $H(X)$ was almost zero. After the test, we are perfectly uncertain bout the subject's health, and thus the entropy $H(X|y_1)$ is 1 bit. Thus the test has increased our uncertainty by approximately 1 bit. Clearly, the two information measures I_1 and I_2 differ a great deal in their interpretation of the test results.

A responsible clinician would want to repeat the test. Suppose the second test on the same subject comes out negative, $z = z_0$. Then we know that either the first test gave a false positive or the second test gave a false negative result. Both possibilities are equally likely, independently of whether the patient is truly sick. Thus, the probability $p(x_1|y_1, z_0)$ of the subject being sick is exactly what it was before we performed any tests, namely ε:

$$p(x_1|y_1, z_0) = 1 - p(x_0|y_1, z_0) = \varepsilon.$$

What do our specific information measures say about the information obtained from this second test?

$$\begin{aligned} I_1(X; z_0|y_1) &= p(x_0|y_1, z_0)\log\left[\frac{p(x_0|y_1, z_0)}{p(x_0|y_1)}\right] + p(x_1|y_1, z_0)\log\left[\frac{p(x_1|y_1, z_0)}{p(x_1|y_1)}\right] \\ &= (1-\varepsilon)\log[2(1-\varepsilon)] + \varepsilon\log[2\varepsilon] \\ &\approx 1, \qquad \text{for} \quad \varepsilon \ll 1, \end{aligned}$$

whereas

$$\begin{aligned} I_2(X; z_0|y_1) &= -p(x_0|y_1)\log p(x_0|y_1) - p(x_1|y_1)\log p(x_1|y_1) \\ &\quad + p(x_0|y_1, z_0)\log p(x_0|y_1, z_0) + p(x_1|y_1, z_0)\log p(x_1|y_1, z_0) \\ &= -\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{2}\log\tfrac{1}{2} + \varepsilon\log\varepsilon + (1-\varepsilon)\log(1-\varepsilon) \\ &\approx 1, \qquad \text{for} \quad \varepsilon \ll 1. \end{aligned}$$

Both I_1 and I_2 suggest that information was gained from this second test. As measured by I_2, the information gained in the second test is exactly equal and opposite to the information lost during the first test. Thus the total information gained from both tests is zero. This agrees with the fact that we know precisely as much about the patient's health after these two tests as we did before. As measured by I_1, on the other hand, the accumulated information gained from these two tests is large, particularly when $\varepsilon \ll 1$, which blatantly contradicts the facts.

This example illustrates an important difference between I_1 and I_2: I_2 is additive whereas I_1 is not. If we use measure I_2, then the information obtained from two observations, y_j and z_k, is equal to that obtained from y_j plus that obtained from z_k given what was already known from y_j:

$$\begin{aligned} I_2(X; \{y_j, z_k\}) &= H(X) - H(X|y_j, z_k) \\ &= H(X) - H(X|y_j) + H(X|y_j) - H(X|y_j, z_k) \\ &= I_2(X; y_j) + I_2(X; z_k|y_j). \end{aligned}$$

On the other hand, this is not the case for I_1 (Watanabe 1969, p 533):

$$\begin{aligned} I_1(X;\{y_j,z_k\}) &= \sum_i p(x_i|y_j,z_k)\log\left[\frac{p(x_i|y_j,z_k)}{p(x_i)}\right] \\ I_1(X;y_j)+I_1(X;z_k|y_j) &= \sum_i p(x_i|y_j)\log\left[\frac{p(x_i|y_j)}{p(x_i)}\right] \\ &\quad +\sum_i p(x_i|y_j,z_k)\log\left[\frac{p(x_i|y_j,z_k)}{p(x_i|y_j)}\right] \\ &\neq I_1(X;\{y_j,z_k\}). \end{aligned} \tag{8}$$

3. A unique measure of 'specific information'

The lack of additivity presents a serious problem for the candidacy of I_1 as a measure of specific information. First of all, it conflicts with our intuitive notion that information accumulates additively over a sequence of observations, such that the total obtained in all steps is equal to what we would calculate if we considered all the events as a single observation. Secondly, the additivity for entropy and information measures is at the heart of Shannon's information theory. In fact, as reviewed above, additivity is the defining criterion that leads to the logarithmic form of Shannon entropy, on which the entire formalism is based (Shannon 1948a, Khinchin 1957, Aczel and Daroczy 1975, Guiasu 1977). Finally, a measure of specific information $I(X;y_j)$ is useful only because it allows us to make comparisons across symbols y_j. For example, analyses of neural coding often compare the information obtained from different response patterns (Eckhorn and Pöpel 1974, 1975, de Ruyter van Steveninck and Bialek 1988, Bialek and Zee 1990) or the information transmitted by different stimuli (Fuller and Looft 1984, Optican and Richmond 1987, Richmond and Optican 1990, Theunissen and Miller 1991, Panzeri and Treves 1996, Rolls *et al* 1996, 1998, Buracas *et al* 1998). With increasing availability of simultaneous recordings from an entire neuronal population (Krüger and Aiple 1988, Wilson and McNaughton 1993, Warland *et al* 1997), it is possible to compare the specific information each neuron provides with the population code. In all these cases one presumes that the various specific informations represent the contribution of each observed symbol to the communication process, and that these contributions sum to the overall information flow. If, instead, the components do not sum to the whole, then a comparison between the components is pointless. Thus we are led to reject I_1 as a measure of specific information.

I_2 appears as a suitable alternative that satisfies additivity. An intriguing feature of $I_2(X;y_j)$ is that it can be negative, as for the first diagnostic test in the above example. This is because certain observations y_j do, in fact, increase our uncertainty about the state of the variable X. However, the average of $I_2(X;y_j)$ over all possible y_j is equal to the mutual information (4), which is never negative (Shannon and Weaver 1963, Cover and Thomas 1991). Furthermore, if one insists on additivity, then the specific information—however it is defined—must on occasion take on negative values: as illustrated in the above example, two subsequent observations may combine to produce zero information, and thus they cannot both make positive contributions. One can go further and prove that $I_2(X;y_j)$ is, in fact, a unique definition of specific information: it is the only expression that satisfies additivity

$$I(X;\{y_j,z_k\}) = I(X;y_j)+I(X;z_k|y_j) \tag{9}$$

and averages to the mutual information

$$\sum_j p(y_j)I(X;y_j) = I(X,Y). \tag{10}$$

A proof is given in the appendix.

Because additivity is central to the practical utility of an information measure, as discussed above, we conclude that $I_2(X; y_j)$ is the preferred definition of specific information, and will, from now on, denote it simply by $I(X; y_j)$:

$$I(X; y_j) = \text{information about } X \text{ from observing } y_j$$
$$= I_2(X; y_j) = -\sum_i p(x_i) \log p(x_i) + \sum_i p(x_i|y_j) \log p(x_i|y_j). \tag{11}$$

4. A unique positive measure: the 'specific surprise'

Nevertheless, $I_1(X; y_j)$ is a potentially useful measure related to the effects of observing y_j. We suggest that this measure be termed the 'surprise' about X from observation of a specific y_j, and will, from now on, denote it by $S(X; y_j)$:

$$S(X; y_j) = \text{surprise about } X \text{ from observing } y_j$$
$$= I_1(X; y_j) = \sum_i p(x_i|y_j) \log \left[\frac{p(x_i|y_j)}{p(x_i)} \right].$$

Note that $S(X; y_j)$ is particularly large when $p(x|y_j)$ dominates in regions of X where $p(x)$ is small. In that case the observation has moved our estimate of x towards values that seemed very unlikely prior to the observation: a 'surprising' result. For example, this applies to the first diagnostic test in the above scenario, whose positive outcome suddenly makes it much more likely that the subject is sick. The comparison of S and I in that example confirms that one can experience plenty of surprise without gaining any information. More generally, we would not expect the surprise to be additive. For example, if we were given the outcome of the two diagnostic tests as a combined observation, we could immediately conclude that the two tests are inconsistent; while this itself may be unusual, the combined tests say nothing about the health of the subject, and consequently there is no surprise about that input variable. Thus it is clear that the surprise $S(X; y_j)$ and the specific information $I(X; y_j)$ measure two very different aspects of the communication process.

Another natural expectation is that 'surprise' should be a positive number. At worst, it might be zero, namely when the observation of y changes nothing about the distribution of x. It is difficult to associate an intuitive meaning with negative surprise. As a matter of fact, it is well known that the quantity $S(X; y_j)$ is never negative (Fano 1961, Cover and Thomas 1991). Furthermore, one can show that $S(X; y_j)$ is unique in this regard: it is the only quantity specific to observation of y_j that is strictly non-negative

$$S(X; y_j) \geqslant 0 \qquad \text{for all} \quad y_j \tag{12}$$

and whose average amounts to the mutual information

$$\sum_j p(y_j) S(X; y_j) = I(X; Y). \tag{13}$$

A proof is given in the appendix.

5. Information and surprise in a neural code for visual motion

To illustrate the developments of the preceding sections we now analyse the encoding of visual motion information by neurons in the middle temporal (MT) cortical area of an alert macaque monkey (for experimental details see Buracas *et al* (1998)). In one experiment, eight different visual stimuli x_i were presented within the neuron's receptive field, consisting of a windowed

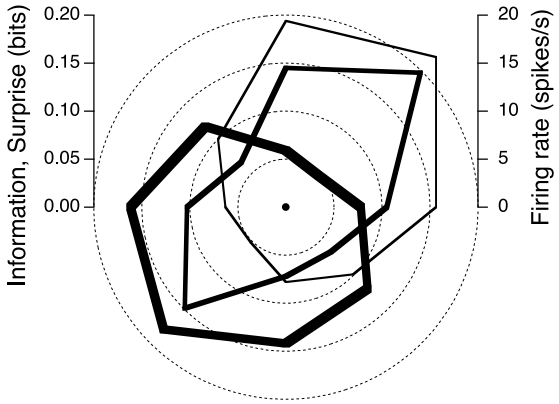

Figure 2. Information and surprise about the neural response obtained from a specific stimulus. A neuron in macaque area MT was stimulated with constant visual motion in eight different directions (for experimental details, see Buracas *et al* (1998)). For each stimulus direction, this polar plot shows the specific information I (thick line) and the specific surprise S (medium line) conveyed about the response by that stimulus. The average firing rate is plotted by the thin line.

grating moving in one of eight possible directions. The neural response y_j was measured as the number of spikes during 1 s of stimulus presentation. Figure 2 illustrates how the average firing rate varies with the stimulus. Like many neurons in the MT area, this cell fires most vigorously to motion in a particular direction (the 'preferred' direction) and least to motion in the opposite (or 'anti-preferred') direction, with intermediate responses for other directions. On any given trial, however, the spike count may vary considerably. The reliability of the response resulting from any given stimulus is captured by the specific information $I(x_i; Y)$. A third curve plots the specific surprise $S(x_i; Y)$ that knowledge of the stimulus conveys about the response. Averaging either the specific information or the specific surprise over all directions results in the full mutual information, as guaranteed by (10) and (13).

The three curves are quite distinct: the firing rate is maximal in the preferred direction, the specific information peaks in the anti-preferred direction, and the specific surprise is bimodal. These qualitative features are easily explained: in response to constant motion stimuli, the variance in the spike count during a given time window is roughly proportional to the average spike count (Shadlen and Newsome 1994, Buracas *et al* 1998). Thus the mean and variance of the spike count distribution $p(y_j|x_i)$ are greatest in response to motion in the preferred direction, and least for motion in the anti-preferred direction. Averaging over all eight equally likely directions ($p(x_i) = \frac{1}{8}$), one obtains a very broad spike count distribution $p(y_j)$ that is peaked at intermediate values. The specific information $I(x_i; Y)$ measures the difference in entropy between this average spike count distribution and the distribution for a specified direction $p(y_j|x_i)$. Since the distribution is most narrow when the stimulus moves in the anti-preferred direction, this corresponds to the greatest specific information. On the other hand, the specific surprise $S(x_i; Y)$ measures the degree of separation between the two spike count distributions. Because the average distribution is peaked at intermediate spike count values, the surprise is large in both the preferred and anti-preferred directions where the overlap is small. This example illustrates that information and surprise describe very different aspects of the communication process in neural systems. Neither of these aspects is contained in the tuning curve, which measures only the average response, not its trial-to-trial variation, and thus reveals nothing about the fidelity of the neural code.

Whereas the above analysis measures the knowledge gained from a specific stimulus about

Table 1. Information and surprise about the stimulus obtained from a specific neural response. A neuron in macaque area MT was stimulated with a moving grating (for experimental details, see Buracas *et al* (1998)): every 16.7 ms, the grating stepped either in the neuron's preferred direction (x_1) or the opposite direction (x_0). The resulting spike count was either above threshold (y_1) or below threshold (y_0). (*a*) The statistical relationships between stimuli x_i and responses y_j. (*b*) The specific information I and specific surprise S obtained about the stimulus from the presence (y_1) or the absence (y_0) of a spike.

(*a*)

x_i	$p(x_i)$	$p(x_i\|y_0)$	$p(x_i\|y_1)$
x_0	0.84	0.89	0.10
x_1	0.16	0.11	0.90

(*b*)

y_j	$I(X; y_j)$	$S(X; y_j)$
y_0	0.12	0.01
y_1	0.14	1.93

the neuron's response, a second experiment served to analyse how much a specific response conveys about the stimulus. Here the stimulus consisted of a grating that moved randomly, executing a short step every 16.7 ms either in the cell's preferred direction (stimulus x_1) or the anti-preferred direction (x_0) (Buracas *et al* 1998). For many MT cells, it was found that 'preferred' steps of the grating elicited action potentials with a specific latency, though not all such steps produced a spike. Thus, the response to each step could be categorized into a binary variable by performing a weighted spike count over a brief time window, and testing whether the result was above a certain threshold (response y_1) or below (y_0). This procedure yielded many stimulus-response pairs (x, y), and from their joint distribution the mutual information $I(X; Y)$ was computed as in (2). After optimizing the stimulus parameters and the procedure for categorizing the spike count, it was found that the response conveyed on average ~ 5.5 bits s^{-1} of information about the stimulus (Buracas *et al* 1998).

Does such a neuron convey more information when it fires or when it remains silent? Table 1 shows the statistical relationship between the stimulus directions X and the binary responses Y for a particular MT cell. Note that the grating moved in the preferred direction in only a minority of the steps, a fraction $\sim$0.16. When the neuron did not respond, the conditional probability for a preferred step was somewhat lower still, $\sim$0.11. On occasions where the neuron did respond, a preferred step was very likely, with conditional probability $\sim$0.90. Because the stimulus distribution changes more dramatically when the neuron fires than when it does not fire, the specific surprise is much greater when observing a spike, $S(X; y_1)$, than when not observing a spike, $S(X; y_0)$. On the other hand, the two conditional distributions given a spike, $p(x_i|y_1)$, or no spike, $p(x_i|y_0)$, are almost perfectly reversed, and thus their entropies are almost identical. As a result, the values for the specific information from either observing a spike, $I(X; y_1)$, or not observing a spike, $I(X; y_0)$, are very similar. We conclude that, at least for this neuron, periods of high- and low-firing rate are almost equally informative about the stimulus.

6. Specific surprise and channel capacity

The information-theoretic approach in neuroscience is partly motivated by the belief that neural systems have reached some performance optimum in processing information to cope with their environment. With few notable exceptions (Bialek *et al* 1991, Rieke *et al* 1997) this idea has rarely been tested. Here we show that one test of optimal performance can be obtained from a measurement of the specific surprise. Since prior studies of event-specific coding in the nervous system all measured the specific surprise rather than the specific information (Eckhorn and Pöpel 1974, 1975, Fuller and Looft 1984, Optican and Richmond 1987, de Ruyter van

Steveninck and Bialek 1988, Bialek and Zee 1990, Richmond and Optican 1990, Theunissen and Miller 1991, Panzeri and Treves 1996, Rolls *et al* 1996, 1998, Buracas *et al* 1998), one can thus draw some additional conclusions about the efficiency of coding in these neural systems.

All these studies concern sensory systems, in which information about a set of stimuli X is represented in neural responses Y. As discussed above, the average information that a response event conveys about the stimulus is given by the mutual information (2). This quantity depends on the conditional probability $p(y_j|x_i)$ that a given stimulus x_i will cause a certain response y_j, and also on the frequency $p(x_i)$ with which the various stimuli are used. The relationship $p(y_j|x_i)$ between stimuli and responses is a property of the circuit, and depends on the function of its neurons and their interconnections. On the other hand, the frequency of input stimuli $p(x_i)$ is a property of the environment in which the circuit operates. Given the circuit properties $p(y_j|x_i)$, there is a maximal possible transmission rate, C, achieved by properly adjusting the stimulus distribution $p(x_i)$: this is called the capacity of the communication channel (Shannon 1948a). If a system transmits information at this full capacity, then the distribution of stimuli is optimally matched to the neural code $p(y_j|x_i)$. Do real sensory systems operate at this optimum?

It is not straightforward to determine the optimal input distribution for a given channel, and much of coding theory is dedicated to this problem (Hamming 1986). However, there is a simple test to check whether a given system is transmitting optimally (Shannon 1948a, p 390, Fano 1961, p 136): a system operates at its capacity C if and only if the specific surprise is equal for all output symbols, that is

$$S(X; y_j) = C \qquad \text{for every symbol } y_j. \tag{14}$$

Since the mutual information is symmetric between inputs and outputs, the same must hold for all input symbols,

$$S(x_i; Y) = C \qquad \text{for every symbol } x_i. \tag{15}$$

By applying this theorem to published measurements of the specific surprise, one finds that these neural systems do not operate at capacity. The deviations from constant surprise per symbol (14), (15) are large, though this does not readily quantify the extent to which coding is suboptimal. For example, de Ruyter van Steveninck and Bialek (1988) recorded the spike train of a visual neuron in the fly under a broad ensemble of visual stimuli, and compared the surprise $S(X; y_j)$ about the stimulus X obtained from different kinds of spike pairs y_j (equation (2) in de Ruyter van Steveninck and Bialek (1988)). The surprise varies considerably depending on the interval between the two spikes; it increases dramatically for very short intervals (figure 7 in de Ruyter van Steveninck and Bialek (1988)), because these encode features that are very rare in the stimulus ensemble. Thus, if one considers spike pairs as the symbols of this neural code, the system clearly does not exhaust the available information capacity per symbol. On the other hand, the information content per symbol may not be the relevant criterion of performance. In particular, short inter-spike intervals clearly require less transmission time than long ones. Thus, maximizing the information per symbol does not maximize the information per unit time, which may be the more pressing concern for a fly involved in visually guided pursuit of a potential mate (Land and Collett 1974). In fact, subsequent work showed that the information rate per unit time in this system comes close to the physically achievable limit (Bialek *et al* 1991).

In a different study, Optican and Richmond (1987) computed the surprise $S(x_i; Y)$ that each visual stimulus x_i conveys about the possible responses Y of neurons in the primate cortex. Again, this quantity varied a great deal across stimuli, by at least a factor of ten, whereas it should be constant if the neural code operated at capacity (figures 2–4 in Optican and Richmond (1987), figure 12 in Richmond and Optican (1990)). The same conclusion can

be drawn from a study of olfactory responses in primate cortex (figures 3–6 in Rolls *et al* (1996)). Finally, the above analysis of direction-tuning in primate MT neurons (figure 2) also documented that the surprise varies considerably with the stimulus, although another neuron from the same population showed less variation (figure 2 in Buracas *et al* (1998)). In these cases it is likely that the stimulus ensembles used during experiments are poorly matched to the neural system's capabilities. This is almost certainly a concern in recordings from the visual cortex using stimulation with static Walsh patterns. These stimuli have mathematically appealing properties, but in practice they lead to very small information rates (Optican and Richmond 1987), about 20-fold lower per spike than what cortical neurons can sustain (Buracas *et al* 1998). Clearly it will be of interest how these performance measures fare under natural stimulation encountered during behaviour (Treves *et al* 1999). Particularly in the case of cortical representations, the analysis should be extended to cover multiple neurons in the population that collectively encodes the stimulus.

7. Conclusion

Shannon's quantitative definition of information has been a powerful tool in analysing communication systems. Here we have discussed how one might extend this measure of information so it applies to specific transmitted symbols. The history of this issue is remarkably sparse. In his founding treatise, Shannon (1948a, b) found no need to define the information related to specific events, but operated only with quantities that are averaged over all symbols, such as the entropy $H(X)$ and the mutual information $I(X; Y)$. Brillouin (1956) pointed out that information from individual observations can be negative. He cites the example of a telegram whose last symbol is either 0 = 'all wrong, pay no attention to this message' or 1 = 'telegram is OK, you can use it'. Observation of 0 destroys all the information that accumulated during observation of the preceding symbols and thus should be viewed as carrying negative information. Brillouin concludes that Shannon's theory must be extended to treat such a case. This extension is, in fact, achieved if one defines the information conveyed by a specific symbol y_j in the manner we propose in (11).

Fano's introduction to the theory (1961) begins by defining a 'microscopic' quantity, namely the information between one specific output symbol y_j and a specific input symbol x_i: $I(x_i; y_j) = \log[p(x_i|y_j)/p(x_i)]$. Note that this quantity is symmetric, $I(x_i; y_j) = I(y_j; x_i)$, and additive, $I(x_i; \{y_j, z_k\}) = I(x_i; y_j)+I(x_i; z_k|y_j)$. From this starting point, Fano computes the 'average amount of information provided by y_j about the transmitted x' by averaging $I(x_i; y_j)$ over all x_i, conditional on observation of y_j: $I_1(X; y_j) = \sum_i p(x_i|y_j)I(x_i; y_j)$. He also considers the alternative definition $I_2(X; y_j) = H(X) - H(X|y_j)$, but rejects it. This is surprising, since Fano—like many other authors—considers the property of additivity an essential 'natural' attribute of information (Fano 1961, p 30–31), and $I_1(X; y_j)$ is not additive, as shown above in (8). Subsequent texts (Abramson 1963, Hamming 1986, Mansuripur 1987) generally follow a similar argument in defining the specific information from an output symbol as $I_1(X; y_j)$. The alternative definition $I_2(X; y_j)$ is rarely even considered (Watanabe 1969, Golomb *et al* 1994), though Watanabe (1969, p 533) remarks in passing that the quantity $I_1(X; y_j)$ is not additive.

For any practical purpose, additivity truly is an essential property of information. Abandoning it leads to absurd results when one combines the information from two or more events, as seen in the examples discussed here. We showed that $I_2(X; y_j) = I(X; y_j)$ is the only expression that satisfies this requirement and thus the preferred choice for measuring the 'event-specific information'. The quantity $I_1(X; y_j) = S(X; y_j)$ also has unique properties, in that it is strictly non-negative, and it can be viewed as the 'event-specific surprise'. Information

and surprise generally report different aspects of the neural code (figure 2, table 1). In particular, a comparison of surprise across symbols in the code allows a test for optimal transmission. Of course, a complete understanding of the neural code—including the degree to which it deviates from optimality—requires not just $I(X; y_j)$ and $S(X; y_j)$, but the full joint distribution of stimuli and responses $p(x_i, y_j)$. A satisfying biological understanding will further require some assessment of the subjective value that different events have for behaviour. Nevertheless, mutual information has been a powerful concept in communication science, and we expect that its logically consistent extension to single symbols, as advocated here, will prove useful.

Appendix

A.1. Proof that $I_2(X; y_j)$ is the only additive measure of specific information

Prior to any observation, the probability distribution of the input X is $p(x)$. After observing the specific event y_j, this distribution has changed from $p(x)$ to

$$q(x) = p(x|y_j).$$

As discussed above (3), the specific information $I(X; y_j)$ obtained from this observation must be a functional of the two distributions $p(x)$ and $q(x)$:

$$I(X; y_j) = F[p(x), q(x)]. \tag{16}$$

If subsequently we observe event z_k, the probability distribution of x changes from $q(x)$ to

$$r(x) = q(x|z_k) = p(x|y_j, z_k).$$

The information gained in this step must be the same functional of the initial and final distributions

$$I(X; z_k|y_j) = F[q(x), r(x)].$$

If we consider both events y_j and z_k as a single combined observation then the probability distribution of X changes from $p(x)$ to $r(x)$, and thus the information gained is

$$I(X; \{y_j, z_k\}) = F[p(x), r(x)].$$

The requirement for additivity (9) specifies that

$$F[p(x), r(x)] = F[p(x), q(x)] + F[q(x), r(x)].$$

It follows directly that

$$\begin{aligned} F[q(x), r(x)] &= F[p(x), r(x)] - F[p(x), q(x)] \\ &= -G[r(x)] + G[q(x)]. \end{aligned}$$

So the specific information $I(X; y_j) = F[p(x), p(x|y_j)]$ must be the difference between the values of some functional $G[\ ]$, evaluated first for the prior distribution $p(x)$ and then for the conditional distribution $p(x|y_j)$

$$I(X; y_j) = G[p(x)] - G[p(x|y_j)].$$

We then show that $G[\ ]$ is, in fact, the entropy functional $H[\ ]$. Consider the special case where

$$p(x_i|y_j) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

that is, any observation of y specifies the precise value of x. In that case

$$G[p(x|y_j)] = K$$

has the same constant value for all y_j, because the argument $p(x|y_j)$ is the delta function in each case. Furthermore, since any observation of y leaves no uncertainty about x, $H(X|Y) = 0$ and the mutual information $I(X, Y)$ is simply the entropy $H(X)$. Thus (10) leads to

$$\sum_j p(y_j)(G[p(x)] - G[p(x|y_j)]) = G[p(x)] - K = H[p(x)] = H(X).$$

So

$$G[p(x)] = H[p(x)] + K$$

and one concludes that

$$\begin{aligned} I(X; y_j) &= G[p(x)] - G[p(x|y_j)] = H[p(x)] - H[p(x|y_j)] = H(X) - H(X|y_j) \\ &= I_2(X; y_j). \end{aligned}$$

A.2. Proof that $I_1(X; y_j)$ is the only non-negative measure that averages to the mutual information

We seek a measure $S(X; y_j)$ that is never negative (12) and whose average over observations y_j is the mutual information (13). As argued for $I(X; y_j)$ in (16), the surprise $S(X; y_j)$ must be a functional of the two distributions $p(x)$ and $p(x|y_j)$, simply because these two distributions completely describe the effect of observing y_j. Thus, we can generally write

$$S(X; y_j) = I_1(X; y_j) + A[p(x), p(x|y_j)] \tag{17}$$

where $I_1(X; y_j)$ is given by (5) and $A[\,]$ is some functional of the two distributions. Then (13) requires that

$$\sum_j p(y_j) A[p(x), p(x|y_j)] = 0. \tag{18}$$

Since this property of $A[\,]$ must be satisfied for all possible forms of $p(x)$ and $p(x|y_j)$, it must somehow arise from the constraints on these probability distributions. The only such constraint relevant to the summation in (18) is

$$\sum_j p(y_j)(p(x_i) - p(x_i|y_j)) = 0. \tag{19}$$

Thus $A[\,]$ must be of the form

$$A[p(x), p(x|y_j)] = \sum_i B_i \cdot (p(x_i) - p(x_i|y_j))$$

where $B_i = B_i[p(x)]$ can be any functional of $p(x)$ but does not depend on y_j.

Now consider the case where $p(x|y_j)$ differs very little from $p(x)$,

$$p(x_i|y_j) = p(x_i) + d_i$$

with $|d_i| \ll 1$ and $\sum_i d_i = 0$. Then (5), (17) and (19) lead to

$$S(X; y_j) = \sum_i d_i \cdot (1 + B_i) + (\text{terms of order } d_i^2).$$

For this to be non-negative for all possible $\{d_i\}$ requires that $B_i = -1$ for all i, and consequently

$$A[p(x_i), p(x_i|y_j)] = \sum_i B_i \cdot (p(x_i) - p(x_i|y_j)) = 0.$$

Thus, one concludes that

$$S(X; y_j) = I_1(X; y_j).$$

Acknowledgments

This work was supported by the Sloan Foundation (MD) and grant EY10020 from the National Eye Institute and a Presidential Faculty Fellowship (MM). Thanks to Giedrius Buracas for providing the cortical recordings and to Timothy Holy for comments on the manuscript.

References

Abramson N 1963 *Information Theory and Coding (McGraw-Hill Electronic Science Series)* (New York: McGraw-Hill)

Aczel J and Daroczy Z 1975 *On Measures of Information and their Characterizations* (New York: Academic)

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

Bialek W and Zee A 1990 Coding and computation with neural spike trains *J. Stat. Phys.* **59** 103–15

Brillouin L 1956 *Science and Information Theory* (New York: Academic)

Buracas G T, Zador A M, DeWeese M R and Albright T D 1998 Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex *Neuron* **20** 959–69

Cover T M and Thomas J A 1991 *Elements of Information Theory* (New York: Wiley)

de Ruyter van Steveninck R and Bialek W 1988 Real-time performance of a movement-sensitive neuron in the blowfly visual system: coding and information transfer in short spike sequence *Proc. R. Soc.* B **234** 379–414

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

Eckhorn R and Pöpel B 1975 Rigorous and extended application of information theory to the afferent visual system of the cat: II. Experimental results *Biol. Cybern.* **17** 71–7

Fano R M 1961 *Transmission of Information; A Statistical Theory of Communications* (New York: MIT)

Fuller M S and Looft F J 1984 An information theoretic analysis of cutaneous receptor responses *IEEE Trans. Biomed. Eng.* **31** 377–83

Golomb S W, Peile R E and Scholz R A 1994 *Basic Concepts in Information Theory and Coding: The Adventures of Secret Agent 00111* (New York: Plenum)

Guiasu S 1977 *Information Theory with Applications* (New York: McGraw-Hill)

Hamming R W 1986 *Coding and Information Theory* 2nd edn (Englewood Cliffs, NJ: Prentice-Hall)

Khinchin A I 1957 *Mathematical Foundations of Information Theory* (New York: Dover Publications)

Krüger J and Aiple F 1988 Multimicroelectrode investigation of monkey striate cortex: spike train correlations in the infragranular layers *J. Neurophysiol.* **60** 798–828

Land M F and Collett T S 1974 Chasing behaviour of houseflies (*Fannia canicularis*). A description and analysis *J. Comp. Physiol.* **89** 331–57

Mansuripur M 1987 *Introduction to Information Theory* (Englewood Cliffs, NJ: Prentice-Hall)

Optican L M and Richmond B J 1987 Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex: III. Information theoretic analysis *J. Neurophysiol.* **57** 162–78

Panzeri S and Treves A 1996 Analytical estimates of limited sampling biases in different information measures *Network: Comput. Neural Syst.* **7** 87–107

Richmond B J and Optican L M 1990 Temporal encoding of two-dimensional patterns by single units in primate primary visual cortex: II. Information transmission *J. Neurophysiol.* **64** 370–80

Rieke F, Warland D, de Ruyter van Steveninck R R and Bialek W 1997 *Spikes: Exploring The Neural Code* (Cambridge, MA: MIT Press)

Rolls E T, Critchley H D and Treves A 1996 Representation of olfactory information in the primate orbitofrontal cortex *J. Neurophysiol.* **75** 1982–96

Rolls E T, Treves A, Robertson R G, Georges-Francois P and Panzeri S 1998 Information about spatial view in an ensemble of primate hippocampal cells *J. Neurophysiol.* **79** 1797–813

Shadlen M N and Newsome W T 1994 Noise, neural codes and cortical organization *Curr. Opin. Neurobiol.* **4** 569–79

Shannon C E 1948a A mathematical theory of communication, Part 1 *Bell Syst. Tech. J.* **27** 379–423

——1948b A mathematical theory of communication, Part 2 *Bell Syst. Tech. J.* **27** 623–56

Shannon C E and Weaver W 1963 *The Mathematical Theory of Communication* 2nd edn (Chicago, IL: University of Illinois Press)

Theunissen F E and Miller J P 1991 Representation of sensory information in the cricket cercal sensory system: II. Information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons *J. Neurophysiol.* **66** 1690–703

Treves A, Panzeri S, Rolls E T, Booth M and Wakeman E A 1999 Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli *Neural Comput.* **11** 601–32

Warland D K, Reinagel P and Meister M 1997 Decoding visual information from a population of retinal ganglion cells *J. Neurophysiol.* **78** 2336–50

Watanabe S 1969 *Knowing and Guessing; A Quantitative Study of Inference and Information* (New York: Wiley)

Wilson M A and McNaughton B L 1993 Dynamics of the hippocampal ensemble code for space *Science* **261** 1055–8