4 Information Theory

4.1 Entropy and Mutual Information

Neural encoding and decoding focus on the question "What does the response of a neuron tell us about a stimulus?" In this chapter we consider a related but different question "How much does the neural response tell us about a stimulus?" The techniques of information theory allow us to answer this question in a quantitative manner. Furthermore, we can use them to ask what forms of neural response are optimal for conveying information about natural stimuli. Information theoretic principles play an important role in many of the unsupervised learning methods that are discussed in chapters 8 and 10.

Shannon invented information theory as a general framework for quantifying the ability of a coding scheme or a communication channel (such as the optic nerve) to convey information. It is assumed that the code involves a number of symbols (such as different neuronal responses), and that the coding and transmission processes are stochastic and noisy. The quantities we consider in this chapter, the entropy and the mutual information, depend on the probabilities with which these symbols, or combinations of them, are used. Entropy is a measure of the theoretical capacity of a code to convey information. Mutual information measures how much of that capacity is actually used when the code is employed to describe a particular set of data. Communication channels, if they are noisy, have only limited capacities to convey information. The techniques of information theory are used to evaluate these limits and find coding schemes that saturate them.

In neuroscience applications, the symbols we consider are neuronal responses, and the data sets they describe are stimulus characteristics. In the most complete analyses, which are considered at the end of the chapter, the neuronal response is characterized by a list of action potential firing times. The symbols being analyzed in this case are sequences of action potentials. Computing the entropy and mutual information for spike sequences can be difficult because the frequency of occurrence of many different spike sequences must be determined. This typically requires a large amount of

data. For this reason, many information theory analyses use simplified descriptions of the response of a neuron that reduce the number of possible "symbols" (i.e., responses) that need to be considered. We discuss cases in which the symbols consist of responses described by spike-count firing rates. We also consider the extension to continuous-valued firing rates. Because a reduced description of a spike train can carry no more information than the full spike train itself, this approach provides a lower bound on the actual information carried by the spike train.

Entropy

Entropy is a quantity that, roughly speaking, measures how "interesting" or "surprising" a set of responses is. Suppose that we are given a set of neural responses. If each response is identical, or if only a few different responses appear, we might conclude that this data set is relatively uninteresting. A more interesting set might show a larger range of different responses, perhaps in a highly irregular and unpredictable sequence. How can we quantify this intuitive notion of an interesting set of responses?

We begin by characterizing the responses in terms of their spike-count firing rates (i.e., the number of spikes divided by the trial duration), which can take a discrete set of different values. The methods we discuss are based on the probabilities $P[r]$ of observing a response with a spike-count rate r. The most widely used measure of entropy, due to Shannon, expresses the "surprise" associated with seeing a response rate r as a function of the probability of getting that response, $h(P[r])$, and quantifies the entropy as the average of $h(P[r])$ over all possible responses. The function $h(P[r])$, which acts as a measure of surprise, is chosen to satisfy a number of conditions. First, $h(P[r])$ should be a decreasing function of $P[r]$ because low probability responses are more surprising than high probability responses. Further, the surprise measure for a response that consists of two independent spike counts should be the sum of the measures for each spike count separately. This assures that the entropy and information measures we ultimately obtain will be additive for independent sources. Suppose we record rates r_1 and r_2 from two neurons that respond independently of each other. Because the responses are independent, the probability of getting this pair of responses is the product of their individual probabilities, $P[r_1]P[r_2]$, so the additivity condition requires that

$$h(P[r_1]P[r_2]) = h(P[r_1]) + h(P[r_2])\,. \tag{4.1}$$

The logarithm is the only function that satisfies such an identity for all P. Thus, it only remains to decide what base to use for the logarithm. By convention, base 2 logarithms are used so that information can be compared easily with results for binary systems. To indicate that the base 2 logarithm is being used, information is reported in units of "bits", with

bits

$$h(P[r]) = -\log_2 P[r]\,. \tag{4.2}$$

The minus sign makes h a decreasing function of its argument, as required. Note that information is really a dimensionless number. The bit, like the radian for angles, is not a dimensional unit but a reminder that a particular system is being used.

Expression (4.2) quantifies the surprise or unpredictability associated with a particular response. Shannon's entropy is just this measure averaged over all responses,

entropy

$$H = -\sum_r P[r] \log_2 P[r] \,. \tag{4.3}$$

In the sum that determines the entropy, the factor $h = -\log_2 P[r]$ is multiplied by the probability that the response with rate r occurs. Responses with extremely low probabilities may contribute little to the total entropy, despite having large h values, because they occur so rarely. In the limit when $P[r] \to 0$, $h \to \infty$, but an event that does not occur does not contribute to the entropy because the problematic expression $-0 \log_2 0$ is evaluated as $-\epsilon \log_2 \epsilon$ in the limit $\epsilon \to 0$, which is 0. Very high probability responses also contribute little because they have $h \approx 0$. The responses that contribute most to the entropy have high enough probabilities so that they appear with a fair frequency, but not high enough to make h too small.

Computing the entropy in some simple cases helps provide a feel for what it measures. First, imagine the least interesting situation: when a neuron responds every time by firing at the same rate. In this case, all of the probabilities $P[r]$ are 0, except for one of them, which is 1. This means that every term in the sum of equation (4.3) is 0 because either $P[r] = 0$ or $\log_2 1 = 0$. Thus, a set of identical responses has zero entropy. Next, imagine that the neuron responds in only two possible ways, either with rate r_+ or r_-. In this case, there are only two nonzero terms in equation (4.3), and, using the fact that $P[r_-] = 1 - P[r_+]$, the entropy is

$$H = -(1 - P[r_+]) \log_2 (1 - P[r_+]) - P[r_+] \log_2 P[r_+] \,. \tag{4.4}$$

This entropy, plotted in figure 4.1A, takes its maximum value of 1 bit when $P[r_-] = P[r_+] = 1/2$. Thus, a code consisting of two equally likely responses has one bit of entropy.

Mutual Information

To convey information about a set of stimuli, neural responses must be different for different stimuli. Entropy is a measure of response variability, but it does not tell us anything about the source of that variability. A neuron can provide information about a stimulus only if its response variability is correlated with changes in that stimulus, rather than being purely random or correlated with other unrelated factors. One way to determine whether response variability is correlated with stimulus variability is to compare the responses obtained using a different stimulus on every trial with those measured in trials involving repeated presentations of the same

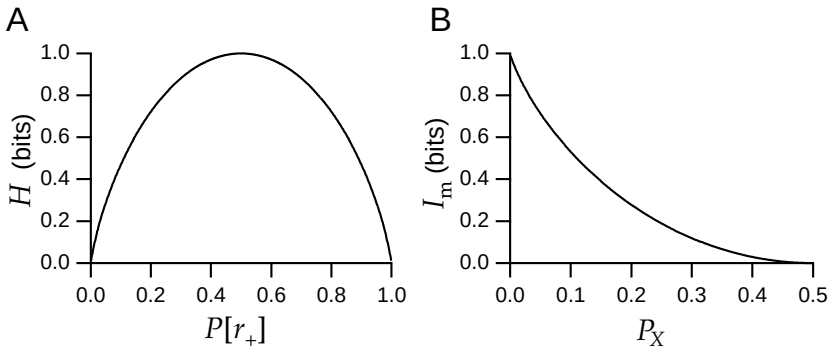

Figure 4.1 (A) The entropy of a binary code. $P[r_+]$ is the probability of a response at rate r_+, and $P[r_-] = 1 - P[r_+]$ is the probability of the other response, r_-. The entropy is maximum when $P[r_-] = P[r_+] = 1/2$. (B) The mutual information for a binary encoding of a binary stimulus. P_X is the probability of an incorrect response being evoked. The plot shows only $P_X \leq 1/2$ because values of $P_X > 1/2$ correspond to an encoding in which the relationship between the two responses and the two stimuli is reversed and the error probability is $1 - P_X$.

stimulus. Responses that are informative about the identity of the stimulus should exhibit larger variability for trials involving different stimuli than for trials that use the same stimulus repetitively. Mutual information is an entropy-based measure related to this idea.

The mutual information is the difference between the total response entropy and the average response entropy on trials that involve repetitive presentation of the same stimulus. Subtracting the entropy when the stimulus does not change removes from the total entropy the contribution from response variability that is not associated with the identity of the stimulus. When the responses are characterized by a spike-count rate, the total response entropy is given by equation 4.3. The entropy of the responses evoked by repeated presentations of a given stimulus s is computed using the conditional probability $P[r|s]$, the probability of a response at rate r given that stimulus s was presented, instead of the response probability $P[r]$ in equation 4.3. The entropy of the responses to a given stimulus is thus

$$H_s = -\sum_r P[r|s] \log_2 P[r|s] \,. \tag{4.5}$$

noise entropy

If we average this quantity over all the stimuli, we obtain a quantity called the noise entropy

$$H_{\text{noise}} = \sum_s P[s] H_s = -\sum_{s,r} P[s] P[r|s] \log_2 P[r|s] \,. \tag{4.6}$$

This is the entropy associated with that part of the response variability that is not due to changes in the stimulus, but arises from other sources. The mutual information is obtained by subtracting the noise entropy from the

full response entropy, which from equations 4.3 and 4.6 gives

$$I_{\mathrm{m}} = H - H_{\mathrm{noise}} = -\sum_r P[r]\log_2 P[r] + \sum_{s,r} P[s]P[r|s]\log_2 P[r|s]\,. \quad (4.7)$$

The probability of a response r is related to the conditional probability $P[r|s]$ and the probability $P[s]$ that stimulus s is presented by the identity (chapter 3),

$$P[r] = \sum_s P[s]P[r|s]\,. \quad (4.8)$$

Using this, and writing the difference of the two logarithms in equation 4.7 as the logarithm of the ratio of their arguments, we can rewrite the mutual information as

mutual information

$$I_{\mathrm{m}} = \sum_{s,r} P[s]P[r|s]\log_2\left(\frac{P[r|s]}{P[r]}\right)\,. \quad (4.9)$$

Recall from chapter 3 that

$$P[r,s] = P[s]P[r|s] = P[r]P[s|r]\,, \quad (4.10)$$

where $P[r,s]$ is the joint probability of stimulus s appearing and response r being evoked. Equation 4.10 can be used to derive yet another form for the mutual information,

$$I_{\mathrm{m}} = \sum_{s,r} P[r,s]\log_2\left(\frac{P[r,s]}{P[r]P[s]}\right)\,. \quad (4.11)$$

This equation reveals that the mutual information is symmetric with respect to interchange of s and r, which means that the mutual information that a set of responses conveys about a set of stimuli is identical to the mutual information that the set of stimuli conveys about the responses. To see this explicitly, we apply equation 4.10 again to write

$$I_{\mathrm{m}} = -\sum_s P[s]\log_2 P[s] + \sum_{s,r} P[r]P[s|r]\log_2 P[s|r]\,. \quad (4.12)$$

This result is the same as equation 4.7, except that the roles of the stimulus and the response have been interchanged. Equation 4.12 shows how response variability limits the ability of a spike train to carry information. The second term on the right side, which is negative, is the average uncertainty about the identity of the stimulus given the response, and reduces the total stimulus entropy represented by the first term.

To provide some concrete examples, we compute the mutual information for a few simple cases. First, suppose that the responses of the neuron are completely unaffected by the identity of the stimulus. In this case, $P[r|s] = P[r]$, and from equation 4.9 it follows immediately that $I_{\mathrm{m}} = 0$. At the other extreme, suppose that each stimulus s produces a unique and

distinct response r_s. Then, $P[r_s] = P[s]$ and $P[r|s]$ is 1 if $r = r_s$ and 0 otherwise. This causes the sum over r in equation 4.9 to collapse to just one term, and the mutual information becomes

$$I_{\rm m} = \sum_s P[s] \log_2 \left(\frac{1}{P[r_s]} \right) = -\sum_s P[s] \log_2 P[s] \,. \tag{4.13}$$

The last expression, which follows from the fact that $P[r_s] = P[s]$, is the entropy of the stimulus. Thus, with no variability and a one-to-one map from stimulus to response, the mutual information is equal to the full stimulus entropy.

Finally, imagine that there are only two possible stimulus values, which we label $+$ and $-$, and that the neuron responds with just two rates, r_+ and r_-. We associate the response r_+ with the $+$ stimulus, and the response r_- with the $-$ stimulus, but the encoding is not perfect. The probability of an incorrect response is $P_{\rm X}$, meaning that for the correct responses $P[r_+|+] = P[r_-|-] = 1 - P_{\rm X}$, and for the incorrect responses $P[r_+|-] = P[r_-|+] = P_{\rm X}$. We assume that the two stimuli are presented with equal probability so that $P[r_+] = P[r_-] = 1/2$, which, from equation 4.4, makes the full response entropy 1 bit. The noise entropy is $-(1 - P_{\rm X}) \log_2 (1 - P_{\rm X}) - P_{\rm X} \log_2 P_{\rm X}$. Thus, the mutual information is

$$I_{\rm m} = 1 + (1 - P_{\rm X}) \log_2 (1 - P_{\rm X}) + P_{\rm X} \log_2 P_{\rm X} \,. \tag{4.14}$$

This is plotted in figure 4.1B. When the encoding is error-free ($P_{\rm X} = 0$), the mutual information is 1 bit, which is equal to both the full response entropy and the stimulus entropy. When the encoding is random ($P_{\rm X} = 1/2$), the mutual information goes to 0.

It is instructive to consider this example from the perspective of decoding. We can think of the neuron as being a communication channel that reports noisily on the stimulus. From this perspective, we want to know the probability that a $+$ was presented, given that the response r_+ was recorded. By Bayes theorem, this is $P[+|r_+] = P[r_+|+]P[+]/P[r_+] = 1 - P_{\rm X}$. Before the response is recorded, the expectation was that $+$ and $-$ were equally likely. If the response r_+ is recorded, this expectation changes to $1 - P_{\rm X}$. The mutual information measures the corresponding reduction in uncertainty or, equivalently, the tightening of the posterior distribution due to the response.

KL divergence

The mutual information is related to a measure used in statistics called the Kullback-Leibler (KL) divergence. The KL divergence between one probability distribution $P[r]$ and another distribution $Q[r]$ is

$$D_{\rm KL}(P, Q) = \sum_r P[r] \log_2 \left(\frac{P[r]}{Q[r]} \right) . \tag{4.15}$$

The KL divergence has a property normally associated with a distance measure, $D_{\rm KL}(P, Q) \geq 0$ with equality if and only if $P = Q$ (proven in appendix A). However, unlike a distance, it is not symmetric with respect to

interchange of P and Q. Comparing the definition 4.15 with equation 4.11, we see that the mutual information is the KL divergence between the distributions $P[r, s]$ and $P[r]P[s]$. If the stimulus and the response were independent of one another, $P[r, s]$ would be equal to $P[r]P[s]$. Thus, the mutual information is the KL divergence between the actual probability distribution $P[r, s]$ and the value it would take if the stimulus and response were independent. The fact that $D_{\rm KL} \geq 0$ proves that the mutual information cannot be negative. In addition, it can never be larger than either the full response entropy or the entropy of the stimulus set.

Entropy and Mutual Information for Continuous Variables

Up to now we have characterized neural responses using discrete spike-count rates. As in chapter 3, it is often convenient to treat these rates instead as continuous variables. There is a complication associated with entropies that are defined in terms of continuous response variables. If we could measure the value of a continuously defined firing rate with unlimited accuracy, it would be possible to convey an infinite amount of information using the endless sequence of decimal digits of this single variable. Of course, practical considerations always limit the accuracy with which a firing rate can be measured or conveyed.

To define the entropy associated with a continuous measure of a neural response, we must include some limit on the measurement accuracy. The effects of this limit typically cancel in computations of mutual information because the mutual information is the difference between two entropies. In this section, we show how entropy and mutual information are computed for responses characterized by continuous firing rates. For completeness, we also treat the stimulus parameter s as a continuous variable. This means that the probability $P[s]$ is replaced by the probability density $p[s]$, and sums over s are replaced by integrals.

For a continuously defined firing rate, the probability of the firing rate lying in the range between r and $r + \Delta r$, for small Δr, is expressed in terms of a probability density as $p[r]\Delta r$. Summing over discrete bins of size Δr, we find, by analogy with equation (4.3),

$$\begin{aligned} H &= -\sum p[r]\Delta r \log_2(p[r]\Delta r) \\ &= -\sum p[r]\Delta r \log_2 p[r] - \log_2 \Delta r \,. \end{aligned} \tag{4.16}$$

To extract the last term we have expressed the logarithm of a product as the sum of two logarithms and used the fact that the sum of the response probabilities is 1. We would now like to take the limit $\Delta r \to 0$ but we cannot, because the $\log_2 \Delta r$ term diverges in this limit. This divergence reflects the fact that a continuous variable measured with perfect accuracy has infinite entropy. However, for reasonable (i.e., Riemann integrable) $p[r]$, everything works out fine for the first term because the sum becomes

an integral in the limit $\Delta r \to 0$. In this limit, we can write

$$\lim_{\Delta r \to 0} \{H + \log_2 \Delta r\} = -\int dr\, p[r] \log_2 p[r]\,. \tag{4.17}$$

continuous entropy

Δr is best thought of as a limit on the resolution with which the firing rate can be measured. Unless this limit is known, the entropy of a probability density for a continuous variable can be determined only up to an additive constant. However, if two entropies computed with the same resolution are subtracted, the troublesome term involving Δr cancels, and we can proceed without knowing its precise value. All of the cases where we use equation 4.17 are of this form. The integral on the right side of equation 4.17 is sometimes called the differential entropy.

differential entropy

The noise entropy, for a continuous variable like the firing rate, can be written in a manner similar to the response entropy 4.17, except that the conditional probability density $p[r|s]$ is used:

continuous noise entropy

$$\lim_{\Delta r \to 0} \{H_{\text{noise}} + \log_2 \Delta r\} = -\int ds \int dr\, p[s] p[r|s] \log_2 p[r|s]\,. \tag{4.18}$$

The mutual information is the difference between the expressions in equations 4.17 and 4.18,

continuous mutual information

$$I_{\text{m}} = \int ds \int dr\, p[s] p[r|s] \log_2 \left(\frac{p[r|s]}{p[r]} \right). \tag{4.19}$$

Note that the factor of $\log_2 \Delta r$ cancels in the expression for the mutual information because both entropies are evaluated at the same resolution.

In chapter 3, we described the Fisher information as a local measure of how tightly the responses determine the stimulus. The Fisher information is local because it depends on the expected curvature of the likelihood $P[\mathbf{r}|s]$ (typically for the responses of many cells) evaluated at the true stimulus value. The mutual information is a global measure in the sense that it depends on the average overall uncertainty in the decoding distribution $p[s|\mathbf{r}]$, including values of s both close to and far from the true stimulus. If the decoding distribution $p[s|\mathbf{r}]$ has a single peak about the true stimulus, the Fisher information and the mutual information are closely related. In particular, for large numbers of neurons, the maximum likelihood estimator tends to have a sharply peaked Gaussian distribution, as discussed in chapter 3. In this case, the mutual information is, up to an additive constant, the logarithm of the Fisher information averaged over the distribution of stimuli.

4.2 Information and Entropy Maximization

Entropy and mutual information are useful quantities for characterizing the nature and efficiency of neural encoding and selectivity. Often, in addition to such characterizations, we seek to understand the computational

implications of an observed response selectivity. For example, we might ask whether neural responses to natural stimuli are optimized to convey as much information as possible. This hypothesis can be tested by computing the response characteristics that maximize the mutual information conveyed about naturally occurring stimuli and comparing the results with responses observed experimentally.

Because the mutual information is the full response entropy minus the noise entropy, maximizing the information involves a compromise. We must make the response entropy as large as possible without allowing the noise entropy to get too big. If the noise entropy is small, maximizing the response entropy, subject to an appropriate constraint, maximizes the mutual information to a good approximation. We therefore begin our discussion by studying how response entropy can be maximized. Later in the discussion, we will consider the effects of noise entropy.

Constraints play a crucial role in this analysis. We have already seen that the theoretical information-carrying capacity associated with a continuous firing rate is limited only by the resolution with which the firing rate can be defined. Even with a finite resolution, a firing rate could convey an infinite amount of information if it could take arbitrarily high values. Thus, we must impose some constraint that limits the firing rate to a realistic range. Possible constraints include limiting the maximum allowed firing rate or holding the average firing rate or its variance fixed.

Entropy Maximization for a Single Neuron

To maximize the response entropy, we must find a probability density $p[r]$ that makes the integral in equation 4.17 as large as possible while satisfying whatever constraints we impose. During the maximization process, the resolution Δr is held fixed, so the $\log_2 \Delta r$ term remains constant, and it can be ignored. As a result, it will not generally appear in the following equations. One constraint that always applies in entropy maximization is that the integral of the probability density must be 1. Suppose that the neuron in question has a maximum firing rate of $r_{\max}$. Then, the integrals in question extend from 0 to $r_{\max}$. To find the $p[r]$ producing the maximum entropy, we must maximize

$$-\int_0^{r_{\max}} dr\, p[r] \log_2 p[r]\,, \tag{4.20}$$

subject to the constraint

$$\int_0^{r_{\max}} dr\, p[r] = 1\,. \tag{4.21}$$

The result, computed using Lagrange multipliers (see the Mathematical Appendix), is that the probability density that maximizes the entropy sub-

ject to this constraint is a constant,

$$p[r] = \frac{1}{r_{\max}}\,, \tag{4.22}$$

independent of r. The entropy for this probability density, for finite firing-rate resolution Δr, is

$$H = \log_2 r_{\max} - \log_2 \Delta r = \log_2 \left(\frac{r_{\max}}{\Delta r}\right)\,. \tag{4.23}$$

histogram equalization

Equation 4.22 is the basis of a signal-processing technique called histogram equalization. Applied to neural responses, this is a procedure for tailoring the neuronal selectivity so that $p[r] = 1/r_{\max}$ in response to a set of stimuli over which the entropy is to be maximized. Suppose a neuron responds to a stimulus characterized by the parameter s by firing at a rate $r = f(s)$. For small Δs, the probability that the continuous stimulus variable falls in the range between s and $s + \Delta s$ is given in terms of the stimulus probability density by $p[s]\Delta s$. This produces a response that falls in the range between $f(s + \Delta s)$ and $f(s)$. If the response probability density takes its optimal value, $p[r] = 1/r_{\max}$, the probability that the response falls within this range is $|f(s + \Delta s) - f(s)|/r_{\max}$. Setting these two probabilities equal to each other, we find that $|f(s + \Delta s) - f(s)|/r_{\max} = p[s]\Delta s$.

Consider the case of a monotonically increasing response so that $f(s + \Delta s) > f(s)$ for positive Δs. Then, in the limit $\Delta s \to 0$, the equalization condition becomes

$$\frac{df}{ds} = r_{\max}\, p[s]\,, \tag{4.24}$$

which has the solution

$$f(s) = r_{\max} \int_{s_{\min}}^{s} ds'\, p[s']\,, \tag{4.25}$$

where $s_{\min}$ is the minimum value of s, which is assumed to generate no response. Thus, entropy maximization requires that the average firing rate of the responding neuron be proportional to the integral of the probability density of the stimulus.

Laughlin (1981) has provided evidence that responses of the large monopolar cell (LMC) in the visual system of the fly satisfy the entropy-maximizing condition. The LMC responds to contrast, and Laughlin measured the probability distribution of contrasts of natural scenes in habitats where the flies he studied live. The solid curve in figure 4.2 is the integral of this measured distribution. The data points in figure 4.2 are LMC responses as a function of contrast. These responses are measured as membrane potential fluctuation amplitudes, not as firing rates, but the analysis presented can be applied without modification. As figure 4.2 indicates, the response as a function of contrast is very close to the integrated probability density, suggesting that the LMC is using a maximum entropy encoding.

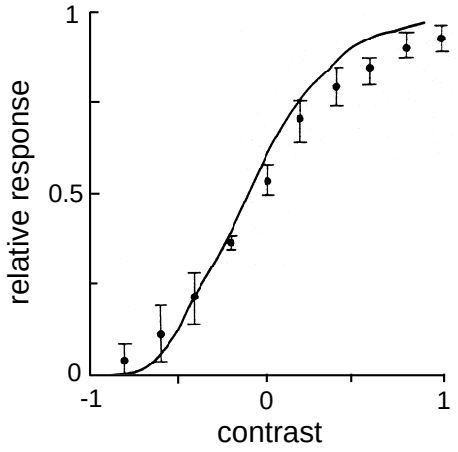

Figure 4.2 Contrast response of the fly LMC (data points) compared to the integral of the natural contrast probability distribution (solid curve). The relative response is the amplitude of the membrane potential fluctuation produced by the onset of a light or dark image with a given level of contrast divided by the maximum response. Contrast is defined relative to the background level of illumination. (Adapted from Laughlin, 1981.)

Even though neurons have maximum firing rates, the constraint $r \leq r_{\max}$ may not always be the factor limiting the entropy. For example, the average firing rate of the neuron may be constrained to values much less than $r_{\max}$, or the variance of the firing rate might be constrained. The reader is invited to show that the entropy-maximizing probability density, if the average firing rate is constrained to a fixed value, is an exponential. A related calculation shows that the probability density that maximizes the entropy subject to constraints on the firing rate and its variance is a Gaussian.

Populations of Neurons

When a population of neurons encodes a stimulus, optimizing their individual response properties will not necessarily lead to an optimized population response. Optimizing individual responses could result in a highly redundant population representation in which different neurons encode the same information. Entropy maximization for a population requires that the neurons convey independent pieces of information (i.e., they must have different response selectivities). Let the vector $\mathbf{r}$ with components r_a for $a = 1, 2, \ldots, N$ denote the firing rates for a population of N neurons, measured with resolution Δr. If $p[\mathbf{r}]$ is the probability of evoking a population response characterized by the vector $\mathbf{r}$, the entropy for the entire population response is

$$H = -\int d\mathbf{r}\, p[\mathbf{r}] \log_2 p[\mathbf{r}] - N \log_2 \Delta r\,. \tag{4.26}$$

Along with the full population entropy of Equation 4.26, we can also con-

sider the entropy associated with individual neurons within the population. If $p[r_a] = \int \prod_{b \neq a} dr_b \, p[\mathbf{r}]$ is the probability density for response r_a from neuron a, its entropy is

$$H_a = -\int dr_a \, p[r_a] \log_2 p[r_a] - \log_2 \Delta r = -\int d\mathbf{r} \, p[\mathbf{r}] \log_2 p[r_a] - \log_2 \Delta r \, . \tag{4.27}$$

The true population entropy can never be greater than the sum of these individual neuron entropies over the entire population,

$$H \leq \sum_a H_a \, . \tag{4.28}$$

To prove this, we note that the difference between the full entropy and the sum of individual neuron entropies is

$$\sum_a H_a - H = \int d\mathbf{r} \, p[\mathbf{r}] \log_2 \left(\frac{p[\mathbf{r}]}{\prod_a p_a[r_a]} \right) \geq 0 \, . \tag{4.29}$$

The inequality follows from the fact that the middle expression is the KL divergence between the probability distributions $p[\mathbf{r}]$ and $\prod_a p[r_a]$, and a KL divergence is always nonnegative. Equality holds only if

$$p[\mathbf{r}] = \prod_a p[r_a] \, , \tag{4.30}$$

that is, if the responses of the neurons are statistically independent. Thus, the full response entropy is never greater than the sum of the entropies of the individual neurons in the population, and it reaches the limiting value when equation 4.30 is satisfied. A code that satisfies this condition is called a factorial code because the probability factorizes into a product of single neuron probabilities. When the population-response probability density factorizes, this implies that the individual neurons respond independently. The entropy difference in equation 4.29 has been suggested as a measure of redundancy.

factorial code

redundancy

Combining this result with the results of the previous section, we conclude that the maximum population-response entropy can be achieved by satisfying two conditions. First, the individual neurons must respond independently, which means that $p[\mathbf{r}] = \prod_a p[r_a]$ must factorize. Second, they must all have response probabilities that are optimal for whatever constraints are imposed (e.g., flat, exponential, or Gaussian). If the same constraint is imposed on every neuron, the second condition implies that every neuron must have the same response probability density. In other words, $p[r_a]$ must be the same for all a values, a property called probability equalization. This does not imply that all the neurons respond identically to every stimulus. Indeed, the conditional probabilities $p[r_a|s]$ must be different for different neurons if they are to act independently. We proceed by considering factorization and probability equalization as general principles of entropy maximization, without imposing explicit constraints.

factorization

probability equalization

Exact factorization and probability equalization are difficult to achieve, especially if the form of the neural response is restricted. These goals are likely to be impossible to achieve, for example, if the neural responses are modeled as having a linear relation to the stimulus. A more modest goal is to require that the lowest-order moments of the population-response probability distribution match those of a fully factorized and equalized distribution. If the individual response probability distributions are equal, the average firing rates and firing rate variances will be the same for all neurons, $\langle r_a \rangle = \langle r \rangle$ and $\langle (r_a - \langle r \rangle)^2 \rangle = \sigma_r^2$ for all a. Furthermore, the covariance matrix for a factorized and probability-equalized population distribution is proportional to the identity matrix,

$$Q_{ab} = \int d\mathbf{r}\, p[\mathbf{r}](r_a - \langle r \rangle)(r_b - \langle r \rangle) = \sigma_r^2 \delta_{ab} \,. \tag{4.31}$$

decorrelation and variance equalization

Finding response distributions that satisfy only the decorrelation and variance equalization condition of equation 4.31 is usually tractable. In the following examples, we restrict ourselves to this easier task. This maximizes the entropy only if the statistics of the responses are Gaussian, but it is a reasonable procedure even in a non-Gaussian case, because it typically reduces the redundancy in the population code and spreads the load of information transmission equally among the neurons.

Application to Retinal Ganglion Cell Receptive Fields

Entropy and information maximization have been used to explain properties of visual receptive fields in the retina, LGN, and primary visual cortex. The basic assumption is that these receptive fields serve to maximize the amount of information that the associated neural responses convey about natural visual scenes in the presence of noise. Information theoretical analyses are sensitive to the statistical properties of the stimuli being represented, so the statistics of natural scenes play an important role in these studies. Natural scenes exhibit substantial spatial and temporal redundancy. Maximizing the information conveyed requires removing this redundancy from the neural responses.

It should be kept in mind that the information maximization approach sets limited goals and requires strong assumptions about the nature of the constraints relevant to the nervous system. In addition, the approach analyzes only the representational properties of neural responses and ignores the computational goals of the visual system, such as object recognition or target tracking. Finally, maximizing other measures of performance, different from the mutual information, may give similar results. Nevertheless, the principle of information maximization is quite successful at accounting for properties of receptive fields early in the visual pathway.

In chapter 2, a visual image was defined by a contrast function $s(x, y, t)$ with a trial-averaged value of 0. For the calculations we present here, it is more convenient to express the x and y coordinates for locations on the

viewing screen in terms of a single vector $\vec{x} = (x, y)$, or sometimes $\vec{y} = (x, y)$. Using this notation, the linear estimate of the response of a visual neuron discussed in chapter 2 can be written as

$$L(t) = \int_0^\infty d\tau \int d\vec{x}\, D(\vec{x}, \tau) s(\vec{x}, t - \tau)\,. \tag{4.32}$$

If the space-time receptive field $D(\vec{x}, \tau)$ is separable, $D(\vec{x}, \tau) = D_s(\vec{x}) D_t(\tau)$, and we can rewrite $L(t)$ as the product of integrals involving temporal and spatial filters. To keep the notation simple, we assume that the stimulus can also be separated, so that $s(\vec{x}, t) = s_s(\vec{x}) s_t(t)$. Then, $L(t) = L_s L_t(t)$ where

$$L_s = \int d\vec{x}\, D_s(\vec{x}) s_s(\vec{x}) \tag{4.33}$$

and

$$L_t(t) = \int_0^\infty d\tau\, D_t(\tau) s_t(t - \tau)\,. \tag{4.34}$$

In the following, we analyze the spatial and temporal components, D_s and D_t, separately by considering the information-carrying capacity of L_s and L_t. We study the spatial receptive fields of retinal ganglion cells in this section, and the temporal response properties of LGN cells in the next. Later, we discuss the application of information maximization ideas to primary visual cortex.

To derive appropriately optimal spatial filters, we consider an array of retinal ganglion cells with receptive fields covering a small patch of the retina. We assume that the statistics of the input are spatially (and temporally) stationary or translation-invariant. This means that all locations and directions in space (and all times), at least within the patch we consider, are equivalent. This equivalence allows us to give all of the receptive fields the same spatial structure, with the receptive fields of different cells merely being shifted to different points within the visual field. As a result, we write the spatial kernel describing a retinal ganglion cell with receptive field centered at the point $\vec{a}$ as $D_s(\vec{x} - \vec{a})$. The linear response of this cell is then

$$L_s(\vec{a}) = \int d\vec{x}\, D_s(\vec{x} - \vec{a}) s_s(\vec{x})\,. \tag{4.35}$$

Note that we are labeling the neurons by the locations $\vec{a}$ of the centers of their receptive fields rather than by an integer index such as i. This is a convenient labeling scheme that allows sums over neurons to be replaced by sums over parameters describing their receptive fields. The vectors $\vec{a}$ for the different neurons take on discrete values corresponding to the different neurons in the population. If many neurons are being considered, these discrete vectors may fill the range of receptive field locations quite densely. In this case, it is reasonable to approximate the large but discrete

set of $\vec{a}$ values with a vector $\vec{a}$ that is allowed to vary continuously. In other words, as an approximation, we proceed as if there were a neuron corresponding to every continuous value of $\vec{a}$. This allows us to treat $L(\vec{a})$ as a function of $\vec{a}$ and to replace sums over neurons with integrals over $\vec{a}$. In the case we are considering, the receptive fields of retinal ganglion cells cover the retina densely, with many receptive fields overlapping each point on the retina, so the replacement of discrete sums over neurons with continuous integrals over $\vec{a}$ is quite accurate.

The Whitening Filter

We will not attempt a complete entropy maximization for the case of retinal ganglion cells. Instead, we follow the approximate procedure of setting the correlation matrix between different neurons within the population proportional to the identity matrix (equation 4.31). The relevant correlation is the average, over all stimuli, of the product of the linear responses of two cells, with receptive fields centered at $\vec{a}$ and $\vec{b}$,

$$Q_{\rm LL}(\vec{a},\vec{b}) = \langle L_{\rm s}(\vec{a})L_{\rm s}(\vec{b})\rangle = \int d\vec{x}\, d\vec{y}\, D_{\rm s}(\vec{x}-\vec{a})D_{\rm s}(\vec{y}-\vec{b})\langle s_{\rm s}(\vec{x})s_{\rm s}(\vec{y})\rangle\,. \tag{4.36}$$

The average here, denoted by angle brackets, is not over trials but over the set of natural scenes for which we believe the receptive field is optimized. By analogy with equation 4.31, decorrelation and variance equalization of the different retinal ganglion cells, when $\vec{a}$ and $\vec{b}$ are taken to be continuous variables, require that we set this correlation function proportional to a δ function,

$$Q_{\rm LL}(\vec{a},\vec{b}) = \sigma_{\rm L}^2\delta(\vec{a}-\vec{b})\,. \tag{4.37}$$

This is the continuous variable analog of making a discrete correlation matrix proportional to the identity matrix (equation 4.31). The δ function with vector arguments is nonzero only when all of the components of $\vec{a}$ and $\vec{b}$ are identical.

The quantity $\langle s_{\rm s}(\vec{x})s_{\rm s}(\vec{y})\rangle$ in equation 4.36 is the correlation function of the stimulus averaged over natural scenes. Our assumption of homogeneity implies that this quantity is only a function of the vector difference $\vec{x}-\vec{y}$ (actually, if all directions are equivalent, it is only a function of the magnitude $|\vec{x}-\vec{y}|$), and we write it as

$$Q_{ss}(\vec{x}-\vec{y}) = \langle s_{\rm s}(\vec{x})s_{\rm s}(\vec{y})\rangle\,. \tag{4.38}$$

To determine the form of the receptive field filter that is optimal, we must solve equation 4.37 for D_s. This is done by expressing $D_{\rm s}$ and Q_{ss} in terms of their Fourier transforms $\tilde{D}_{\rm s}$ and $\tilde{Q}_{ss}$,

$$D_{\rm s}(\vec{x}-\vec{a}) = \frac{1}{4\pi^2}\int d\vec{\kappa}\,\exp\left(-i\vec{\kappa}\cdot(\vec{x}-\vec{a})\right)\tilde{D}_{\rm s}(\vec{\kappa}) \tag{4.39}$$

$$Q_{ss}(\vec{x}-\vec{y}) = \frac{1}{4\pi^2}\int d\vec{\kappa}\,\exp\left(-i\vec{\kappa}\cdot(\vec{x}-\vec{y})\right)\tilde{Q}_{ss}(\vec{\kappa})\,. \tag{4.40}$$

$\tilde{Q}_{ss}$, which is real and nonnegative, is also called the stimulus power spectrum (see chapter 1). In terms of these Fourier transforms, equation 4.37 becomes

$$|\tilde{D}_s(\vec{\kappa})|^2 \tilde{Q}_{ss}(\vec{\kappa}) = \sigma_L^2 \,, \tag{4.41}$$

from which we find

$$|\tilde{D}_s(\vec{\kappa})| = \frac{\sigma_L}{\sqrt{\tilde{Q}_{ss}(\vec{\kappa})}} \,. \tag{4.42}$$

whitening filter

The linear kernel described by equation 4.42 exactly compensates for whatever dependence the Fourier transform of the stimulus correlation function has on the spatial frequency $\vec{\kappa}$, making the product $\tilde{Q}_{ss}(\vec{\kappa})|\tilde{D}_s(\vec{\kappa})|^2$ independent of $\vec{\kappa}$. This product is the power spectrum of L. The output of the optimal filter has a power spectrum that is independent of spatial frequency, and therefore has the same characteristics as white noise. Therefore, the kernel in equation 4.42 is called a whitening filter. Different spatial frequencies act independently in a linear system, so decorrelation and variance equalization require them to be utilized at equal signal strength.

The calculation we have performed determines only the amplitude $|\tilde{D}_s(\vec{\kappa})|$, and not $\tilde{D}_s(\vec{\kappa})$ itself. Thus, decorrelation and variance equalization do not uniquely specify the form of the linear kernel. We study some consequences of the freedom to choose different linear kernels satisfying equation 4.42 later in the chapter.

The spatial correlation function for natural scenes has been measured, with the result that $\tilde{Q}_{ss}(\vec{\kappa})$ is proportional to $1/|\vec{\kappa}|^2$ over the range it has been evaluated. The behavior near $\vec{\kappa} = 0$ is not well established, but the divergence of $1/|\vec{\kappa}|^2$ near $\vec{\kappa} = 0$ can be removed by setting $\tilde{Q}_{ss}(\vec{\kappa})$ proportional to $1/(|\vec{\kappa}|^2 + \kappa_0^2)$ where κ_0 is a constant. The stimuli of interest in the calculation of retinal ganglion receptive fields are natural images as they appear on the retina, not in the photographs from which the natural scenes statistics are measured. An additional factor must be included in $\tilde{Q}_{ss}(\vec{\kappa})$ to account for filtering introduced by the optics of the eye (the optical modulation transfer function). A simple model of the optical modulation transfer function results in an exponential correction to the stimulus correlation function,

optical modulation transfer function

$$\tilde{Q}_{ss}(\vec{\kappa}) \propto \frac{\exp(-\alpha|\vec{\kappa}|)}{|\vec{\kappa}|^2 + \kappa_0^2} \,, \tag{4.43}$$

with α a parameter. Substituting this into equation 4.42 gives the rather peculiar result that the amplitude $|\tilde{D}_s(\vec{\kappa})|$, being proportional to the inverse of the square root of $\tilde{Q}_{ss}$, is predicted to grow exponentially for large $|\vec{\kappa}|$. Whitening filters maximize entropy by equalizing the distribution of response power over the entire spatial frequency range. High spatial frequency components of images are relatively rare in natural scenes and, even if they occur, are greatly attenuated by the eye. The whitening filter compensates for this by boosting the responses to high spatial frequencies. Although this is the result of the entropy maximization calculation, it is not

a good strategy to use in an unrestricted way for visual processing. Real inputs to retinal ganglion cells involve a mixture of true signal and noise coming from biophysical sources in the retina. At high spatial frequencies, for which the true signal is weak, inputs to retinal ganglion cells are likely to be dominated by noise, especially in low-light conditions. Boosting the amplitude of this noise-dominated input and transmitting it to the brain is not an efficient visual encoding strategy.

The problem of excessive boosting of responses at high spatial frequency arises in the entropy maximization calculation because no distinction has been made between the entropy coming from true signals and that coming from noise. To correct this problem, we should maximize the information transmitted by the retinal ganglion cells about natural scenes, rather than maximize the entropy. A full information-maximization calculation of the receptive field properties of retinal ganglion cells can be performed, but this requires introducing a number of assumptions about the constraints that are relevant, and it is not entirely obvious what these constraints should be. Instead, we will follow an approximate procedure that prefilters the input to eliminate as much noise as possible, and then uses the results of this section to maximize the entropy of a linear filter acting on the prefiltered input signal.

Filtering Input Noise

Suppose that the visual stimulus on the retina is the sum of the true stimulus $s_s(\vec{x})$ that should be conveyed to the brain and a noise term $\eta(\vec{x})$ that reflects image distortion, photoreceptor noise, and other signals that are not worth conveying beyond the retina. To deal with such a mixed input signal, we express the Fourier transform of the linear kernel $\tilde{D}_s(\vec{\kappa})$ as a product of two terms: a noise filter, $\tilde{D}_\eta(\vec{\kappa})$, that eliminates as much of the noise as possible; and a whitening filter, $\tilde{D}_w(\vec{\kappa})$, that satisfies equation 4.42. The Fourier transform of the complete filter is then $\tilde{D}_s(\vec{\kappa}) = \tilde{D}_w(\vec{\kappa})\tilde{D}_\eta(\vec{\kappa})$.

To determine the form of the noise filter, we demand that when it is applied to the total input $s_s(\vec{x}) + \eta(\vec{x})$, the result is as close to the signal part of the input, $s_s(\vec{x})$, as possible. The problem of minimizing the average squared difference between the filtered noisy signal and the true signal is formally the same as the problems we solved in chapter 2 (appendix A) and chapter 3 (appendix C) to determine the optimal kernels for rate prediction and for spike decoding (also see the Mathematical Appendix). The general solution is that the Fourier transform of the optimal filter is the Fourier transform of the cross-correlation between the quantity being filtered and the quantity being approximated divided by the Fourier transform of the autocorrelation of the quantity being filtered. In the present example, there is a slight complication that the integral in equation 4.35 is not in the form of a convolution, because D_s is written as a function of $\vec{x} - \vec{a}$ rather than $\vec{a} - \vec{x}$. However, in the case we consider, this ultimately makes no difference to the final answer.

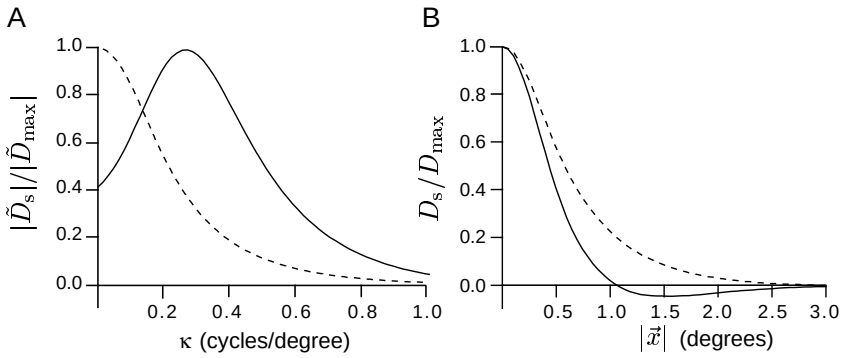

Figure 4.3 Receptive field properties predicted by entropy maximization and noise suppression of responses to natural images. (A) The amplitude of the predicted Fourier-transformed linear filters for low (solid curve) and high (dashed curve) input noise. $|\tilde{D}_s(\vec{\kappa})|$ is plotted relative to its maximum value. (B) The linear kernel as a function of the distance from the center of the receptive field for low (solid curve) and high (dashed curve) input noise. Observe the center-surround structure at low noise. $\tilde{D}_s(\vec{\kappa})$ is taken to be real, and $D_s(|\vec{x}|)$ is plotted relative to its maximum value. Parameter values used were $1/\alpha = 0.16$ cycles/degree, $k_0 = 0.16$ cycles/degree, and $\tilde{Q}_{\eta\eta}/\tilde{Q}_{ss}(0) = 0.05$ for the low-noise case and 1 for the high-noise case.

The calculation simplifies because we assume that the signal and noise terms are uncorrelated, so that $\langle s_s(\vec{x})\eta(\vec{y})\rangle = 0$. Then, the relevant cross-correlation for this problem is

$$\langle (s_s(\vec{x}) + \eta(\vec{x}))s_s(\vec{y})\rangle = Q_{ss}(\vec{x} - \vec{y})\,, \tag{4.44}$$

and the autocorrelation is

$$\langle (s_s(\vec{x}) + \eta(\vec{x}))(s_s(\vec{y}) + \eta(\vec{y}))\rangle = Q_{ss}(\vec{x} - \vec{y}) + Q_{\eta\eta}(\vec{x} - \vec{y})\,, \tag{4.45}$$

where Q_{ss} and $Q_{\eta\eta}$ are, respectively, the stimulus and noise autocorrelations functions. These results imply that the optimal noise filter is real and given, in terms of the Fourier transforms of Q_{ss} and $Q_{\eta\eta}$, by

noise filter

$$\tilde{D}_\eta(\vec{\kappa}) = \frac{\tilde{Q}_{ss}(\vec{\kappa})}{\tilde{Q}_{ss}(\vec{\kappa}) + \tilde{Q}_{\eta\eta}(\vec{\kappa})}\,. \tag{4.46}$$

Because the noise filter is designed so that its output matches the signal as closely as possible, we make the approximation of using the same whitening filter as before (equation 4.42). Combining the two, we find that

$$|\tilde{D}_s(\vec{\kappa})| \propto \frac{\sigma_L\sqrt{\tilde{Q}_{ss}(\vec{\kappa})}}{\tilde{Q}_{ss}(\vec{\kappa}) + \tilde{Q}_{\eta\eta}(\vec{\kappa})}\,. \tag{4.47}$$

Linear kernels resulting from equation 4.47, using equation 4.43 for the stimulus correlation function, are plotted in figure 4.3. For this figure, we have assumed that the input noise is white so that $\tilde{Q}_{\eta\eta}$ is independent of $\vec{\kappa}$. Both the amplitude of the Fourier transform of the kernel (figure 4.3A) and

the actual spatial kernel $D_s(\vec{x})$ (figure 4.3B) are plotted under conditions of low and high noise. The linear kernels in figure 4.3B have been constructed by assuming that $\tilde{D}_s(\vec{\kappa})$ satisfies equation 4.47 and is real, which minimizes the spatial extent of the resulting receptive field. The resulting function $D_s(\vec{x})$ is radially symmetric, so it depends only on the distance $|\vec{x}|$ from the center of the receptive field to the point $\vec{x}$, and this radial dependence is plotted in figure 4.3B. Under low noise conditions (solid lines in figure 4.3), the linear kernel has a bandpass character and the predicted receptive field has a center-surround structure, which matches the retinal ganglion receptive fields shown in chapter 2. This structure eliminates one major source of redundancy in natural scenes: the strong similarity of neighboring inputs owing to the predominance of low spatial frequencies in images.

When the noise level is high (dashed lines in figure 4.3), the structure of the optimal receptive field is different. In spatial frequency terms, the filter is now low-pass, and the receptive field loses its surround. This structure averages over neighboring pixels to extract the true signal obscured by the uncorrelated noise. In the retina, we expect the signal-to-noise ratio to be controlled by the level of ambient light, with low levels of illumination corresponding to the high-noise case. The predicted change in the receptive fields at low illumination (high noise) matches what actually happens in the retina. At low light levels, circuitry changes within the retina remove the opposing surrounds from retinal ganglion cell receptive fields.

Temporal Processing in the LGN

Natural images tend to change relatively slowly over time. This means that there is substantial redundancy in the succession of natural images, suggesting an opportunity for efficient temporal filtering to complement efficient spatial filtering. An analysis similar to that of the previous section can be performed to account for the temporal receptive fields of visually responsive neurons early in the visual pathway. Recall that the predicted linear temporal response is given by $L_t(t)$, as expressed in equation 4.34. The analog of equation 4.37 for temporal decorrelation and variance equalization is

$$\langle L_t(t) L_t(t') \rangle = \sigma_L^2 \delta(t - t') \,. \tag{4.48}$$

This is mathematically identical to equation 4.37 except that the role of the spatial variables $\vec{a}$ and $\vec{b}$ has been replaced by the temporal variables t and t'. The analysis proceeds exactly as above, and the optimal filter is the product of a noise filter and a temporal whitening filter, as before. The temporal linear kernel $D_t(\tau)$ is written in terms of its Fourier transform,

$$D_t(\tau) = \frac{1}{2\pi} \int d\omega \, \exp(-i\omega\tau) \tilde{D}_t(\omega) \,, \tag{4.49}$$

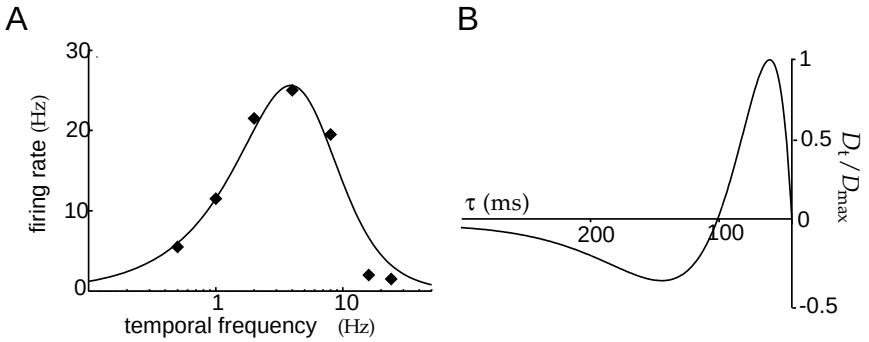

Figure 4.4 (A) Predicted (curve) and actual (diamonds) selectivity of an LGN cell as a function of temporal frequency. The predicted curve is based on the optimal linear filter $\tilde{D}_t(\omega)$ with $\omega_0 = 5.5$ Hz. (B) Causal, minimum phase, temporal form of the optimal filter. (Adapted from Dong and Atick, 1995; data in A from Saul and Humphrey, 1990.)

and $\tilde{D}_t(\omega)$ is given by an equation similar to 4.47,

$$|\tilde{D}_t(\omega)| \propto \frac{\sigma_L \sqrt{\tilde{Q}_{ss}(\omega)}}{\tilde{Q}_{ss}(\omega) + \tilde{Q}_{\eta\eta}(\omega)} . \tag{4.50}$$

In this case, $\tilde{Q}_{ss}(\omega)$ and $\tilde{Q}_{\eta\eta}(\omega)$ are the power spectra of the signal and the noise in the temporal domain.

Dong and Atick (1995) analyzed temporal receptive fields in the LGN in this way, under the assumption that a substantial fraction of the temporal redundancy of visual stimuli is removed in the LGN rather than in the retina. They determined that the temporal power spectrum of natural scenes has the form

$$\tilde{Q}_{ss}(\omega) \propto \frac{1}{\omega^2 + \omega_0^2} , \tag{4.51}$$

where ω_0 is a constant. The resulting filter, in both the temporal frequency and the time domains, is plotted in figure 4.4. Figure 4.4A shows the predicted and actual frequency responses of an LGN cell. This is similar to the plot in figure 4.3A, except that the result has been normalized to a realistic response level so that it can be compared with data. Because the optimization procedure determines only the amplitude of the Fourier transform of the linear kernel, $D_t(\tau)$ is not uniquely specified. To determine the temporal kernel, we require it to be causal ($D_t(\tau) = 0$ for $\tau < 0$) and impose a technical condition known as minimum phase, which assures that the output changes as rapidly as possible when the stimulus varies. Figure 4.4B shows the resulting form of the temporal filter. The space-time receptive fields shown in chapter 2 tend to change sign as a function of τ. The temporal filter in figure 4.4B has exactly this property.

An interesting test of the notion of optimal coding was carried out by Dan, Atick, and Reid (1996). They used both natural scene and white-noise stimuli while recording cat LGN cells. Figure 4.5A shows the power

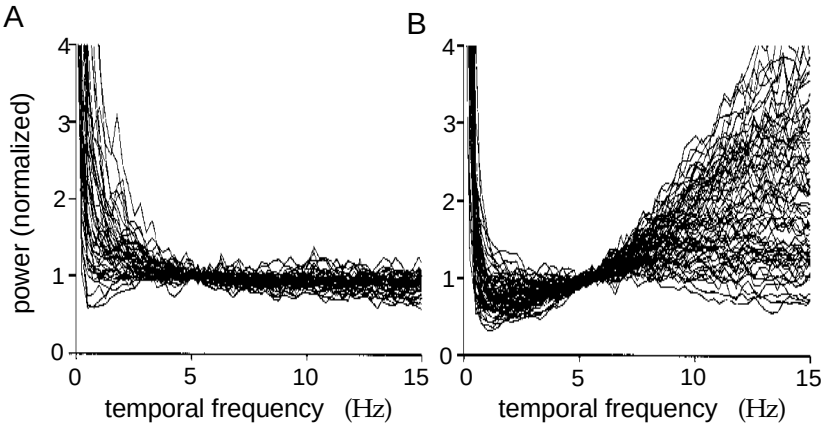

Figure 4.5 (A) Power spectra of the spike trains of 51 cat LGN cells in response to presentation of the movie *Casablanca*, normalized to their own values between 5 and 6 Hz. B) Equivalently normalized power spectra of the spike trains of 75 LGN cells in response to white-noise stimuli. (Adapted from Dan et al., 1996.)

spectra of spike trains of cat LGN cells in response to natural scenes (the movie *Casablanca*), and figure 4.5B shows power spectra in response to white-noise stimuli. The power spectra of the responses to natural scenes are quite flat above about $\omega = 3$ Hz. In response to white noise, on the other hand, they rise with ω. This is exactly what we would expect if LGN cells are acting as temporal whitening filters. In the case of natural stimuli, the whitening filter evenly distributes the output power over a broad frequency range. Responses to white-noise stimuli increase at high frequencies due to the boosting of inputs at these frequencies by the whitening filter.

Cortical Coding

Computational concerns beyond mere linear information transfer are likely to be relevant at the level of cortical processing of visual images. For one thing, the primary visual cortex has many more neurons than the LGN, yet they can collectively convey no more information about the visual world than they receive. As we saw in chapter 2, neurons in primary visual cortex are selective for quantities, such as spatial frequency and orientation, that are of particular importance in relation to object recognition but not for information transfer. Nevertheless, the methods described in the previous section can be used to understand restricted aspects of receptive fields of neurons in primary visual cortex, namely, the way that their multiple selectivities are collectively assigned. For example, cells that respond best at high spatial frequencies tend to respond more to low temporal frequency components of images, and vice versa.

The stimulus power spectrum written as a function of both spatial and temporal frequency has been estimated as $\tilde{Q}_{ss}(\vec{\kappa}, \omega) \propto 1/\left(|\vec{\kappa}|^2 + \alpha^2\omega^2\right)$,

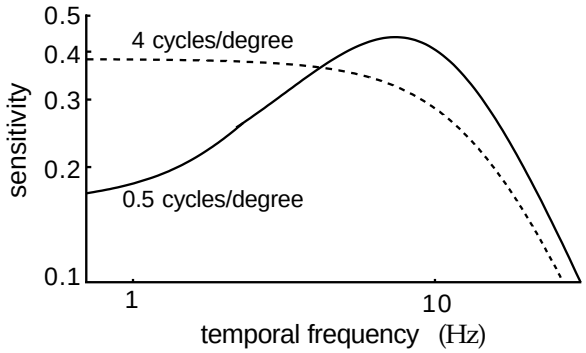

Figure 4.6 Dependence of temporal frequency tuning on preferred spatial frequency for space-time receptive fields derived from information maximization in the presence of noise. The curves show a transition from partial whitening in temporal frequency for low preferred spatial frequency (solid curve, 0.5 cycles/degree) to temporal summation for high preferred spatial frequency (dashed curve, 4 cycles/degree). (Adapted from Li, 1996.)

where $\alpha = 0.4$ cycle seconds/degree. This correlation function decreases both for high spatial and high temporal frequencies. Figure 4.6 shows how temporal selectivity for a combined noise and whitening filter, constructed using this stimulus power spectrum, changes for different preferred spatial frequencies. The basic idea is that components with fairly low stimulus power are boosted by the whitening filter, while those with very low stimulus power get suppressed by the noise filter. As shown by Li (1996), if a cell is selective for high spatial frequencies, the input signal rapidly falls below the noise (treated as white) as the temporal frequency of the input is increased. As a result, the noise filter of equation 4.46 causes the temporal response to be largest at 0 temporal frequency (dashed curve of figure 4.6). If instead the cell is selective for low spatial frequencies, the signal dominates the noise up to higher temporal frequencies, and the whitening filter causes the response to increase as a function of temporal frequency up to a maximum value where the noise filter begins to suppress the response (solid curve in figure 4.6). Receptive fields with preference for high spatial frequency thus act as low-pass temporal filters, and receptive fields with selectivity for low spatial frequency act as bandpass temporal filters.

Similar conclusions can be drawn concerning other joint selectivities. For example, color-selective (chrominance) cells tend to be selective for low temporal frequencies, because their input signal-to-noise ratio is lower than that for broadband (luminance) cells. There is also an interesting predicted relationship between ocular dominance and spatial frequency tuning due to the nature of the correlations between the two eyes. Optimal receptive fields with low spatial frequency tuning (for which the input signal-to-noise ratio is high) have enhanced sensitivity to differences between inputs coming from the two eyes. Receptive fields tuned to intermediate and high spatial frequencies suppress ocular differences.

4.3 Entropy and Information for Spike Trains

Computing the entropy or information content of a neuronal response characterized by spike times is much more difficult than computing these quantities for responses described by firing rates. Nevertheless, these computations are important, because firing rates are incomplete descriptions that can lead to serious underestimates of the entropy and information. In this section, we discuss how the entropy and mutual information can be computed for spike trains. Extensive further discussion can be found in the book by Rieke et al. (1997).

Spike-train entropy calculations are typically based on the study of long-duration recordings consisting of many action potentials. The entropy or mutual information typically grows linearly with the length of the spike train being considered. For this reason, the entropy and mutual information of spike trains are reported as entropy or information rates. These are the total entropy or information divided by the duration of the spike train. We write the entropy rate as $\dot{H}$ rather than H. Alternatively, entropy and mutual information can be divided by the total number of action potentials and reported as bits per spike rather than bits per second.

entropy and information rates

To compute entropy and information rates for a spike train, we need to determine the probabilities that various temporal patterns of action potentials appear. These probabilities replace the factors $P[r]$ or $p[r]$ that occur when discrete or continuous firing rates are used to characterize a neural response. The temporal pattern of a group of action potentials can be specified by listing either the individual spike times or the sequence of intervals between successive spikes. The entropy and mutual information calculations we present are based on a spike-time description, but as an initial example we consider an approximate computation of entropy using interspike intervals.

The probability of an interspike interval falling in the range between τ and $\tau + \Delta\tau$ is given in terms of the interspike interval probability density by $p[\tau]\Delta\tau$. Because the interspike interval is a continuous variable, we must specify a resolution $\Delta\tau$ with which it is measured to define the entropy. If the different interspike intervals are statistically independent, the entropy associated with the interspike intervals in a spike train of average rate $\langle r \rangle$ and of duration T is the number of intervals, $\langle r \rangle T$, times the integral over τ of $-p[\tau]\log_2(p[\tau]\Delta\tau)$. The entropy rate is obtained by dividing this result by T, and the entropy per spike requires dividing by the number of spikes, $\langle r \rangle T$. The assumption of independent interspike intervals is critical for obtaining the spike-train entropy solely in terms of $p[\tau]$. Correlations between different interspike intervals reduce the total entropy, so the result obtained by assuming independent intervals provides an upper bound on the true entropy of a spike train. Thus, in general, the entropy rate $\dot{H}$ for a spike train with interspike interval distribution $p[\tau]$ and average rate $\langle r \rangle$

satisfies

$$\dot{H} \leq -\langle r \rangle \int_0^\infty d\tau \, p[\tau] \log_2 (p[\tau] \Delta\tau) \, . \tag{4.52}$$

Poisson entropy rate

If a spike train is described by a homogeneous Poisson process with rate $\langle r \rangle$, we have $p[\tau] = \langle r \rangle \exp(-\langle r \rangle \tau)$, and the interspikes are statistically independent (chapter 1). Equation 4.52 is then an equality and, performing the integrals,

$$\dot{H} = \frac{\langle r \rangle}{\ln(2)} \left(1 - \ln(\langle r \rangle \Delta\tau)\right) \, . \tag{4.53}$$

We now turn to a more general calculation of the spike-train entropy. To make entropy calculations practical, a long spike train is broken into statistically independent subunits, and the total entropy is written as the sum of the entropies for the individual subunits. In the case of equation 4.52, the subunit was the interspike interval. If interspike intervals are not independent, and we wish to compute a result and not merely a bound, we must work with larger subunit descriptions. Strong et al. (1998) proposed a scheme that uses spike sequences of duration T_s as these basic subunits. Note that the variable T_s is used here to denote the duration of the spike sequence being considered, while T, which is much larger than T_s, is the duration of the entire spike train.

The time that a spike occurs is a continuous variable, so, as in the case of interspike intervals, a resolution must be specified when spike train entropies are computed. This can be done by dividing time into discrete bins of size Δt. We assume that the bins are small enough so that not more than one spike appears in a bin. Depending on whether or not a spike occurred within it, each bin is labeled by a 0 (no spike) or a 1 (spike). A spike sequence defined over a block of duration T_s is thus represented by a string of $T_s/\Delta t$ zeros and ones. We denote such a sequence by $B(t)$, where B is a $T_s/\Delta t$ bit binary number, and t specifies the time of the first bin in the sequence being considered. Both T_s and t are integer multiples of the bin size Δt.

The probability of a sequence B occurring at any time during the entire response is denoted by $P[B]$. This can be obtained by counting the number of times the sequence B occurs anywhere within the spike trains being analyzed (including overlapping cases). The spike-train entropy rate implied by this distribution is

$$\dot{H} = -\frac{1}{T_s} \sum_B P[B] \log_2 P[B] \, , \tag{4.54}$$

where the sum is over all the sequences B found in the data set, and we have divided by the duration T_s of a single sequence to obtain an entropy rate.

If the spike sequences in nonoverlapping intervals of duration T_s are independent, the full spike-train entropy rate is also given by equation 4.54.

However, any correlations between successive intervals (if $B(t + T_s)$ is correlated with $B(t)$, for example) reduce the total spike-train entropy, causing equation 4.54 to overestimate the true entropy rate. Thus, for finite T_s, this equation provides an upper bound on the true entropy rate. If T_s is too small, $B(t + T_s)$ and $B(t)$ are likely to be correlated, and the overestimate may be severe. As T_s increases, we expect the correlations to get smaller, and equation 4.54 should provide a more accurate value. For any finite data set, T_s cannot be increased past a certain point, because there will not be enough spike sequences of duration T_s in the data set to determine their probabilities. Thus, in practice, T_s must be increased until the point where the extraction of probabilities becomes problematic, and some form of extrapolation to $T_s \to \infty$ must be made.

Statistical mechanics arguments suggest that the difference between the entropy for finite T_s and the true entropy for $T_s \to \infty$ should be proportional to $1/T_s$ for large T_s. Therefore, the true entropy can be estimated, as in figure 4.7, by linearly extrapolating a plot of the entropy rate versus $1/T_s$ to the point $1/T_s = 0$. In figure 4.7 (upper line), this has been done for data from the motion-sensitive H1 neuron of the fly visual system. The plotted points show entropy rates computed for different values of $1/T_s$, and they vary linearly over most of the range of the plot. However, when $1/T_s$ goes below about 20/s (or $T_s > 50$ ms), the dependence suddenly changes. This is the point at which the amount of data is insufficient to extract even an overestimate of the entropy. By linearly extrapolating the linear part of the series of computed points in figure 4.7, Strong et al. estimated that the H1 spike trains had an entropy rate of 157 bits/s when the resolution was $\Delta t = 3$ ms.

To compute the mutual information rate for a spike train, we must subtract the noise entropy rate from the full spike-train entropy rate. The noise entropy rate is determined from the probabilities of finding various sequences B, given that they were evoked by the same stimulus. This is done by considering sequences $B(t)$ that start at a fixed time t. If the same stimulus is used in repeated trials, sequences that begin at time t in every trial are generated by the same stimulus. Therefore, the conditional probability of the response, given the stimulus, is in this case the distribution $P[B(t)]$ for response sequences beginning at time t. This is obtained by determining the fraction of trials on which $B(t)$ was evoked. Note that $P[B(t)]$ is the probability of finding a given sequence at time t within a set of spike trains obtained on trials using the same stimulus. In contrast, $P[B]$, used in the spike-train entropy calculation, is the probability of finding the sequence B at any time within these trains. Determining $P[B(t)]$ for a sufficient number of spike sequences may take a large number of trials using the same stimulus.

The full noise entropy is computed by averaging the noise entropy at time t over all t values. The average over t plays the role of the average over

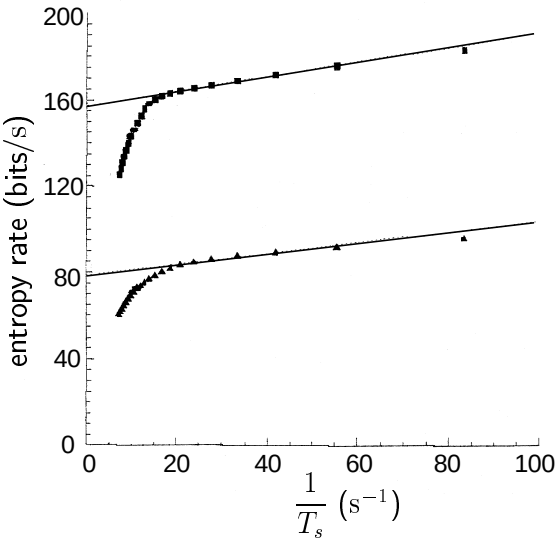

Figure 4.7 Entropy and noise entropy rates for the H1 visual neuron in the fly responding to a randomly moving visual image. The filled circles in the upper trace show the full spike-train entropy rate computed for different values of $1/T_s$. The straight line is a linear extrapolation to $1/T_s = 0$, which corresponds to $T_s \to \infty$. The lower trace shows the spike train noise entropy rate for different values of $1/T_s$. The straight line is again an extrapolation to $1/T_s = 0$. Both entropy rates increase as functions of $1/T_s$, and the true spike-train and noise entropy rates are overestimated at large values of $1/T_s$. At $1/T_s \approx 20/s$, there is a sudden shift in the dependence. This occurs when there is insufficient data to compute the spike sequence probabilities. The difference between the y intercepts of the two straight lines plotted is the mutual information rate. The resolution is $\Delta t = 3$ ms. (Adapted from Strong et al., 1998.)

stimuli in equation 4.6. The result is

$$\dot{H}_{\text{noise}} = -\frac{\Delta t}{T} \sum_t \left(\frac{1}{T_s} \sum_B P[B(t)] \log_2 P[B(t)] \right), \qquad (4.55)$$

where $T/\Delta t$ is the number of different t values being summed.

If equation 4.55 is based on finite-length spike sequences, it provides an upper bound on the noise entropy rate. The true noise entropy rate is estimated by performing a linear extrapolation in $1/T_s$ to $1/T_s = 0$, as was done for the spike-train entropy rate. The result, shown in figure 4.7, is a noise entropy of 79 bits/s for $\Delta t = 3$ ms. The information rate is obtained by taking the difference between the extrapolated values for the spike-train and noise entropy rates. The result for the fly H1 neuron used in figure 4.7 is an information rate of 157 - 79 = 78 bits/s or 1.8 bits/spike. Values in the range 1 to 3 bits/spike are typical results of such calculations for a variety of preparations.

Both the spike-train and noise entropy rates depend on Δt. The leading dependence, coming from the $\log_2 \Delta t$ term discussed previously, cancels

in the computation of the information rate, but the information can still depend on Δt through nondivergent terms. This reflects the fact that more information can be extracted from accurately measured spike times than from poorly measured spike times. Thus, we expect the information rate to increase with decreasing Δt, at least over some range of Δt values. At some critical value of Δt that matches the natural degree of noise jitter in the spike timings, we expect the information rate to stop increasing. This value of Δt is interesting because it tells us about the degree of spike timing accuracy in neural encoding.

The information conveyed by spike trains can be used to compare responses to different stimuli and thereby reveal stimulus-specific aspects of neural encoding. For example, Rieke et al. (1995) compared the information conveyed by single neurons in a peripheral auditory organ (the amphibian papilla) of the bullfrog in response to broadband noise or to noise filtered to have an amplitude spectrum close to that of natural bullfrog calls (although the phases for each frequency component were chosen randomly). They determined that the cells conveyed on average of 46 bits per second (1.4 bits per spike) for broadband noise and 133 bits per second (7.8 bits per spike) for stimuli with call-like spectra, despite the fact that the broadband noise had a higher entropy. The spike trains in response to the call-like stimuli conveyed information with near maximal efficiency.

4.4 Chapter Summary

Shannon's information theory can be used to determine how much a neural response tells both us and, presumably, the animal in which the neuron lives, about a stimulus. Entropy is a measure of the uncertainty or surprise associated with a stochastic variable, such as a stimulus. Mutual information quantifies the reduction in uncertainty associated with the observation of another variable, such as a response. The mutual information is related to the Kullback-Leibler divergence between two probability distributions. We defined the response and noise entropies for probability distributions of discrete and continuous firing rates, and considered how the information transmitted about a set of stimuli might be optimized. The principles of entropy and information maximization were used to account for features of the receptive fields of cells in the retina, LGN, and primary visual cortex. This analysis introduced probability factorization and equalization, and whitening and noise filters. Finally, we discussed how the information conveyed about dynamic stimuli by spike sequences can be estimated.

4.5 Appendix

Positivity of the Kullback-Leibler Divergence

Jensen's inequality

The logarithm is a concave function, which means that $\log_2\langle z\rangle \geq \langle \log_2 z\rangle$, where the angle brackets denote averaging with respect to some probability distribution and z is any positive quantity. The equality holds only if z is a constant. If we consider this relation, known as Jensen's inequality, with $z = Q[r]/P[r]$ and the average defined over the probability distribution $P[r]$, we find

$$-D_{\mathrm{KL}}(P,Q) = \sum_r P[r]\log_2\left(\frac{Q[r]}{P[r]}\right) \leq \log_2\left(\sum_r P[r]\frac{Q[r]}{P[r]}\right) = 0\,. \tag{4.56}$$

The last equality holds because $Q[r]$ is a probability distribution and thus satisfies $\sum_r Q[r] = 1$. Equation 4.56 implies that $D_{\mathrm{KL}}(P,Q) \geq 0$, with equality holding if and only if $P[r] = Q[r]$. A similar result holds for the Kullback-Leibler divergence between two probability densities,

$$D_{\mathrm{KL}}(p,q) = \int dr\, p[r]\log_2\left(\frac{p[r]}{q[r]}\right) \geq 0\,. \tag{4.57}$$

4.6 Annotated Bibliography

Information theory was created by Shannon (see **Shannon & Weaver, 1949**) largely as a way of understanding communication in the face of noise. **Cover & Thomas (1991)** provides a review, and **Rieke et al. (1997)** gives a treatment specialized to neural coding. Information theory and theories inspired by it, such as histogram equalization, were adopted in neuroscience and psychology as a way of understanding sensory transduction and coding, as discussed by **Barlow (1961)** and **Uttley (1979)**. We followed a more recent set of studies, inspired by Linkser (1988) and Barlow (1989), which have particularly focused on optimal coding in early vision; Atick & Redlich (1990), Plumbley (1991), Atick et al. (1992), **Atick (1992)**, van Hateren (1992; 1993), Li & Atick (1994a), Dong & Atick (1995), and Dan et al. (1996). Li & Atick (1994b) discuss the extension to joint selectivities of cells in V1; and Li & Atick (1994a) and Li (1996) treat stereo and motion sensitivities as examples.

The statistics of natural sensory inputs is reviewed by **Field (1987)**. Campbell & Gubisch (1966) estimated the optimal modulation transfer function.

We followed the technique of Strong et al. (1998) for computing the mutual information about a dynamical stimulus in spike trains. Bialek et al. (1993) presents an earlier approach based on stimulus reconstruction.