Primer

Information theory in the brain

Pamela Reinagel

Claude Shannon's classic 1948 paper introduced a general theory for measuring the transmission of information from a source, across a noisy channel, to a receiver (Figure 1). This theory became known as *information theory*. Shannon illustrated his theory with such examples as messages sent over telegraph lines by Morse code. But the same principles can be applied to neurons: a neuron transmits information along its axon to other neurons, using a neural code. Information theory has proved to be a powerful tool for quantifying the communication of information by neurons.

Bits

Shannon argued that information should be measured in the now-familiar units of bits. A bit is a binary digit (0 or 1), or the amount of information needed to give an answer to a single yes-or-no question. The outcome of one coin toss could be reported in one bit (`0` for heads, `1` for tails). The result of two coin tosses could be reported in two bits, which can represent four possible combinations (`00`, `01`, `10`, `11`). In general, with n coin tosses there are $N=2^n$ possible outcomes. Turning this around, if there are 2^n possible combinations or messages, then there are $\log_2(2^n) = n$ bits of information. The reason $\log_2$ shows up so often in information theory is that the unit of information is the binary (base 2) digit.

Information content

The information content in a source can be measured as the minimum number of bits that could represent the messages from that source. This quantity is called the *entropy* of the source, because its equation has the same form as the entropy in statistical thermodynamics. In the simplest case, when there are N possible messages and all are equally likely, the entropy is $\log_2 N$, as for the coin tosses above.

The situation is more complicated if two coins are weighted, so that each of them comes up heads in 95% of tosses. At first, it seems we still need two bits to represent the four possibilities. But we could instead use a short code, `1`, to designate the usual outcome, and longer codes for the three uncommon outcomes, for example `0` as a prefix for rare outcomes followed by `00`, `01` or `10` to distinguish them.This scheme would use 1 bit $0.95 \times 0.95 \approx 90\%$ of the time, and 3 bits 10% of the time, hence on average using 1.2 bits instead of 2. But is 1.2 the minimum number of bits that could represent this source?

According to information theory, a message (outcome) that occurs with probability P has an entropy of $\log_2(1/P)$ bits. For a simple coin toss, each outcome has a probability P=1/2 and thus an entropy of $\log_2(2) = 1$ bit. In the case of the two weighted coins, the outcome two heads is so likely that its actual occurrence does not contain much new information: P=0.9, and $\log_2(1/0.9) \approx 0.15$ bits. By contrast, the outcome two tails would be very surprising, and thus highly informative: P=1/400, and $\log_2(400) \approx 8.6$ bits. Shannon proved that the entropy of an information source is the sum over all possible messages of the entropy of the message, $\log_2(1/P)$, weighted by how often it occurs (P):

$$\text{Entropy} = \sum \text{P} \log_2\left(\frac{1}{\text{P}}\right)$$

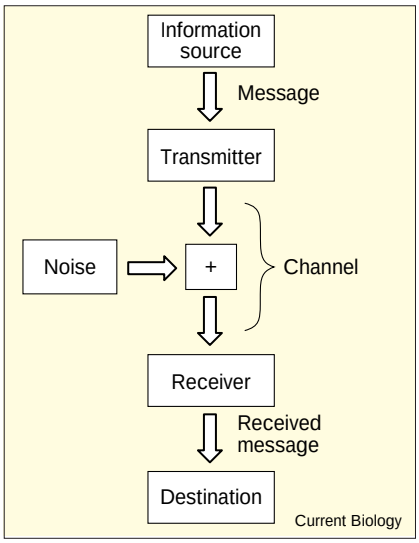
In the case of the two weighted coins above, the entropy turns out to be only 0.6 bits. This means an optimal code could be twice as efficient as the code proposed above, which used 1.2 bits. Note that the entropy of the two weighted coins is much less than two fair coins (2 bits). In general, a system has maximum entropy when every possible message is equally likely. When probabilities are unequal, the entropy is always reduced. At the extreme, when one message has a probability P=1 and all others P=0, the entropy is 0.

Finally, we have been assuming that each message is independent of the others. Entropy is reduced if the probability of a message depends on previous ones. For example, the letter u is fairly uncommon in English text, but after a q it is almost always present. Thus the letter u has high information in general, but almost no information when following q. When messages are not independent, it is necessary to measure the probabilities of sequences of messages to determine the entropy of the source.

Information transmission

Between the source and the receiver, information may become

Figure 1



Shannon's schematic of a general communication system. An information source produces a message, which is then transmitted over a noisy channel to the receiver. The message received at the destination is a corrupted version of the original message.

corrupted by noise (Figure 1). The amount of information that gets through can be measured by how much information the sent and received messages have in common — their *mutual information*. To choose an example from neuroscience, a visual display on which a stimulus is shown could be considered an information source. All the optical and neural events leading up to the firing of a neuron could be taken as the channel, and the neuron's firing rate could be taken as the message received at the destination. We can then ask: how much information about the stimulus is represented by the cell's firing rate?

Suppose the stimulus, S, has two possible values — black and white —chosen randomly on each experimental trial with equal probability. The minimum number of bits required to represent this stimulus is 1 bit, so this is the stimulus entropy. One could classify the cell's response, R, as either on or off in each trial, and choose the threshold in such a way that the cell is considered on in exactly half the trials. Defined this way, the response also has an entropy of 1 bit.

To find out how much information this response contains about the stimulus, we tabulate the probability of each combination of S and R (Figure 2). The response of the neuron in Figure 2a is completely independent of the stimulus, as we might find if we were recording from a purely auditory neuron in response to this visual stimulus. The probability of any combination of stimulus and response is the same as expected by chance. Thus, even though the response has 1 bit of entropy, it contains no information about the stimulus. The response of the neuron in Figure 2b, however, is a perfectly reliable indicator of the stimulus. It contains all the information in the stimulus, 1 bit. A noisy visual neuron, whose response was partially determined by the stimulus, would fall between these extremes.

Mutual information

Shannon showed that the mutual information between R and S can be calculated from the joint probability distribution (Figure 2) by the equation:

$$I = \sum_S \sum_R P(S,R) \log_2 \left(\frac{P(S,R)}{P(S) \times P(R)} \right)$$

This seems more complicated than it is. The joint probability of stimulus S and response R (one shaded square) is written P(S,R). The ratio compares this joint probability to what might happen by chance – the product of the two individual probabilities $P(S) \times P(R)$. If these two are equal, as for the neuron in Figure 2a, then the ratio is 1, and $\log_2(1) = 0$ bits. If the joint probability differs from chance, information is encoded. For example, for the neuron in Figure 2b the joint probability of getting an on response with a black stimulus is 0.5, compared to the expected $0.5 \times 0.5 = 0.25$, so the ratio is 2, and $\log_2(2) = 1$ bit. Finally, $\Sigma_S \Sigma_R P(S,R)$ simply indicates taking a sum over all stimuli and all responses, weighted according to how often the combination occurs.

Figure 2

(a)

Response / Stimulus	Off $p = 0.5$	On $p = 0.5$
White $p = 0.5$	0.25	0.25
Black $p = 0.5$	0.25	0.25

(b)

Response / Stimulus	Off $p = 0.5$	On $p = 0.5$
White $p = 0.5$	0.5	0.0
Black $p = 0.5$	0.0	0.5

Current Biology

Responses of two neurons during presentation of a random, binary stimulus. The neuron in **(a)** codes no information, whereas the cell in **(b)** is a perfect encoder.

In most neurobiology experiments, both the stimuli and the responses are more complicated than this example. Some stimuli have hundreds or thousands of bits of information per second. For example, a stimulus could have many distinct shades of gray, with different shades at different spatial locations, and the shades could be changed hundreds of times per second, to make a black-and-white movie. Neural responses are also much richer than simply on and off. For example, a response could be defined by the exact time of each action potential fired by each cell in a large population of sensory neurons. Even for fairly simple experiments, it is usually impossible to collect enough experimental data to fill in a table like those in Figure 2. Fortunately, several sophisticated tricks have been devised to estimate the mutual information in other ways (for one example, see Figure 3).

Cracking neural codes

Information theory has ben used to study how neurons encode information. Firing rates have long been known to be important for neural codes; the firing rates of single neurons often correlate systematically with experimental parameters we can vary or measure in the laboratory, such as the intensity of a sensory stimulus or the magnitude of a muscular contraction. Using information theory, we can put

a number on the amount of information encoded by the firing rate of a neuron. We can then ask whether additional information is transmitted by other features of the neural response.

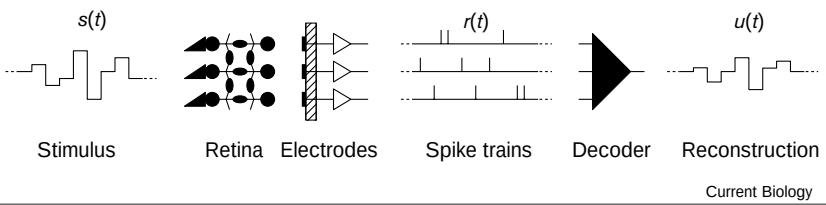
The most striking lesson to emerge over the past decade is that the exact timing of action potentials is important in neural coding, particularly when the encoded signal is itself rapidly varying. For the most part, this is an extension of the idea that firing rates code information, but we now think of a neuron's rate (probability of firing) as varying on the timescale of milliseconds rather than seconds. There has been intense interest in whether information is encoded in other aspects of neural responses, such as temporal patterns of firing, or firing patterns involving multiple cells. In recent years, a few examples of such temporal and population coding have been demonstrated, although so far the amount of information involved has been modest. It remains controversial whether these codes will turn out to be functionally important.

From information theory, we know that the most efficient code to transmit a signal depends on the statistics of the signal. This is why Morse code uses a short symbol (•) for the frequent letter *e*, and a clumsy long one (— — • —) for the rare *q*. But for a language with few *e*s and many *q*s, this would be an especially bad code. Similarly, a neural code that represented one class of sensory stimuli efficiently would necessarily be inefficient for other stimuli. It has been proposed that sensory neurons evolved to send as much information to the brain as possible under real-world conditions. Early tests of this theory have confirmed the prediction that sensory neural codes seem to be specifically adapted for the statistics of their natural stimuli.

Decoding a neural code

One approach to studying neural codes is to try to decode a neural response and reconstruct the original signal. Any information about the stimulus that can be successfully reconstructed is thereby demonstrated to be encoded by the neuron(s). The successful decoding algorithm may also provide insight into how that information is represented, and how other neurons might extract it. For example, suppose one has recorded the spike trains of several visual neurons in response to a randomly modulating stimulus (Figure 3). One could try to find a decoder that would estimate the visual stimulus from the responses of the cells. With judicious choice of the input signal, efficient methods can be used to compare the reconstruction to the original signal, to obtain an information estimate without recourse to a joint probability table (Figure 2). In this way mutual information can be estimated with comparatively little data.

Figure 3



The luminance of a visual stimulus *s*(*t*) is chosen independently at each time step. The responses of retinal ganglion cells *r*(*t*) are recorded on multiple electrodes. A decoder is sought that can produce a reconstruction, *u*(*t*), that matches the original stimulus as well as possible. (Adapted with permission from Warland DK, Reinagel P, Meister M: *J Neurophysiol* 1997, **78:**2336-2350.)

Broader applications

Information theory is a completely general method to measure the transfer of information from one place to another, and is particularly suited to describing many aspects of neural function. Information theory has been applied most widely in the field of sensory coding, but it is equally applicable to the transmission of neural commands for motor output patterns. One can also measure the mutual information between the spike train of one neuron and that of its postsynaptic target, or between a single cell's synaptic currents and its action potentials.

The term *information* suggests to us something about the intention of a sender or the value to the receiver. But, in the uses of information theory presented above, neither is implied. For example, there is more information content (entropy) in a randomly flickering visual stimulus than in a real-world scene, even though the flickering is meaningless. In the future it will be important for the field to address such additional issues as: the value of different stimulus information for the animal; the role of active exploration of the enviroment in selection stimuli for encoding; and the mechanisms in the brain for decoding the information carried by neural responses.

Key references

Shannon CE: **A mathematical theory of communication.** *Bell System Tech J* 1948, **27:**379-423;623-656. (Reprinted in *Claude Elwood Shannon: Collected Papers.* Edited by Sloane NJA and Wyner AD. New York: IEEE Press; 1993. Also available at http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html)

Rieke FM, Warland DK, de Ruyter van Steveninck R, Bialek W: *Spikes: Exploring the neural code.* Cambridge, Massachusetts: MIT Press; 1997

Cover TM, Thomas JA: *Elements of Information Theory.* New York: Wiley; 1991.

Address: Department of Neuroscience, Harvard Medical School, 220 Longwood Avenue, Boston, Massachusetts 02115, USA.