# Chapter 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 theory is a general framework for quantifying the ability of a coding scheme to convey information. It is assumed that the code involves a number of symbols, and the quantities we consider in this chapter, the entropy and mutual information, depend on the frequencies 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 applied to describe a particular set of data.

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 with different numbers of action potentials or responses described by different firing rates. Since 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. To be specific, imagine that we are given a discrete set of neural responses described, for instance, by the number of spikes fired during a fixed period of stimulus presentation. The entropy depends only on these neural responses and not on the stimulus, so, at this point, we do not need to know anything about the stimuli. Knowledge of the stimulus will be needed later when we consider the mutual information. If each response has the same number of spikes, or if only a few different spike counts appear, we might conclude that the data set is relatively uninteresting. A more interesting set might show a larger range of spike numbers, perhaps in a highly irregular and unpredictable sequence. How can we quantify this intuitive notion of an interesting set of responses?

The methods we discuss are based on the probabilities $P[n]$ of observing a response consisting of $n$ spikes. The most widely used measure of entropy, due to Shannon, expresses the 'surprise' associated with seeing an $n$-spike response as a function of the probability of getting that response, $h(P[n])$, and quantifies the entropy as the average over all possible responses of $h(P[n])$. The function $h(P[n])$, which acts as a measure of surprise, is chosen to satisfy a number of conditions. Among these are that $h(P[n])$ should be a decreasing function of $P[n]$, since low probability responses are more surprising than high probability responses. Further, the surprise measure for a complex 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 $n_1$ and $n_2$ spikes from two neurons that respond independently of each other. Because the responses are independent, the probability of getting a pair of responses $n_1$ and $n_2$ is the product of their individual probabilities, $P[n_1]P[n_2]$, so the additivity condition requires that

$$h(P[n_1]P[n_2]) = h(P[n_1]) + h(P[n_2]) . \tag{4.1}$$

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

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

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

$$H = \langle -\log_2 P[n] \rangle = -\sum_n P[n] \log_2 P[n] \,. \tag{4.3}$$

In the sum that determines the entropy, the factor $h = -\log_2 P[n]$ is multiplied by the probability that the response with $n$ spikes occurs. Thus, responses with extremely low probabilities may contribute little to the total entropy, despite having large $h$ values, because they occur so rarely. 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 the same number of spikes. In this case, all of the probabilities $P[n]$ are zero except for one of them which is one. This means that every term in the sum of equation (4.3) is zero because either $P[n] = 0$ or $\log_2(1) = 0$. (The problematic expression $0 \log_2(0)$ is evaluated as $\epsilon \log_2(\epsilon)$ in the limit $\epsilon \to 0$ and is zero.) Thus, a set of identical responses has zero entropy. Next, imagine that the spike counting period is so small that the responses consist of either zero or one spike. In this case, there are only two nonzero terms in equation (4.3), and, using the fact that $P[0] = 1 - P[1]$ in this case, the entropy is

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

This entropy, plotted in figure 4.1A, takes its maximum value of one bit when $P[0] = P[1] = 1/2$. 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 only provide information about a set of stimuli if its response variability is correlated with changes in those stimuli, rather than being purely noise 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

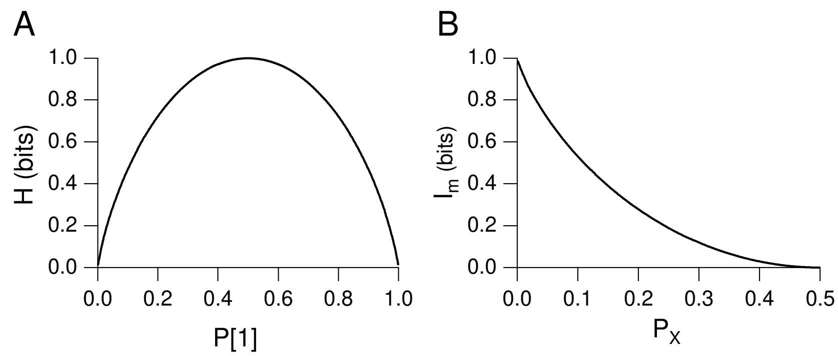



Figure 4.1: A) The entropy of a binary code. $P[1]$ is the probability of a response consisting of one spike and $P[0] = 1 - P[1]$ is the probability of a response with no spikes. The entropy is maximum when $P[0] = P[1] = 1/2$. B) The mutual information for a binary encoding of a binary stimulus. $P_X$ is the probability of an incorrect response being evoked. The plot only shows $P_X \leq 1/2$ 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 $1 - P_X$.

with those measured in trials involving repeated presentations of the same 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 number, 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 $P[n|s]$, the probability of an $n$-spike response given that stimulus $s$ was presented, instead of the response probability $P[n]$ in equation 4.3. The entropy of the responses to a given stimulus is thus

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

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

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

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 the entropy associated with the stimulus, and it is obtained by subtracting the noise entropy from the full response entropy. From equations 4.3 and 4.6, we find that the mutual information $I_{\mathrm{m}}$ is

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

The probability of an $n$-spike response is related to the conditional probability $P[n|s]$ and the probability $P[s]$ that stimulus $s$ is presented by the identity

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

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

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

By Bayes theorem,

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

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

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

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

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

This result is the same as equation 4.7 except that the the role of the stimulus and the response have been interchanged. It provides another way of computing the mutual information (see Rieke *et al*, 1997).

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, $P[n|s] = P[n]$, 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 $n_s$. Then, $P[n_s]{=}P[s]$ and $P[n|s]{=}1$ if $n{=}n_s$ and $P[n|s]{=}0$

otherwise. This causes the sum over $n$ in equation 4.9 to collapse to just one term, and the mutual information becomes

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

The last expression, which follows from the fact that $P[n_s] = P[s]$, 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 $-$. These are presented with equal probabilities, $P[+] = P[-] = 1/2$. Also imagine, as we did in our sample calculation of the entropy, that the neural response consists of either one or no spikes. We associate the one-spike response with the $+$ stimulus, and the no-spike response with the $-$ stimulus. However, the encoding is not perfect, the probability of an incorrect response is $P_{\rm X}$. This means that for the correct responses $P[1|+] = P[0|-] = 1 - P_{\rm X}$, and for the incorrect responses $P[1|-] = P[0|+] = P_{\rm X}$. In this case, $P[0] = P[1] = 1/2$, so, from equation 4.4, the full response entropy is one 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 one 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 zero.

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

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

The KL divergence has a property normally associated with a distance measure, $D_{\rm KL} \ge 0$ 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 $P[n, s]$ and $P[n]P[s]$. If the stimulus and the response were independent of each other, $P[n, s]$ would be equal to $P[n]P[s]$. Thus, the mutual information is the KL divergence between the actual probability distribution $P[n, s]$, and the value it would take if the stimulus and response were independent. The fact that $D_{\rm KL} \ge 0$ 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 spike counts. It is possible to use other measures of the response, but a complication arises if this measure takes continuous values, rather than the integer values of a spike count. Consider, for example a continuous-valued firing rate. If we could measure such a 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. This limit typically cancels in computations of mutual information because these involve taking differences between two entropies.

For a continuously defined firing rate, the probability of the firing rate lying in the range between $r$ and $r + \Delta r$, for small $\Delta r$, is expressed in terms of a probability density as $p[r]\Delta r$. Summing over discrete bins of size $\Delta r$ 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 the the sum of the response probabilities is one. 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.* Reimann integrable) $p[r]$, everything works out fine for the first term since the sum becomes an integral in the limit $\Delta r \to 0$. In this limit, we can write

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

$\Delta r$ 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 only be determined up to an additive constant. However, if two entropies computed with the same resolution are subtracted, the troublesome term involving $\Delta r$ 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.

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 instead of the conditional probability $P[n|s]$,

$$\lim_{\Delta r \to 0} \left\{ H_{\text{noise}} + \log_2(\Delta r) \right\} = -\sum_s \int dr\, p[r|s] \log_2 p[r|s] \,, \tag{4.18}$$

and the mutual information is then

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

Note that the factor of $\log_2 \Delta r$ has canceled in the expression for the mutual information because it is the difference of two entropies. We can also consider stimuli that are described by a continuous variable. In that case, the probability $P[s]$ in equations 4.19 is replaced by the density $p[s]$ and the sum over $s$ is replaced by an integral.

## 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, which means that the mutual information is limited by the resolution with which the response can be read out rather than by extraneous sources of noise, 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 variable like a 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 involve limiting the maximum allowed firing rate or holding fixed the average firing rate or its variance.

## 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 are imposed. During the maximization process, the resolution $\Delta r$ 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 one. Suppose that the neuron in question has a maximum firing rate of $r_{\max}$. Then, the integrals in question extend from 0 to $r_{\max}$, and 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 constraint can be imposed by using a Lagrange multiplier (exercise 1). The result is that the probability density that maximizes the entropy subject to this constraint is a constant,

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

independent of $r$. The entropy for this probability density is

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

The observation that the maximum entropy probability density for a response with a finite range is constant is the basis of a signal processing technique called histogram equalization. What form of neural response assures that $p[r] = 1/r_{\max}$, thereby maximizing the response entropy? Suppose a neuron responds to stimuli characterized by a continuous parameter $s$ by firing at a rate $r = f(s)$, where $f(s)$ is the response tuning curve. The probability that the continuous stimulus variable falls in the range between $s$ and $s + \Delta s$, for small $\Delta s$, 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 $\Delta s$. Then, in the limit $\Delta s \to 0$, the 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 implies that the tuning curve of the responding neuron should be proportional to the integral of the probability density of the stimulus.

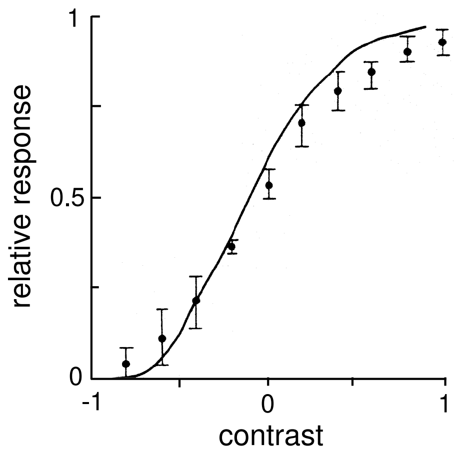



Figure 4.2: Contrast response of the fly LMC (data points) compared to the integral of the natural contrast probability distribution (solid curve). The response is the amplitude of the membrane potential fluctuation produced by the onset of a light or dark image with a given level of contrast. Contrast is defined relative to the level of illumination that produced a half-maximal response, $\bar{I}$. For a light intensity $I$, the contrast is defined as $(I - \bar{I})/\bar{I}$. The relative response is the actual response divided by the maximum response for a contrast of one. (Adapted from Laughlin, 1981)

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 lived. 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 above applies equally to this case. As figure 4.2 indicates, the response tuning curve as a function of contrast is very close to the integrated probability density, suggesting that the LMC is using a maximum entropy encoding.

Even though neurons have maximum firing rates, the constraint $r \leq r_{\max}$ may not always be the factor that limits the entropy. For example, the average firing rate of the neuron may be constrained to values much less than $r_{\max}$, or a constraint on the variance of the firing rate might be the relevant condition. Exercise 2 shows that the entropy maximizing probability density if the average firing rate is constrained to a fixed value is an exponential. A related calculation (see exercise 3) 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 representation in which different neurons encode the same information, even though they do this efficiently. Entropy maximization for a population of neurons requires that the neurons convey independent pieces of information. 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. 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

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

Barlow (1961, 1989) suggested a measure of redundancy for a population of neurons equal to the difference between the sum of the entropies of the individual neurons in the population and the population response entropy of equation 4.26. The entropy for neuron $a$ is $-\int dr_a\, p[r_a] \log_2 p[r_a] - \log_2 \Delta r$ so the Barlow redundancy is

$$B = -\sum_{a=1}^{N} \int dr_a\, p[r_a] \log_2 p[r_a] + \int d\mathbf{r}\, p[\mathbf{r}] \log_2 p[\mathbf{r}]\,. \tag{4.27}$$

We can combine these terms by using the identity

$$p[r_a] = \int \prod_{b \neq a} dr_b\, p[\mathbf{r}] \tag{4.28}$$

to obtain

$$B = \int d\mathbf{r}\, p[\mathbf{r}] \log_2 \left( \frac{p[\mathbf{r}]}{\prod_a p[r_a]} \right)\,. \tag{4.29}$$

This is the KL divergence between the probability densities $p[\mathbf{r}]$ and $\prod_a p[r_a]$, which is always non-negative and is only zero if these two probability densities are equal. Thus, the full population response entropy is never greater than the sum of the entropies of the individual neurons in the population, and it reaches its maximum allowed value when

$$p[\mathbf{r}] = \prod_{a=1}^{N} p[r_a]\,. \tag{4.30}$$

This is exactly the condition that the individual responses are independent of each other. A code that satisfies this condition is called a factorial code because the probability factorizes into a product of single neuron probabilities.

Combining the result on factorization with the results of the previous section, we conclude that the maximum population response entropy can be achieved if the individual neurons act independently, and if they all have identical response probabilities that are flat, exponential, or of some other form that maximize their individual response entropies subject to whatever constraint is being imposed. To maximize the entropy of a population response explicitly, we need to know the constraints that are applied to the neuronal responses. However for a wide class of constraints, entropy maximization involves two principles: *factorization*, making $p[\mathbf{r}] = \prod_a p[r_a]$, and *probability equalization*, making $p[r_a]$ the same for all of the neurons (all $a$ values). The exact form of the optimal individual response distribution depends on the nature of the constraint imposed. Note that the fact that the neurons have the same response probability densities $p[r_a]$ does not imply that they respond identically to every stimulus because the conditional probabilities $p[r_a|s]$ can be different for different neurons.

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 for the neurons in the population will all take the same value and have the same variances, $\langle r_a \rangle = \overline{r}$ and $\langle (r_a - \overline{r})^2 \rangle = \sigma_r^2$ for all $a$. The covariance matrix for a factorized and probability equalized population distribution is thus proportional to the identity matrix,

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

Finding response distributions that satisfy the decorrelation and variance equalization condition of equation 4.31 is usually tractable. In the following examples, we will 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 more equally among the neurons.

## Application to Retinal Ganglion Cell Receptive Fields

Entropy and information maximization have been used to derive the filtering properties of visual receptive fields in the retina, LGN, and primary visual cortex. The basic assumption is that these receptive fields have a form that maximizes 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 since visual objects are extensive in space and time. Maximizing the information conveyed requires removing this redundancy.

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 only analyzes the representational properties of neural responses and ignores the computational goals of the visual system, such as object recognition or target tracking. Nevertheless, it is quite successful at accounting for properties of receptive fields early in the visual pathway.

In chapter 2, a linear estimate of the response of a visual neuron to a stimulus described by the contrast $s(x, y, t)$ was written as

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

(Recall that, following the convention established in chapters 1 and 2, $\langle s \rangle = 0$.) If the spacetime receptive field $D(x, y, \tau)$ is separable, so that $D(x, y, \tau) = D_\mathrm{s}(x, y) D_t(\tau)$, we can rewrite this 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 $s(x, y, t) = s(x, y)s(t)$. Then, $L(t) = L_\mathrm{s} L_t(t)$ where

$$L_\mathrm{s} = \int dxdy\, D_\mathrm{s}(x, y) s(x, y) \tag{4.33}$$

and

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

In the following, we analyze the spatial and temporal components, $D_\mathrm{s}$ and $D_t$, separately by considering the information carrying capacity of $L_\mathrm{s}$ 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.

For the present purposes, it is useful to alter the notation in equation 4.33 slightly by writing the spatial coordinates as a two-component vector, so that the pair of coordinates $(x, y)$ is written as a single vector $\vec{x}$ or $\vec{y}$. 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, meaning that all locations and directions in space (or time), 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

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

Note that we are labeling the neurons by the locations $\vec{a}$ of the centers of their receptive fields. It is convenient to allow $\vec{a}$ to vary continuously rather than taking only the discrete values associated with specific neurons in the population being described. This allows us to replace sums over neurons with integrals over $\vec{a}$. Because the receptive fields of retinal ganglion cells cover the retina densely, with many receptive fields overlapping each point on the retina, 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, but rather will 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, in this case, is the average over all stimuli of the product of the linear responses of two cells, one with receptive field centered at $\vec{a}$ and the other at $\vec{b}$,

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

The average here, denoted by angle brackets, is 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, requires that we set this correlation function proportional to a $\delta$ function,

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

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

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

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

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 $D_s$ and $Q_{ss}$ in terms of their Fourier transforms $\tilde{D}_s$ and $\tilde{Q}_{ss}$,

$$D_s(\vec{x}-\vec{a}) = \frac{1}{4\pi^2}\int d\vec{k}\, \exp\left(-i\vec{k}\cdot(\vec{x}-\vec{a})\right)\tilde{D}_s(\vec{k}) \tag{4.39}$$

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

and writing the $\delta$ function as

$$\delta(\vec{a}-\vec{b}) = \frac{1}{4\pi^2}\int d\vec{k}\, \exp\left(-i\vec{k}\cdot(\vec{a}-\vec{b})\right) \tag{4.41}$$

(we choose to insert the minus sign in front of the $i$ in the integral for the $\delta$ function for convenience). The Fourier transform $\tilde{D}_s$ is in general complex, but because $D_s$ is a real function, we are guaranteed that $\tilde{D}_s(-\vec{k}) = \tilde{D}_s^*(\vec{k})$, where the asterisk denotes complex conjugation. The Fourier transform $\tilde{Q}_{ss}(\vec{k})$ is real and non-negative. This follows from the fact that $Q_{ss}$ is only a function of $|\vec{x}-\vec{y}|$ and is also non-negative, in the sense that for all suitably integrable functions $f(\vec{x})$,

$$\int d\vec{x}d\vec{y}\, f(\vec{x})Q_{ss}(\vec{x}-\vec{y})f(\vec{y}) = \left\langle \int\!\!\int d\vec{x}d\vec{y}\, f(\vec{x})s(\vec{x})s(\vec{y})f(\vec{y})\right\rangle \tag{4.42}$$

$$= \left\langle \left(\int d\vec{x}\, f(\vec{x})s(\vec{x})\right)^2\right\rangle \geq 0\,. \tag{4.43}$$

$\tilde{Q}_{ss}$ is also called the stimulus power spectrum.

Writing 4.36 in terms of Fourier transforms, substituting the resulting expression into 4.37, and performing some standard calculations involving the Fourier representation of $\delta$ functions, we find that

$$\frac{1}{4\pi^2}\int dk\, \exp\left(-i\vec{k}\cdot(\vec{a}-\vec{b})\right)|\tilde{D}_s(\vec{k})|^2\tilde{Q}_{ss}(\vec{k}) = \frac{\sigma_L^2}{4\pi^2}\int dk\, \exp\left(-i\vec{k}\cdot(\vec{a}-\vec{b})\right)\,. \tag{4.44}$$

Because this must hold for all values of $\vec{a}-\vec{b}$, we must have

$$|\tilde{D}_s(\vec{k})|^2\tilde{Q}_{ss}(\vec{k}) = \sigma_L^2 \tag{4.45}$$

from which we find

$$|\tilde{D}_s(\vec{k})| = \frac{\sigma_L}{\sqrt{\tilde{Q}_{ss}(\vec{k})}} . \tag{4.46}$$

The linear kernel described by equation 4.46 exactly compensates for whatever dependence the Fourier transform of the stimulus correlation function has on the spatial frequency $\vec{k}$ such that the product $\tilde{Q}_{ss}(\vec{k})|\tilde{D}_s(\vec{k})|^2$ is independent of $\vec{k}$. This product is the power spectrum of $L$. Because the output of the optimal filter has a power spectrum that is independent of spatial frequency (or white), the kernel in equation 4.46 is called a whitening filter.

Note that the calculation we have performed only determines the amplitude $|\tilde{D}_s(\vec{k})|$ and not $\tilde{D}_s(\vec{k})$ itself. Thus, decorrelation and variance equalization do not uniquely specify the form of the linear kernel. In fact, there is a whole spectrum of appropriate receptive fields that satisfies equation 4.37. Additional solutions can be generated by orthogonal transformations using functions $U(\vec{a}, \vec{c})$ that satisfy the integral condition

$$\int d\vec{c}\, U(\vec{a}, \vec{c}) U(\vec{b}, \vec{c}) = \delta(\vec{a} - \vec{b}) . \tag{4.47}$$

Suppose that we have a filter in the form of equation 4.35 that satisfies equation 4.37. We write a new filter in terms of this old one as

$$L'_s(\vec{a}) = \int d\vec{c}\, U(\vec{a}, \vec{c}) L_s(\vec{c}) . \tag{4.48}$$

The correlation function for this new filter is

$$\langle L'(\vec{a}) L'(\vec{b}) \rangle = \int d\vec{c} d\vec{d}\, U(\vec{a}, \vec{c}) U(\vec{b}, \vec{d}) \langle L_s(\vec{c}) L_s(\vec{d}) \rangle . \tag{4.49}$$

By equation 4.37,

$$\begin{aligned} \langle L'(\vec{a}) L'(\vec{b}) \rangle &= \int d\vec{c} d\vec{d}\, U(\vec{a}, \vec{c}) U(\vec{b}, \vec{d}) \sigma_L^2 \delta(\vec{c} - \vec{d}) \\ &= \sigma_L^2 \int d\vec{c}\, U(\vec{a}, \vec{c}) U(\vec{b}, \vec{c}) = \sigma_L^2 \delta(\vec{a} - \vec{b}) \end{aligned} \tag{4.50}$$

where we have used equation 4.47 in the last line. Thus, if $L_s$ is decorrelated, so is $L'_{|s}$. For an orthogonal transformation that satisfies $U(\vec{a}, \vec{c}) = U(\vec{a} - \vec{c})$, the shapes of the receptive fields of all cells will remain the same. However, it is also possible to choose $U(\vec{a}, \vec{c})$ so that the cells have different receptive fields. An important example of this is a multiscale wavelet representation with different cells tuned to different spatial frequencies in a way that achieves translation and scale invariance. We study consequences of such a representation later in the chapter.

The spatial correlation function for natural scenes has been measured with the result that $\tilde{Q}_{ss}(\vec{k})$ is proportional to $1/|\vec{k}|^2$ over the range it has been evaluated (Field, 1987; Ruderman and Bialek, 1994). The behavior near $\vec{k} = 0$ is not well established, but the divergence of $1/|\vec{k}|^2$ near $\vec{k} = 0$ can be removed by setting $\tilde{Q}_{ss}(\vec{k})$ proportional to $1/(|\vec{k}|^2 + k_0^2)$ where $k_0$ 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 $\tilde{Q}_{ss}(\vec{k})$ 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 (Campbell & Gubisch, 1966) results in an exponential correction to the stimulus correlation function

$$\tilde{Q}_{ss}(\vec{k}) \propto \frac{e^{-\alpha|\vec{k}|}}{|\vec{k}|^2 + k_0^2} \, . \tag{4.51}$$

Putting this into equation 4.46 gives a rather peculiar result; the amplitude $|\tilde{D}_s(\vec{k})|$, being proportional to the inverse of the square root of $\tilde{Q}_{ss}$, is predicted to grow exponentially for large $|\vec{k}|$. Whitening filters maximize entropy by equalizing the distribution of response power over the entire spatial frequency range. High spatial frequency components of images are greatly attenuated by the eye and are relatively rare in natural scenes, and the whitening filter compensates for this by boosting the sensitivity to high spatial frequencies. Although this is the result of the entropy maximization calculation, it is not a good strategy to use in visual processing. Real inputs to retinal ganglion cells involve a mixture of true signal and noise coming from the photoreceptors of the retina. At high spatial frequencies, where 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, and not just maximize the entropy. A full information maximization calculation of the receptive field properties of retinal ganglion cells can be performed (see exercise 4), but this requires introducing a number of assumptions about the constraints that are relevant, and it is not 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 pre-filtered input signal.

### Filtering Input Noise

Suppose that the visual stimulus on the retina is the sum of the true stimulus $s(\vec{x})$ 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 $\tilde{D}_{\rm s}(\vec{k})$ as a product of two terms, a whitening filter, $\tilde{D}_{\rm w}(\vec{k})$, that satisfies equation 4.46 and a noise filter, $\tilde{D}_\eta(\vec{k})$, that eliminates as much of the noise as possible. The Fourier transform of the complete optimal filter is then $\tilde{D}_{\rm s}(\vec{k}) = \tilde{D}_{\rm w}(\vec{k})\tilde{D}_\eta(\vec{k})$.

To determine the form of the noise filter, we demand that when it is applied to the total input $s(\vec{x}) + \eta(\vec{x})$, the result is as close to the signal part of the input, $s(\vec{x})$, as possible. As in the previous section, it is easiest to perform the necessary calculations in the Fourier-transformed spatial frequency domain. Thus, we express the difference between the filtered input and the true signal in terms of the Fourier transforms of the signal and noise, $\tilde{s}(\vec{k})$ and $\tilde{\eta}(\vec{k})$. The Fourier transform of the output of the noise filter is $\tilde{D}_\eta(\vec{k})(\tilde{s}(\vec{k}) + \tilde{\eta}(\vec{k}))$, and we want to make this as close as possible to the Fourier transform of the pure signal, $\tilde{s}(\vec{k})$. To do this, we minimize the integral over $\vec{k}$ of the squared amplitude of the difference between these two quantities,

$$\left\langle \int d\vec{k} \left| \tilde{D}_\eta(\vec{k}) \left( \tilde{s}(\vec{k}) + \tilde{\eta}(\vec{k}) \right) - \tilde{s}(\vec{k}) \right|^2 \right\rangle = \text{minimum} \, , \tag{4.52}$$

averaged over natural scenes. Setting the derivative of this expression with respect to $\tilde{D}^*_\eta(\vec{k}')$ to zero gives

$$\int d\vec{k}\, \tilde{D}_\eta(\vec{k}) \left( \left\langle \tilde{s}(\vec{k})\tilde{s}^*(\vec{k}') \right\rangle + \left\langle \tilde{\eta}(\vec{k})\tilde{\eta}^*(\vec{k}') \right\rangle \right) = \int d\vec{k} \left\langle \tilde{s}(\vec{k})\tilde{s}^*(\vec{k}') \right\rangle . \tag{4.53}$$

In evaluating this expression, we have assumed that the signal and noise terms are statistically independent so that $\langle \tilde{s}(\vec{k})\tilde{\eta}^*(\vec{k}')\rangle = \langle \tilde{\eta}(\vec{k})\tilde{s}^*(\vec{k}')\rangle = 0$. The remaining averages are related to the Fourier transforms of the stimulus-stimulus and noise-noise correlation functions (assuming spatial stationarity in both the stimulus and the noise) by the identities

$$\langle \tilde{s}(\vec{k})\tilde{s}^*(\vec{k}')\rangle = \tilde{Q}_{ss}(\vec{k})\delta(\vec{k} - \vec{k}') \text{ and } \langle \tilde{\eta}(\vec{k})\tilde{\eta}^*(\vec{k}')\rangle = \tilde{Q}_{\eta\eta}(\vec{k})\delta(\vec{k} - \vec{k}') \, . \tag{4.54}$$

Substituting these expressions into equation 4.53 gives

$$\tilde{D}_\eta(\vec{k}) \left( \tilde{Q}_{ss}(\vec{k}) + \tilde{Q}_{\eta\eta}(\vec{k}) \right) = \tilde{Q}_{ss}(\vec{k}) \, , \tag{4.55}$$

which has the solution

$$\tilde{D}_\eta(\vec{k}) = \frac{\tilde{Q}_{ss}(\vec{k})}{\tilde{Q}_{ss}(\vec{k}) + \tilde{Q}_{\eta\eta}(\vec{k})} \, . \tag{4.56}$$

This defines both the phase and magnitude of the noise filter. Combining $\tilde{D}_\eta$ with the whitening filter of equation 4.46, we find that

$$|\tilde{D}_s(\vec{k})| \propto \frac{\sigma_L \sqrt{\tilde{Q}_{ss}(\vec{k})}}{\tilde{Q}_{ss}(\vec{k}) + \tilde{Q}_{\eta\eta}(\vec{k})} . \tag{4.57}$$

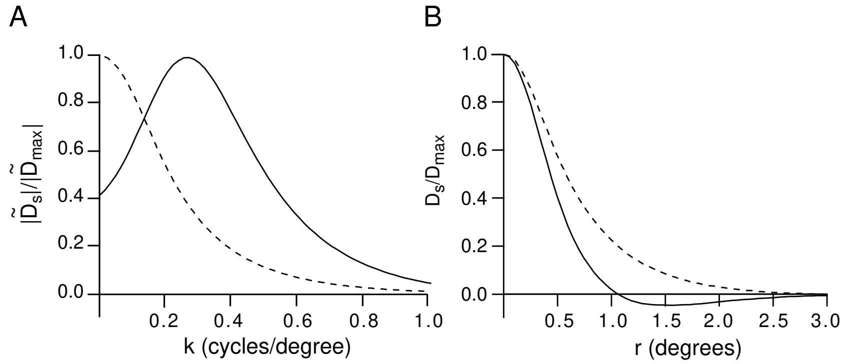



Figure 4.3: Receptive field properties predicted by entropy maximization of responses to natural images. A) The amplitude of the predicted Fourier-transformed linear filters for low (solid curve) and high (dashed curve) input noise. $|\tilde{D}_s(\vec{k})|$ is plotted relative to its maximum value. B) 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. Note the center-surround structure at low noise. $\tilde{D}_s(\vec{k})$ is taken to be real, and $D_s(r)$ is plotted relative to its maximum value. Parameter values used were $\alpha = 0.16$ 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.

Linear kernels resulting from equation 4.57 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 $\vec{k}$. 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 4.57B have been constructed by assuming that $\tilde{D}_s(\vec{k})$ satisfies equation 4.57 and is real, which minimizes the spatial extent of the resulting receptive field. The function $D_s(\vec{x})$ is radially symmetric so it only depends on the distance $r$ 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. 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. 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

Many aspects of our visual world change relatively slowly. This means that there is substantial redundancy in the succession of natural images, which suggests 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

$$\langle L_t(t) L_t(t') \rangle = \sigma_L^2 \delta(t - t') \,. \tag{4.58}$$

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

$$D_t(\tau) = \frac{1}{2\pi} \int d\omega \, \exp(-i\omega\tau) \tilde{D}_t(\omega) \tag{4.59}$$

and $\tilde{D}_t(\omega)$ is given by an equation similar to 4.57,

$$|\tilde{D}_t(\omega)| \propto \frac{\sigma_L \sqrt{\tilde{Q}_{ss}(\omega)}}{\tilde{Q}_{ss}(\omega) + \tilde{Q}_{\eta\eta}(\omega)} \,, \tag{4.60}$$

but, 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

$$\tilde{Q}_{ss}(\omega) \propto \frac{1}{\omega^2 + \omega_0^2} \tag{4.61}$$

where $\omega_0$ is a constant. The resulting filter, in both the temporal frequency and time domains, is plotted in figure 4.4. Figure 4.4A shows the predicted

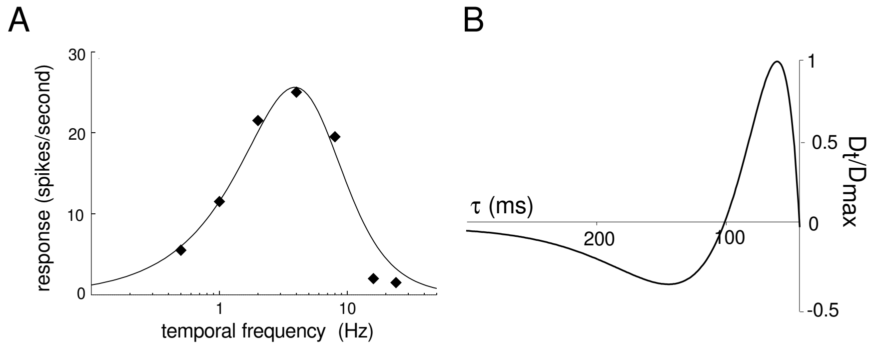



Figure 4.4: A) Predicted (curve) and actual (diamonds) selectivity of an LGN cell as a function of temporal frequency. The predicted curve is based on the optimal linear filter $\tilde{D}_t(\omega)$ with $\omega_0 = 5.5$ Hz. The data are from Saul and Humphrey (1990). B) Causal, minimum phase, temporal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

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 only determines the amplitude of the Fourier transform of the linear kernel, $D_t(\tau)$ is not uniquely specified. To uniquely specify the temporal kernel we require it to be causal ($D_t(\tau) = 0$ for $\tau < 0$) and impose a condition known as minimal phase, which assures that the output will change as rapidly as possible when the stimulus varies. Figure 4.4B shows the resulting form of the temporal filter. The spacetime receptive fields shown in chapter 2 tend to change sign as a function of $\tau$. 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 spectra of spike trains of cat LGN cells in response to natural scenes (the movie *Casablanca*) and figure 4.5B shows power spectra for 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 $\omega$. 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 information transfer are likely to be relevant at the level of cortical processing of visual images. For one thing,

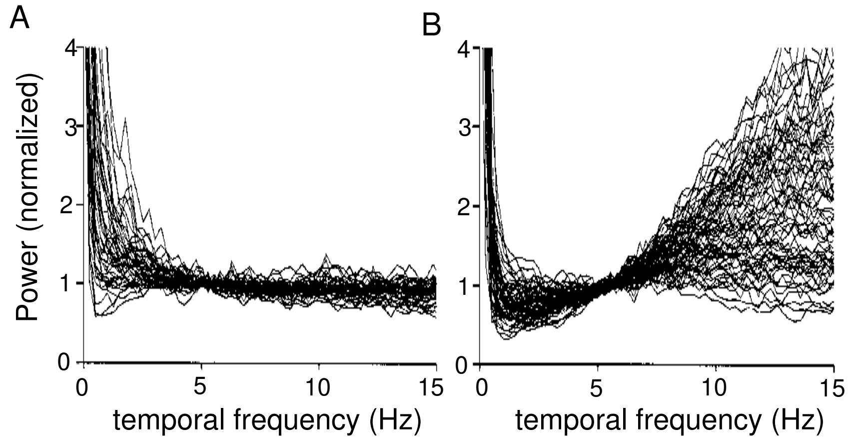



Figure 4.5: A) Power spectra of the spike trains of 51 cat LGN cells in response to presentation of the movie *Casablanca*, normalized to their own values between 5 and 6 Hz. B) Equivalently normalized power spectra of the spike trains of 75 LGN cells in response to white-noise stimuli. (Adapted from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

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. 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 the 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*.

A simple form of the stimulus power spectrum as a function of both spatial and temporal frequency is

$$Q_{ss}(\vec{k}, \omega) \propto \frac{1}{|\vec{k}|^2 + \xi^2 \omega^2} \tag{4.62}$$

where $\xi = 0.4$ 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 small 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 4.56 causes the temporal response to be largest at zero 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

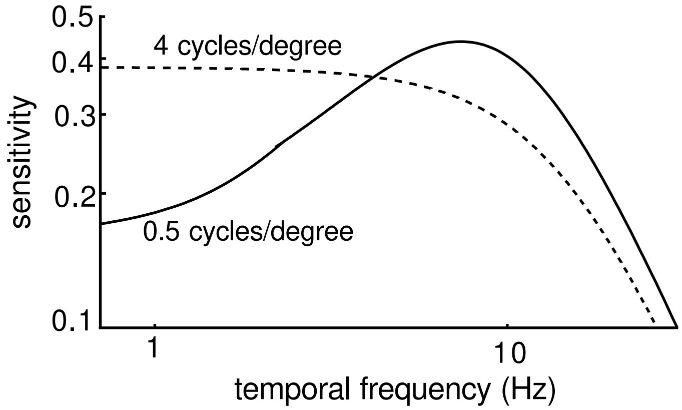



Figure 4.6: Dependence of temporal frequency tuning on preferred spatial frequency for spacetime receptive fields derived from information maximization in the presence of noise. The curves show a transition from partial whitening in temporal frequency for low preferred spatial frequency (solid curve,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

filter causes the response to increase as a function of temporal frequency up to a maximum value where the noise filter begins to supress the response (solid curve in figure 4.6). Model 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 band-pass temporal filters.

Similar conclusions can be drawn concerning other joint selectivities. For instance, color selective (chrominance) cells tend to be selective for low temporal frequencies, because their input signal to noise ratio is lower than that for broad-band (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 should suppress ocular differences.

## 4.3 Entropy and Information for Spike Trains

Computing the entropy or information content of a neuronal response characterized by a sequence of action potentials is much more difficult than computing these quantities for responses described by spike counts or firing rates. Nevertheless, these computations are important because spike counts and firing rates are incomplete descriptions of the full neuronal response. Incomplete descriptions of neural responses can lead to serious underestimates of the information they carry. In this section, we describe two approaches that can be used to compute spike train entropy and information content. Other approaches and 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 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 being considered. Alternately, entropy and mutual information rates can be divided by the total number of action potentials and reported in units of bits per spike rather than bits per second.

To compute the entropy or mutual information rate for a spike train, we need to determine the probabilities that various temporal patterns of action potentials appear. These probabilities replace the factors $P[n]$ or $p[r]$ that occur when spike counts or firing rates are used to characterize a neural response. The temporal pattern of a group of action potentials can be specified by listing the sequence of intervals between successive spikes. For example, we can characterize the temporal pattern of a sequence of $M+1$ action potentials by listing the successive interspike intervals $\tau_1, \tau_2, \ldots, \tau_M$. The probability density needed to compute the entropy or mutual information rate on the basis of $M$-interval sequences is $p[\tau_1, \tau_2, \ldots, \tau_M]$. Because the interspike interval is a continuous variable, we must specify a resolution $\Delta\tau$ with which interspike intervals are measured.

The entropy associated with an $M$-interval distribution is the integral of $p[\tau_1, \tau_2, \ldots, \tau_M] \log_2(p[\tau_1, \tau_2, \ldots, \tau_M](\Delta\tau)^M)$ over all the different $\tau$ values. The average number of interspike intervals that appear in a train per unit time is equal to the average firing rate, $\bar{r}$. The rate at which $M$-interval sequences appear in such a train is then $\bar{r}/M$. Using this rate, the full spike train entropy rate computed from $M$-interval sequences is

$$\left\langle \frac{dH_M}{dt} \right\rangle = -\frac{\bar{r}}{M} \int_0^\infty \prod_{i=1}^{M} d\tau_i \, p[\tau_1, \tau_2, \ldots, \tau_M] \log_2 \left( p[\tau_1, \tau_2, \ldots, \tau_M](\Delta\tau)^M \right) , \tag{4.63}$$

and the noise entropy rate is

$$\begin{aligned} \left\langle \frac{dH_{\text{noise}}}{dt} \right\rangle &= -\frac{\bar{r}}{M} \sum_s P[s] \int_0^\infty \prod_{i=1}^{M} d\tau_i \, p[\tau_1, \tau_2, \ldots, \tau_M | s] \times \\ & \quad \log_2 \left( p[\tau_1, \tau_2, \ldots, \tau_M | s](\Delta\tau)^M \right) . \end{aligned} \tag{4.64}$$

The mutual information is the difference between these two entropies. The entropy and mutual information computed using $M$-interval sequences depend on the value of $M$ used. Formally, the exact answer is obtained by taking the limit $M \to \infty$. In practice, this is never possible and extrapolation methods must be employed.

The difficult task in any spike train entropy calculation is determining the probability densities for $M$-interval sequences. These densities can be approximated as histograms by dividing interspike durations into bins of size $\Delta\tau$ and equating $p[\tau_1, \tau_2, \ldots, \tau_M](\Delta\tau)^M$ to the frequency with which a spike sequence with $\tau$ values lying in a given set of $N$ bins occurs in the data set. In counting the frequency of occurrence of different sequences, two sequences are considered the same if they fall within the same bin in the $M$-dimensional space of $\tau$ values.

Some care must be used in choosing $\Delta\tau$. If $\Delta\tau$ is too small, there will be insufficient numbers of 'identical' spike sequences in the data to estimate their probability of occurrence. If $\Delta\tau$ is too large, we may call spike sequences identical that actual signify different things, and thereby underestimate the information carried by the neuronal response. However, for a reasonable choice of $\Delta\tau$ and given enough data, the calculation of the entropy of a set of spike sequences is straightforward. Once the frequencies of occurrence for different spike sequences in the data set have been determined, the integral in equation 4.63 is re-interpreted as a sum over bins and the occurrence frequencies are used as probabilities in the sum to compute the entropy rate. The noise entropy is computed in a similar way using the frequencies of occurrence in repeated trials involving the same stimulus in place of the conditional probabilities in equation 4.64.

Equations 4.63 and 4.64 give the spike train and noise entropy rates computed from $M$-interval sequences. These formulae are based on the assumption that the different $M$-interval sequences in the complete train are independent of each other so that the total entropy is simply the sum of their individual contributions. If there are correlations between different $M$-interval sequences in the train, this will reduce the total spike train entropy, and equations 4.63 and 4.64 will overestimate the entropy rates. For finite $M$, these equations provide upper bounds on the true entropy rates. If $M$ is too small, there are likely to be correlations between different $M$-interval sequences and the overestimate may be severe. As $M$ increases, we expect the correlations to get smaller, and equations 4.63 and 4.64 should provide more accurate values. For any finite data set, $M$ cannot be increased past a certain point, because there will not be enough $M$-interval sequences in the data to determine their probabilities. Thus, in practice, $M$ must be increased until the point where the extraction of probabilities becomes problematic, and an extrapolation to $M \to \infty$ must be made.

A simple upper bound on the spike train entropy can be obtained by taking $M = 1$, which means that we are considering each interspike interval as if it were independent of all the others. The entropy computed using

$M = 1$ in equation 4.63 gives the bound

$$\left\langle \frac{dH}{dt} \right\rangle \le \left\langle \frac{dH_1}{dt} \right\rangle = -\bar{r} \int_0^\infty d\tau \, p[\tau] \log_2(p[\tau]\Delta\tau) \, . \qquad (4.65)$$

If the spike train is described by a homogeneous process with rate $\bar{r}$, this become an equality, because in this case all the interspike intervals really are independent of each other. For a Poisson train, $p[\tau] = \bar{r}\exp(-\bar{r}\tau)$, so

$$\left\langle \frac{dH}{dt} \right\rangle = \frac{\bar{r}}{\ln(2)} \left(1 - \ln(\bar{r}\Delta\tau)\right) \, . \qquad (4.66)$$

To obtain more accurate results than the $M = 1$ upper bound, we must consider a series of larger $M$ values. An example of such a computation is shown in figure 4.7A where the entropy rate is plotted as a function of $1/M$. Note that as $M$ increases (and $1/M$ decreases), the computed entropy rate decreases. When $M$ reaches a value of around 5 ($1/M = 0.2$), there is a sudden transition and the entropy rate starts to decrease dramatically for larger $M$ values. This is the point where the limited size of the data set starts to affect the calculation. In figure 4.7A, an estimate of the true entropy rate is given by a linear extrapolation of the results obtained for $M \le 5$. The estimated entropy is the $y$-intercept of the straight line in figure 4.7A.

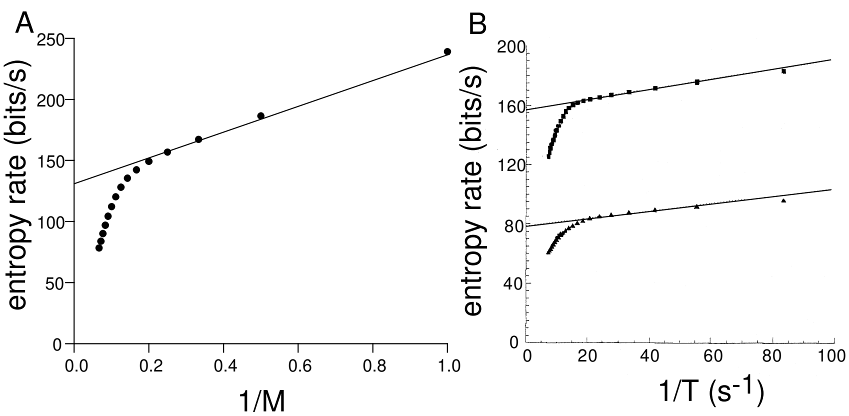



Figure 4.7: A) Entropy rates computed by considering interspike intervals of $M$-interval sequences for different values of $M$. In this example, the spike train was generated using correlated input to a simple spiking model. The resolution was $\Delta\tau = 3$ ms. B) Entropy and noise entropy rates for the H1 visual neuron in the fly responding to a random moving visual image. The upper trace shows the full spike train entropy rate computed for different values of $T$. The straight line is an extrapolation to $T \to \infty$, which is the $y$ intercept. The lower trace shows the spike train noise entropy rate for different values of $T$. The straight line is again an extrapolation to $T \to \infty$. The difference between the $y$ intercepts of the two straight lines is the computed mutual information rate. The resolution was $\Delta t = 3$ ms. (A adapted from Garibay and Wang, 1999; B adapted from Strong *et al*, 1998)

Strong, Koberle, de Ruyter van Steveninck, and Bialek (1998) have proposed an alternative scheme in which calculations are done on the basis of spike sequences lasting for a specific duration $T$, rather than having a given number of spikes. In all other respects, their calculations are like the one described above. Strong *et al* divided time up into bins of size $\Delta t$. They then studied all intervals of duration $T$ by examining all sequences of $T/\Delta t$ time bins in the data set. Spike sequences were characterized by specifying whether or not a spike occurred within each time bin. Probabilities were then extracted by determining the frequency of occurrence of different spiking patterns. As in the case of finite $M$, calculations for finite $T$ overestimate the entropy rates because they ignore correlations over time periods longer than $T$. An estimate of the exact answer is obtained by extrapolating to $T \to \infty$.

Results for the motion sensitive H1 neuron in the fly visual system are shown in figure 4.7B. The two traces in figure 4.7B show the full spike train entropy (upper trace) and noise entropy (lower trace) rates of the response to randomly generated moving visual images. Note that both the full and the noise entropies are overestimated if a small value of $T$ is used. However, both entropies decrease when $T$ increases, until there is a sudden shift to a much steeper fall off. As in figure 4.7A, this transition occurs when $T$ gets large enough so that there is insufficient data to compute the spike sequence probabilities. The solid lines in figure 4.7B show linear extrapolations of the results above this transition point to the left-hand axis which represents the limit $T \to \infty$. The $y$ intercepts of these lines give the best estimates of the entropies. The difference between the full entropy and the noise entropy in figure 4.7B gives the rate at which the spike train carries information about the stimulus being used. This information is 78 bits/s or 1.8 bits/spike. Values in the range 1 to 3 bits/spike are typical results of such calculations.

Assessments of the information conveyed by spike trains can be used to compare responses to different stimuli and thereby reveal aspects of neural encoding. For instance, Rieke, Bodnar & Bialek (1995) compared the information conveyed by single neurons in a peripheral auditory organ (the amphibian papilla) of the bullfrog in response to broad-band 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 46 bits per second (1.4 bits per spike) for broad-band noise and 133 bits per second (7.8 bits per spike) for stimuli with call-like spectra, despite the fact that the broad-band noise has a higher entropy. The spike trains in response to the call-like stimuli convey information with near maximal efficiency.

## 4.4 Appendix - Positivity of the Kullback-Leibler Divergence

Since log is a *concave* function, Jensen's inequality

$$\sum_n P[n] \log f(n) \leq \log \sum_n P[n] f(n) \tag{4.67}$$

holds for any probability distribution $\sum_n P[n] = 1$ and positive function $f(n) > 0$, with equality if and only if $f(n)$ is constant. Applying this to

$$f(n) = \frac{P[n]}{Q[n]} \tag{4.68}$$

where $Q[n]$ is any other probability distribution (with the property that $Q[n] = 0$ implies that $P[n] = 0$) implies that the Kullback-Leibler divergence $D_{\text{KL}}$ satisfies

$$\begin{aligned} -D_{\text{KL}}(P, Q) &= \sum_n P[n] \log \frac{Q[n]}{P[n]} && (4.69) \\ &\leq \log \sum_n Q[n] && (4.70) \\ &= 0 && (4.71) \end{aligned}$$

since $Q[n]$ is a probability distribution with $\sum_n Q[n] = 1$. Further, equality holds if and only if $P[n] = Q[n]$.

## 4.5 Further Reading and References

**Further Reading**

Atick, JJ (1992) Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems* **3**:213-251.

Barlow, HB (1961) Possible principles underlying the transformation of sensory messages. In WA Rosenblith, editor, *Sensory Communication*. Cambridge, MA:MIT Press.

Cover, TM & Thomas, JA (1991) *Elements of Information Theory.* New York, NY: Wiley.

Field, DJ (1987) Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America* **A4**:2379-2394.

Rieke, FM, Warland, D, de Ruyter van Steveninck, R & Bialek, W (1997) *Spikes: Exploring the Neural Code*. Cambridge, MA:MIT Press.

Shannon, CE & Weaver, W (1949) *The Mathematical Theory of Communications*. Urbana:University of Illinois Press.

Uttley, AM (1979) *Information Transmission in the Nervous System*. London:Academic Press.

### References

Atick, JJ, Li, Z & Redlich, AN (1992) Understanding retinal color coding from first principles. *Neural Computation* **4**:559-572.

Atick, JJ & Redlich, AN (1990) Mathematical model of the simple cells in the visual cortex. *Biological Cybernetics* **63**:99-109.

Atick, JJ & Redlich, AN (1990) Towards a theory of early visual processing. *Neural Computation* **2**:308-320.

Atick, JJ & Redlich, AN (1991) Predicting ganglion and simple cell receptive field organizations. *International Journal of Neural Systems* **1**:305-315.

Atick, JJ & Redlich, AN (1992) What does the retina know about natural scenes? *Neural Computation* **4**:196-210.

Attneave, F (1954). Informational aspects of visual perception. *Psychological Review* **61**:183-193.

Barlow, HB (1989) Unsupervised learning. *Neural Computation* **1**:295-311.

Campbell, FW & Gubisch, RW (1966) Optical quality of the human eye. *Journal of Physiology* **186**:558-578.

Dan, Y, Atick, JJ & Reid, RC (1996) Efficient coding of natural scenes in the lateral geniculate nucleus: Experimental test of a computational theory. *The Journal of Neuroscience* **16**:3351-3362.

Dong, DW & Atick, JJ (1995) Temporal decorrelation: A theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems* **6**:159-178.

Laughlin, S (1981) A simple coding procedure enhances a neuron's information capacity. *Z Naturforsch* **36**:910-912.

Li, Z (1996) A theory of the visual motion coding in the primary visual cortex. *Neural Computation* **8**:705-730.

Li, Z & Atick, JJ (1994a) Toward a theory of the striate cortex. *Neural Computation* **6**:127-146.

Li, Z & Atick, JJ (1994b) Efficient stereo coding in the multiscale representation. *Network: Computation in Neural Systems* **5**:157-174.

Linsker, R (1988) Self-organization in a perceptual network. *Computer* **21**:105-117.

Plumbley, MD (1991) *On Information Theory and Unsupervised Neural Networks*. CUED/F-INFENG/TR.78, Cambridge University Engineering Department, Cambridge, England.

Plumbley, MD & Fallside, F (1991) The effect of receptor signal-to-noise levels on optimal filtering in a sensory system. In *ICASSP-91*, **4**:2321-2324.

Rieke F, Bodnar, DA & Bialek, W (1995) Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **262**:259-265.

Ruderman, DL & Bialek, W (1994) Statistics of natural images: Scaling in the woods. *Physical Review Letters* **73**:814-817.

Strong, SP, Koberle, R, de Ruyter can Steveninck, RR & Bialek, W (1998) Entropy and information in neural spike trains. *Physical Review Letters* **80**:197-200.

Snyder, AW, Laughlin, SB & Stavenga, DG (1977) Information capacity of eyes. *Vision Research,* **17**:1163-1175.

Srinivisan, MV, Laughlin, SB & Dubs, A (1982) Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London, Series B: Biological Sciences* **216**:427-459.

van Hateren, JH (1992) Theoretical prediction of spatiotemporal receptive field of fly LMCs and experimental validation. *Journal of Comparative Physiology* **A171**:157-170.

van Hateren, JH (1992) A theory of maximizing sensory information. *Biological Cybernetics* **68**:23-29.

van Hateren, JH (1993a) Three modes of spatiotemporal preprocessing by eyes. *Journal of Comparative Physiology, A* **172**:583-591.

van Hateren, JH (1993b) Spatial, temporal and spectral pre-processing for colour vision. *Proceedings of the Royal Society of London, Series B* **251**:61-68.

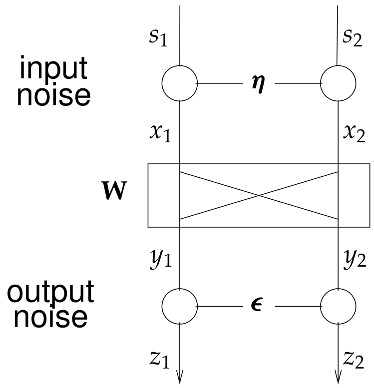
## 4.6 Exercises

1. Show that the distribution that maximizes the entropy when the firing rate is constrained to lie in the range $0 \leq r \leq r_{\text{max}}$ is given by equation 4.22 and its entropy for a fixed resolution $\Delta r$ is given by equation 4.23.

2. Show that the distribution that maximizes the entropy when the mean of the firing rate is held fixed is an exponential, and compute its entropy for a fixed resolution $\Delta r$. Assume that the firing rate can fall anywhere in the range from zero to infinity.

3. Show that the distribution that maximizes the entropy when the mean and variance of the firing rate are held fixed is a Gaussian, and compute its entropy for a fixed resolution $\Delta r$. To simplify the mathematics, allow the firing rate to take any value between minus and plus infinity.

4. Consider the coding problem illustrated in the figure (Plumbley, 1991):




$$\langle \mathbf{s}\mathbf{s}^T \rangle = \mathbf{C}$$
$$\langle \eta_i \eta_j \rangle = \tau^2 \delta_{ij}$$
$$\mathbf{x} = \mathbf{s} + \boldsymbol{\eta}$$
$$\mathbf{y} = \mathbf{W}\mathbf{x}$$
$$\langle \epsilon_i \epsilon_j \rangle = \sigma^2 \delta_{ij}$$
$$\mathbf{z} = \mathbf{y} + \boldsymbol{\epsilon}$$

where $\mathbf{s}, \boldsymbol{\eta}, \boldsymbol{\epsilon}$ have zero-mean Gaussian distributions with the given covariance matrices, and the task is to maximize the information $\mathbf{I}[\mathbf{s};\mathbf{z}]$ given a constraint on the output power $\sum_i \langle y_i^2 \rangle \leq P$.

(a) Show that the matrix form of the function to be optimized is

$$M(\mathbf{W}, \phi) = \log \left| \mathbf{WQW}^T + \sigma^2 \mathbb{I} \right| - \log \left| \tau^2 \mathbf{WW}^T + \sigma^2 \mathbb{I} \right| + \phi \mathrm{tr}(\mathbf{WQW}^T)$$

where $\mathbf{Q} = \mathbf{C} + \tau^2 \mathbb{I}$, $\phi$ is a Lagrange multiplier enforcing the power constraint and $|\mathbf{A}|$ is the determinant of matrix $\mathbf{A}$.

(b) Show that this is rotationally invariant – *i.e.* if $\mathbf{W}$ is a solution then so is $\mathbf{UW}$ for any orthogonal matrix $\mathbf{UU}^T = \mathbf{I}$. Prove that this implies we can choose to express $\mathbf{W}$ as

$$\mathbf{W} = \mathbf{U} \begin{pmatrix} w_1 & 0 \\ 0 & w_2 \end{pmatrix} \mathbf{V}$$

where $\mathbf{V}$ is the collection of eigenvectors of matrix $\mathbf{C}$ and $w_i$ are the filter coefficients.

(c) Re-express $M(\mathbf{W}, \phi)$ in terms of $w_i$, $\sigma^2$, $\tau^2$ and the eigenvalues of $\mathbf{C}$, $c_i^2$.

(d) If $\tau^2 = 0$, show that the optimal solution involves pure whitening: $w_i \propto 1/c_i$

(e) If $\tau^2 \neq 0$, show that for a given $\phi$, $w_i$ satisfies a quadratic equation, and show that $w_i = 0$ unless the input signal/noise ratio satisfies

$$\frac{c_i^2}{\tau^2} > \frac{1}{\frac{1}{-\phi\sigma^2} - 1}$$

5. Using the analysis of exercise 4, plot the optimal values of $w_1$ and $w_2$ as a function of $c_2^2$ for $c_1^2 = 10$, $\sigma^2 = 1$, $\tau^2 = 1$, $P = 1$. Translate the approximate filter design of equation 4.57 to the present case and show the results on the same diagram.

6. Consider the case of four inputs coming from two different spatial positions in each of two eyes, with:

$$\mathbf{C} = \begin{array}{c c} & \begin{array}{cccc} s_1^R & s_2^R & s_1^L & s_2^L \end{array} \\ \begin{array}{c} s_1^R \\ s_2^R \\ s_1^L \\ s_2^L \end{array} & \begin{pmatrix} 1 & r & s & t \\ r & 1 & t & s \\ s & t & 1 & r \\ t & s & r & 1 \end{pmatrix} \end{array}$$

   Show that a coordinate system derived from considering spatial summation ('low' spatial frequency) and difference ('high' spatial frequency) channels:

$$s_+^L = s_1^L + s_2^L \qquad s_+^R = s_1^R + s_2^R$$
$$s_-^L = s_1^L - s_2^L \qquad s_-^R = s_1^R - s_2^R$$

   can be used to diagonalize $\mathbf{C}$ and, in the case that $r = 0.4$, $s = 0.6$, $t = 0.1$, $\sigma^2 = 1$, $\tau^2 = 0.2$, $P = 1$, show that the eigenvalues $c_i^2$ and optimal output weights are given by:

| | sum | | difference | |
|---|---|---|---|---|
| | $*=+$ | | $*=-$ | |
| mode | $c^2$ | $w$ | $c^2$ | $w$ |
| $s_*^L + s_*^R$ | 2.1 | 0.53 | 1.1 | 0.88 |
| $s_*^L - s_*^R$ | 0.7 | 0.63 | 0.1 | 0 |

   Interpret the difference in outcome for the modes in terms of the joint coding for spatial frequency and ocularity.

7. Construct an integrate-and-fire neuron model, and drive it with an injected current with a complex time course. One way to generate the external current is to sum two or more sine waves with incommensurate frequencies. Compute the rate of information about the injected current contained in the spike train produced by this model neuron using one of the methods discussed in the text.