Chapter 3

Neural Decoding

3.1 Encoding and Decoding

In chapters 1 and 2, we considered the problem of predicting neural responses to known stimuli. The nervous system faces the reverse problem, determining what is going on in the real world from neuronal spiking patterns. It is interesting to attempt such computations ourselves, using the responses of one or more neurons to identify a particular stimulus or to extract the value of a stimulus parameter. We will assess the accuracy with which this can be done primarily by using optimal decoding techniques, regardless of whether the computations involved seem biologically plausible. Biophysically realistic implementations for some of these techniques are discussed in chapter 7. Optimal decoding allows us to determine limits on the accuracy and reliability of neuronal encoding. In addition, it is useful for estimating the information content of neuronal spike trains, as discussed in the next chapter.

As Rieke, Warland, de Ruyter van Steveninck and Bialek (1997) have noted, the encoding and decoding of neural responses are related through a basic identity of probability theory, Bayes theorem. Let $\mathbf{r}$ represent the response of a neuron or a population of neurons to a stimulus s. Throughout this chapter, $\mathbf{r} = (r_1, r_2, \ldots, r_N)$ for N neurons is a list of neuronal firing rates. However, for the present discussion it could be any set of parameters describing the neuronal response, such as a list of spike times. $P[s]$ and $P[\mathbf{r}]$ denote the probabilities that stimulus s appears and that response $\mathbf{r}$ is recorded, respectively. To study neural encoding and decoding, we need, in addition, two conditional probabilities: the probability of evoking a response $\mathbf{r}$ given that the stimulus s was presented, denoted by $P[\mathbf{r}|s]$; and the probability that a given stimulus s was presented given that a particular response $\mathbf{r}$ was recorded, $P[s|\mathbf{r}]$. The encoding map is characterized by the probabilities $P[\mathbf{r}|s]$ which reflect features of the encoding process as discussed in chapter 1. The decoding map is defined by the probabilities, $P[s|\mathbf{r}]$. When the responses or stimuli are described by continuous variables, these probabilities must be replaced by the corresponding probability densities, $p[\mathbf{r}]$ or $p[s]$ for example, but all the results we discuss are equally valid.

The different probabilities we have introduced are related. Let $P[\mathbf{r}, s]$ be the probability that $\mathbf{r}$ and s both occur in a trial. This can be expressed as either the probability of the response given the stimulus times the probability of the stimulus, or the probability of the stimulus given the response times the probability of the response:

$$P[\mathbf{r}, s] = P[\mathbf{r}|s]P[s] = P[s|\mathbf{r}]P[\mathbf{r}] \,. \tag{3.1}$$

This is the basis of Bayes theorem relating $P[s|\mathbf{r}]$ to $P[\mathbf{r}|s]$,

$$P[s|\mathbf{r}] = \frac{P[\mathbf{r}|s]P[s]}{P[\mathbf{r}]} \,. \tag{3.2}$$

Unlike the conditional probability relevant to encoding, $P[\mathbf{r}|s]$, the conditional probability $P[s|\mathbf{r}]$ needed for decoding depends on the stimulus probability $P[s]$ (which is called the *prior* probability). As a result, decoding requires knowledge of the statistical properties of experimentally or naturally occurring stimuli.

In the following sections, we discuss several examples of decoding that involve both single neurons and neuronal populations. We first study discrimination between two different stimulus values on the basis of the response of a single cell. We then consider extracting the value of a parameter that characterizes a static stimulus from the time-averaged responses of populations of neurons. As a final example, we return to single neurons and discuss the decoding of time-varying stimuli on the basis of the spike trains they evoke.

3.2 Discrimination

To introduce the notion of *discriminability* and the *receiver operating characteristic* that lie at the heart of discrimination analysis, we will discuss a fascinating study performed by Britten, Shadlen, Newsome and Movshon (1992). In their experiments, a monkey was trained to discriminate between two directions of motion of a visual stimulus. The stimulus was a pattern of dots on a video monitor that jumped from random initial locations to new locations every 45 ms. To introduce a sense of directed movement at a particular velocity, a certain percentage of the dots, chosen randomly at each time step, moved together by a fixed amount in a fixed direction (figure 3.1). The remaining dots moved to random new locations. The percentage of dots that moved together in the fixed direction is called the correlation level. At 0% correlation, the image appears chaotic with no sense of any particular direction of motion. As the correlation increases

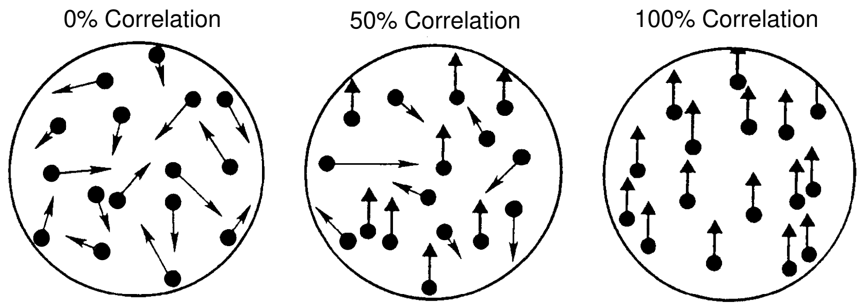

Figure 3.1: The moving random-dot stimulus for different levels of correlation. The visual image consists of randomly placed dots which jump every 45 ms according to the scheme described in the text. At 0% correlation the dots move randomly. At 50% correlation, half the dots move randomly and half move coherently (upwards in this example). At 100% correlation all the dots move coherently. (Adapted from Britten *et al*, 1992)

from zero, a sense of movement in a particular direction appears in the image, until, at 100% correlation, the entire pattern moves coherently on the monitor. By varying the degree of correlation, the task of detecting the movement direction can be made more or less difficult.

The experiment of Britten *et al* (1992) combined neural recording with behavioral measurements. In the behavioral part of the experiment, the monkey had to report the direction of motion in the random dot images. During the same task, recordings were made from neurons in area MT. Only two different possible directions of coherent movement of the dots were used while a particular neuron was being recorded, either the direction that produced the maximum response in that neuron, or the opposite direction. The monkey's task was to discriminate between these two directions. The filled circles and solid curve in figure 3.2A show the proportion of correct responses in a typical experiment. Below 1% correlation the responses were near chance (fraction correct = 0.5) while above 10% correlation the monkey approached perfect performance (fraction correct = 1).

Figure 3.2B shows histograms of firing rates in response to different levels of movement correlation. The firing rates plotted are the number of spikes recorded during the 2 s period that the stimulus was presented, divided by 2 s. The neuron tended to fire more spikes when the motion was in its preferred direction, which we will call the plus (or '+') direction (hatched histogram), than in the other, minus (or '−') direction (solid histogram). For large correlations, the firing-rate distributions corresponding to the two directions are fairly well separated, while for low correlations, they merge. These two distributions are approximately Gaussian, with the same variance, σ_r^2, and different means, $\langle r \rangle_+$ for the plus direction and $\langle r \rangle_-$ for the minus direction. In this case, a convenient measure of the separa-

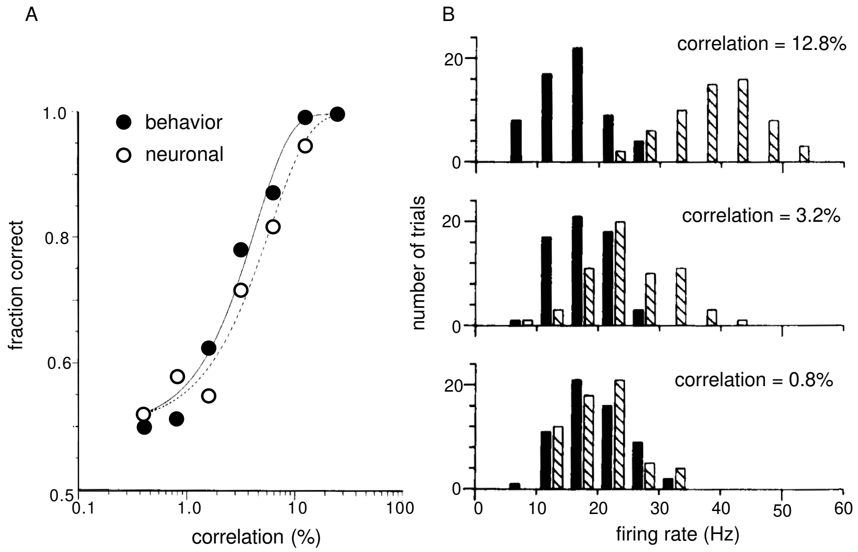

Figure 3.2: Behavioral and electrophysiological results from the random dot motion discrimination task. A) The filled circles show the fraction of correct discriminations made by the monkey as a function of the degree of correlation of the motion. The open circles show the discrimination accuracy that an ideal observer could achieve on the analogous two-alternative forced choice discrimination task using the responses of a single neuron. B) Firing rate histograms for three different levels of discrimination. Hatched rectangles show the results for motion in the plus direction and solid rectangles for motion in the minus direction. The histograms have been thinned for clarity so that not all the bins are shown. (Adapted from Newsome *et al*, 1989)

tion between the distributions is the *discriminability*

$$d' = \frac{\langle r\rangle_+ - \langle r\rangle_-}{\sigma_r} \tag{3.3}$$

which is the distance between the means in units of the common standard deviation. The larger d', the more separated the distributions.

How could we use the neural response to determine which direction the stimulus moved? A simple procedure is to determine the firing rate r during a trial and compare this to a threshold number z. If $r \geq z$ we report plus; otherwise we report minus. Figure 3.2B suggests that if we choose z to lie somewhere between the two distributions, this procedure will give the correct answer at high correlation, but will have difficultly distinguishing the two directions for low correlation. This difficulty is clearly related to the degree to which the two distributions in figure 3.2B overlap, and thus to the discriminability. The probability that this procedure will generate the correct answer when the stimulus is moving in the plus direction is the conditional probability that $r \geq z$ given a plus stimulus, $P[r \geq z|+]$. The probability that it will give the answer plus when the stimulus is actually moving in the minus direction (called a false alarm) is similarly $P[r \geq z|-]$.

These two probabilities completely determine the performance of the decoding procedure because the probabilities for the other two cases, reporting minus when the correct answer is plus and reporting minus when the true answer is minus, are $1 - P[r \geq z|+]$ and $1 - P[r \geq z|-]$ respectively. In signal detection theory, the quantity used to perform the discrimination, r in our case, is called the test and the two probabilities corresponding to reporting a plus answer have specific names:

$$\begin{aligned} \alpha(z) &= P[r \geq z|-] \quad \text{is the } \textit{size} \text{ of the test} \\ \beta(z) &= P[r \geq z|+] \quad \text{is the } \textit{power} \text{ of the test} . \end{aligned} \tag{3.4}$$

The following table shows how the probabilities of the test giving right and wrong answers in the different cases depend on α and β.

	probability	
stimulus	correct	incorrect
$+$	β	$1-\beta$
$-$	$1-\alpha$	α

For a good test, there is a threshold such that the size is near 0 and the power is near 1.

If the probability densities for the firing rates r in the plus and minus cases are Gaussian functions with means $\langle r \rangle_+$ and $\langle r \rangle_-$ and common variance σ_r^2,

$$\alpha(z) = \frac{1}{2}\mathrm{erfc}\left(\frac{z - \langle r \rangle_-}{\sqrt{2}\sigma_r}\right) \quad \text{and} \quad \beta(z) = \frac{1}{2}\mathrm{erfc}\left(\frac{z - \langle r \rangle_+}{\sqrt{2}\sigma_r}\right) \tag{3.5}$$

where erfc is the complementary error function defined by

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty dy \exp(-y^2) . \tag{3.6}$$

The performance of the decoding procedure we have been discussing depends critically on the value of the threshold z to which the rate r is compared. In general, it is impossible to choose the threshold so that both the size and power of the test are optimized; a compromise must be made. A logical optimization criterion is to maximize the probability of getting a correct answer, which is equal to $(\beta(z) + 1 - \alpha(z))/2$ if the plus and minus stimuli occur with equal probability. While this is a possible approach for the experiment we are studying, the analysis of receiver operating characteristics that we present is a powerful technique that makes better use of the full range of recorded data and can be generalized to tasks where the optimal strategy is unknown.

ROC Curves

The receiver operating characteristic (ROC) curve provides a way of evaluating how test performance depends on the choice of the threshold z.

Each point on an ROC curve corresponds to a different value of z. The x coordinate of the point is α, the size of the test for this value of z, and the y coordinate is β, its power. As the threshold is varied continuously, these points trace out the ROC plot. If $z=0$, the firing rate will always be greater than or equal to z, so the decoding procedure will always give the answer plus and $\alpha=\beta=1$, producing a point at the upper-right corner of the ROC plot. At the other extreme, if z is very large, r will always be less than z, the test will always report minus, and $\alpha=\beta=0$. This produces a point at the bottom-left corner of the plot. Between these extremes, a curve is traced out as a function of z.

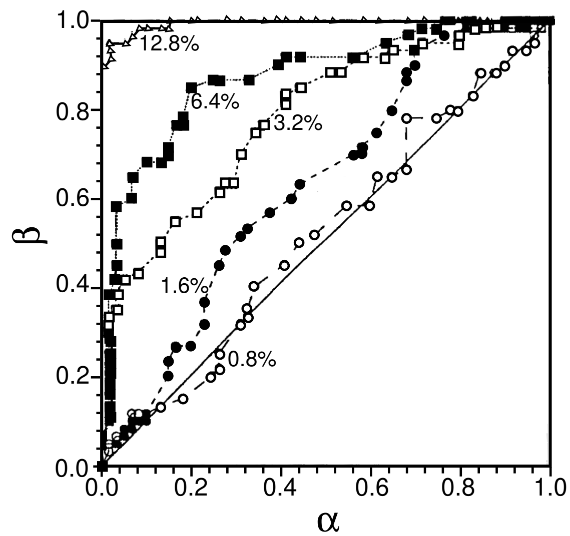

Figure 3.3: ROC curves for a variety of motion correlation levels. Each curve is the locus of points $(\alpha(z), \beta(z))$ for all z values. The values of α and β were computed from histograms like those in figure 3.2B. The diagonal line is the ROC curve for random guessing. (Adapted from Britten *et al*, 1992)

Figure 3.3 shows ROC curves computed by Britten *et al* for several different values of the stimulus correlation. For high correlations, when the task is easy, the ROC curve rises rapidly as the threshold is lowered, and the probability $\beta(z)$ of a correct plus answer quickly approaches 1 without a concomitant increase in α. As the threshold is lowered further, the probability of giving the answer plus when the correct answer is minus also rises, and $\alpha(z)$ increases. When the task is difficult, the curve rises more slowly; and if the task is impossible, so that the test merely gives random answers, the curve will lie along the diagonal $\alpha=\beta$ because the probabilities of answers being right and wrong are equal. This is exactly the trend of the ROC curves at different correlation levels shown in figure 3.3.

Although the ROC curve gives a description of the properties of a test, it does not provide a unique measure of the ability of an observer to perform the discrimination task given the value of r because of the freedom of the observer to choose the threshold z. Considering an alternative task, called two-alternative forced choice, removes this freedom. Here, the stimulus is presented twice, once with motion in the plus direction and once in the minus direction. The task is to decide which presentation corresponded to the plus direction given the firing rates on both trials, r_1 and r_2. A natural extension of the test procedure we have been discussing is to answer trial 1 if $r_1 \geq r_2$ and otherwise answer trial 2. This removes the variable threshold from consideration.

When the ROC curve in figure 3.3 lies along the diagonal, the area underneath it is $1/2$, and this is equal to the the probability of a correct answer in this case. When the task is easy and the ROC curve hugs the left and upper axes in figure 3.3, the area under the curve approaches one. Consideration of the two-alternative forced choice experiment formalizes a relationship between the performance of a test and the area under its ROC curve. Even though the ROC curve is defined in terms of a single observation, the area under it quantifies the probability of error in a two-alternative forced choice task involving two observations.

In the two-alternative forced choice task, the value of r on one trial serves as the threshold for the other trial. For example, if the order of stimulus presentation is plus then minus, the comparison procedure we have outlined will report the correct answer if $r_1 \geq z$ where $z = r_2$, and this has probability $P[r_1 \geq z|+] = \beta(z)$ with $z = r_2$. To determine the probability of getting the correct answer, we need to integrate this probability over all possible values of r_2 weighted by their probability of occurrence. For small Δz, the probability that r_2 takes a value in the range between z and $z + \Delta z$ when the second trial has a minus stimulus is $p[z|-]\Delta z$ where $p[z|-]$ is the conditional firing rate probability density for a firing rate $r = z$. Integrating over all values of z gives the probability of getting the correct answer:

$$P[\text{correct}] = \int dz p[z|-]\beta(z)\,. \tag{3.7}$$

Because the two-alternative forced choice test is symmetric, this is also the probability of being correct if the order of the stimuli is reversed.

Since $\alpha(z)$ is the probability that $r \geq z$ for a minus stimulus, it can be written as an integral of the conditional firing-rate probability density $p[r|-]$,

$$\alpha(z) = \int_z^\infty dr\, p[r|-]\,. \tag{3.8}$$

Taking the derivative of this equation with respect to z we find that

$$\frac{d\alpha}{dz} = -p[z|-]\,. \tag{3.9}$$

This allows us to make the replacement $dz\, p[z|-] \rightarrow -d\alpha$ in the integral of equation 3.7 and to change the integration variable from z to α. Noting that $\alpha = 1$ when $z = 0$ and $\alpha = 0$ when $z = \infty$, we find

$$P[\text{correct}] = \int_0^1 d\alpha\, \beta \tag{3.10}$$

This is exactly the area under the ROC curve, and thus the area under the ROC curve provides a measure of the quality of a test.

Consider the case we mentioned above where $p[r|+]$ and $p[r|-]$ are both Gaussian functions with means $\langle r \rangle_+$ and $\langle r \rangle_-$, and a common variance σ_r^2. In the two-alternative forced choice test, the answer will be correct if the order of presentation is plus then minus and $r_1 > r_2$ or if the order is minus then plus and $r_2 > r_1$. In either case, the probability is

$$P[\text{correct}] = \frac{1}{2}\text{erfc}\left(\frac{\langle r \rangle_- - \langle r \rangle_+}{2\sigma_r}\right) = \frac{1}{2}\text{erfc}\left(-\frac{d'}{2}\right) \tag{3.11}$$

where d' is the discriminability defined in equation 3.3. In the case that the distributions are equal-variance Gaussians, the relationship between the discriminability and the area under the ROC curve is invertible because the complementary error function is monotonic. It is common to quote d' values even for non-Gaussian distributions by inverting the relationship between $P[\text{correct}]$ and d' in equation 3.11.

To interpret their experiment as a two-alternative forced choice case, Britten *et al* imagined that, in addition to being given the firing rate of the recorded neuron during stimulus presentation, the observer is given the firing rate of a hypothetical 'antineuron' having exactly the opposite response characteristics from the recorded neuron. In reality, the responses of this antineuron to a plus stimulus were just those of the recorded neuron to a minus stimulus, and vice versa. The idea of using the responses of a single neuron to opposite stimuli as if they were the simultaneous responses of two different neurons reappears again in our discussion of spike train decoding.

The result of this analysis is that the areas under the ROC curves shown in figure 3.3 give the probability that an ideal observer will successfully decode the neural responses and correctly identify the movement direction, given the firing rates of the recorded neuron and its hypothetical antineuron partner. Figure 3.2A shows a typical result for one recorded neuron. Amazingly, the ability of the ideal observer to perform the discrimination task using a single neuron-antineuron pair is equal to the ability of the monkey to do the task. This seems remarkable because the monkey presumably has access to a large population of such neurons. While the average performance of an ideal observer matches the performance of the animal, the two do not agree on a trial-by-trial basis. Nevertheless it is impressive and somewhat mystifying that by decoding a pair of neurons we can discriminate, on average, as well as the monkey.

The discrimination test we have considered compares the firing rate to a threshold value. Could an observer do better than this already remarkable performance by comparing some other test function to a threshold? What is the best test function to use for this purpose? The Neymann-Pearson lemma (proven in Appendix A) shows that it is impossible to do better than to chose the test function to be the ratio of probability densities (or, where appropriate, probabilities),

$$l(r) = \frac{p[r|+]}{p[r|-]}\,, \tag{3.12}$$

known as the likelihood ratio. The test function used above, r itself, is not equal to the likelihood ratio. However, if the likelihood is a monotonically increasing function of r, as it is for the data of Britten *et al*, the firing-rate threshold test is equivalent to using the likelihood ratio and is indeed optimal. Similarly, any monotonic function of the likelihood ratio will provide as good a test as the likelihood itself, and the logarithm is frequently used.

There is a direct relationship between the likelihood ratio and the ROC curve. As in equations 3.8 and 3.9, we can write

$$\beta(z) = \int_z^\infty dr\; p[r|+] \qquad \frac{d\beta}{dz} = -p[z|+]\,. \tag{3.13}$$

Combining this result with 3.9, we find that

$$\frac{d\beta}{d\alpha} = \frac{d\beta}{dz}\frac{dz}{d\alpha} = \frac{p[z|+]}{p[z|-]} = l(z)\,, \tag{3.14}$$

so the gradient of the ROC curve at the point defined by a given z is the likelihood ratio for that threshold value.

Another way of seeing that comparing the likelihood ratio to a threshold value is an optimal decoding procedure for discrimination uses a Bayesian approach based on associating a cost or penalty with getting the wrong answer. The penalty associated with answering minus when the true answer is plus will be quantified by the loss parameter $\mathcal{L}_-$. Similarly, the loss for answering plus when the correct answer is minus is $\mathcal{L}_+$. For convenience, we will assume that there is neither loss nor gain for answering correctly. The probabilities that the correct answer is plus or minus given the firing rate r are $P[+|r]$ and $P[-|r]$ respectively. These probabilities are related to the conditional firing-rate probability densities by Bayes Theorem,

$$P[+|r] = \frac{p[r|+]P[+]}{p[r]} \quad \text{and} \quad P[-|r] = \frac{p[r|-]P[-]}{p[r]}\,. \tag{3.15}$$

The average loss expected for a plus answer when the firing rate is r is the loss associated with being wrong times the probability of being wrong,

$$\text{Loss}_+ = \mathcal{L}_+ P[-|r]\,. \tag{3.16}$$

Similarly the expected loss when answering minus is

$$\text{Loss}_- = \mathcal{L}_- P[+|r]\,. \tag{3.17}$$

A reasonable strategy is to 'cut the losses', answering plus if $\text{Loss}_+ \leq \text{Loss}_-$ and minus otherwise. Combining the above three equations, we find that this strategy gives the response plus if

$$l(r) = \frac{p[r|+]}{p[r|-]} \geq \frac{\mathcal{L}_+}{\mathcal{L}_-}\frac{P[-]}{P[+]}\,. \tag{3.18}$$

This shows that the strategy of comparing the likelihood ratio to a threshold is a way of minimizing the expected loss. The right hand side of this inequality gives an explicit formula for the value of the threshold that should be used, and reflects two factors. One is the relative losses for the two sorts of possible errors. The other is the relative prior probabilities that the answer is plus or minus. Interestingly, it is possible to change the thresholds that human subjects use in discrimination tasks by manipulating these two factors.

If the condition probability densities $p[r|+]$ and $p[r|-]$ are Gaussians with means r_+ and r_- and identical variances σ^2, and $P[+] = P[-] = 1/2$, the probability $P[+|r]$ is a sigmoidal function of r

$$P[+|r] = \frac{1}{1 + \exp(-d'(r - r_{\text{ave}})/\sigma)} \tag{3.19}$$

where $r_{\text{ave}} = (r_+ + r_-)/2$. This provides an alternate interpretation of the parameter d', often used in the psychophysics literature. It determines the slope of a sigmoidal fit to $P[+|r]$.

We have so far considered discriminating between two quite distinct stimulus values, plus and minus. Often we are interested in discriminating between two stimulus values $s+\Delta s$ and s that are very close to each other. In this case, the likelihood ratio is

$$\begin{aligned} \frac{p[r|s+\Delta s]}{p[r|s]} &\approx \frac{p[r|s] + \Delta s \partial p[r|s]/\partial s}{p[r|s]} \\ &= 1 + \Delta s \frac{\partial \ln p[r|s]}{\partial s} \end{aligned}$$

For small Δs, a test that compares

$$Z(r) = \frac{\partial \ln p[r|s]}{\partial s} \tag{3.20}$$

to a threshold $(z-1)/\Delta s$ is equivalent to the likelihood ratio test for the discrimination.

Although we mainly discussed discrimination between stimuli on the basis of the response of a single neuron, neurons typically represent information collectively through the activities of large populations. If a whole population of neurons responds to a stimulus with rates $\mathbf{r} = (r_1, r_2, \ldots, r_N)$, the only essential difference is that terms such as $p[r|s]$ are replaced by $p[\mathbf{r}|s]$, the conditional probability density of the population response $\mathbf{r}$. ROC analysis, likelihood ratio tests, and the Neymann-Pearson lemma continue to apply in exactly the same way. In the next section, we discuss models of $p[\mathbf{r}|s]$ for population decoding of stimulus values.

3.3 Decoding

Decoding requires us to estimate the value of a stimulus from the response that it evokes. In this section, we study how the value of a parameter associated with a static stimulus can be decoded from the time-averaged firing rates of a population of neurons. In such a population, individual neurons typically have different but overlapping selectivities so that many neurons, but not necessarily all, respond to a given stimulus. The use of large numbers of neurons to perform tasks is one of the basic operating principles of most nervous systems. Population coding has a number of advantages, including reduction of uncertainty due to neuronal variability, and the ability to represent a number of different stimulus attributes simultaneously.

Encoding and Decoding Direction

The cercal system of the cricket is an interesting example of population coding involving a relatively small number of neurons. Crickets and other related insects have two appendages called cerci extending from their hind ends. These are covered with hairs that are deflected by air currents. Each hair is attached to a neuron that fires when the hair is deflected. Thousands of these primary sensory neurons send axons to a set of interneurons that relay the sensory information to the rest of the cricket's nervous system. The cercal system is used to report the direction of incoming air currents as a warning of approaching predators. Theunissen and Miller (1991) measured both the mean and the variance of responses of cercal interneurons while blowing air currents at the cerci. At low wind velocities, information about wind direction is encoded by just four interneurons.

Figure 3.4 shows average firing rate tuning curves for the four relevant interneurons as a function of wind direction. These neurons are sensitive primarily to the angle of the wind s around the vertical axis and not to its elevation above the horizontal plane. Wind speed was held constant in these experiments so we do not discuss how it is encoded. The interneuron tuning curves are well approximated by rectified cosine functions. Neuron a (where $a = 1, 2, 3, 4$) responds with a maximum average firing rate when the angle of the wind direction is s_a, the preferred direction angle for that neuron. The tuning curve for interneuron a, that gives the average firing rate $\langle r_a \rangle = f_a(s)$ in response to wind direction s, normalized to its

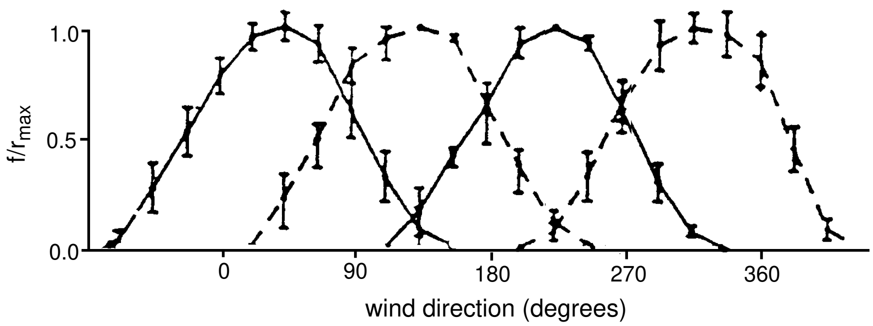

Figure 3.4: Tuning curves for the four low-velocity interneurons of the cricket cercal system. Each neuron responds with a firing rate that closely approximated by a rectified cosine function. The preferred directions of the neurons are located 90° from each other, and $r_{\max}$ values are typically around 40 Hz. Error bars show standard deviations. (Adapted from Theunissen and Miller, 1991)

maximum, can be written as

$$\left\langle \left(\frac{r}{r_{\max}} \right)_a \right\rangle = \left(\frac{f(s)}{r_{\max}} \right)_a = [(\cos(s - s_a)]_+ \tag{3.21}$$

where the rectification forces the rate to be positive. Here $r_{\max}$, which may be different for each neuron, is a constant that determined the maximum average firing rate. The fit can be improved somewhat by introducing a small offset rate, but the simple cosine is adequate for our purposes. No single interneuron of the cercal system responds for all wind directions since each tuning curve covers only half the full range of angles. Rather, the four interneurons encode the wind direction collectively as a population.

To determine the wind direction from the firing rates of the cercal interneurons it is useful to change the notation somewhat. In place of the angle s, we can represent wind direction by a spatial vector $\vec{v}$ pointing parallel to the wind velocity and having unit length, $|\vec{v}| = 1$ (we use over-arrows to denote spatial vectors). Similarly, we can represent the preferred wind direction for each interneuron by a vector $\vec{c}_a$ of unit length pointing in the direction specified by the angle s_a. Since the dot product and cosine are equal for unit vectors, $\vec{v} \cdot \vec{c}_a = \cos(s - s_a)$ and the average firing rate is proportional to the rectified projection of the wind direction vector onto the preferred direction axis of the neuron,

$$\left(\frac{f(s)}{r_{\max}} \right)_a = \left[\vec{v} \cdot \vec{c}_a \right]_+ . \tag{3.22}$$

Decoding the cercal system is particularly easy because of the close relationship between the representation of wind direction it provides and a two-dimensional Cartesian coordinate system. In a Cartesian system, vectors are parameterized by their projections onto x and y axes, v_x and v_y.

These projections can be written as dot products of the vector being represented, $\vec{v}$, with vectors of unit length $\vec{x}$ and $\vec{y}$ lying along the x and y axes, $v_x = \vec{v} \cdot \vec{x}$ and $v_y = \vec{v} \cdot \vec{y}$. Except for the rectification, these equations are identical to 3.22. Furthermore, the preferred directions of the four interneurons, like the x and y axes of Cartesian coordinates, lie along two perpendicular directions (figure 3.5A). Four neurons are required, rather than two, because firing rates cannot represent negative projections. The cricket discovered the Cartesian coordinate system long before Descartes, but failed to invent negative numbers! Perhaps credit should also be given

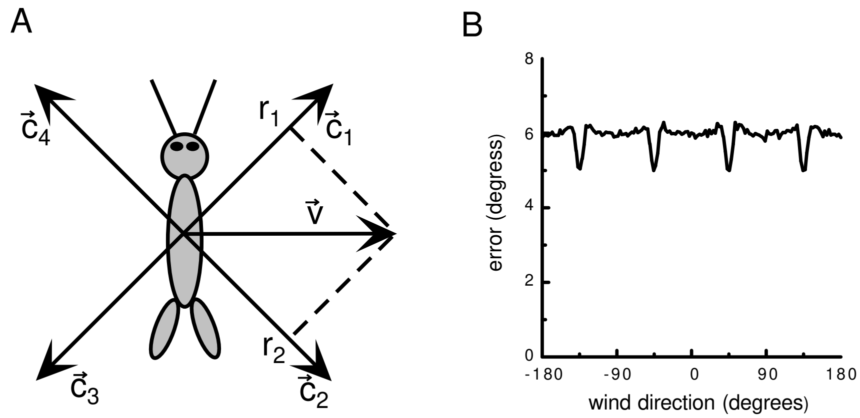

Figure 3.5: A) Preferred directions of four cercal interneurons in relation to the cricket's body. The average firing rate of each neuron for a fixed wind speed is proportional to the projection of the wind velocity vector $\vec{v}$ onto the preferred direction axis of the neuron. The projection directions $\vec{c}_1, \vec{c}_2, \vec{c}_3$ and $\vec{c}_4$ for the four neurons are separated by 90° and they collectively form a Cartesian coordinate system. B) The root-mean-square error in the wind direction determined by vector decoding of the firing rates of four cercal interneurons. These result were obtained through simulation by randomly generating interneuron responses to a variety of wind directions, with the average values and trial-to-trial variability of the firing rates matched to the experimental data. The generated rates were then decoded using equation 3.23 and compared to the wind direction used to generate them. (Adapted from Salinas and Abbott, 1994)

to the leech, for Lewis and Kristan (1997) have shown that the direction of touch sensation in its body segments is encoded by four neurons in a virtually identical arrangement.

A vector $\vec{v}$ can be reconstructed from its Cartesian components through the component-weighted vector sum $\vec{v} = v_x\vec{x} + v_y\vec{y}$. Since the firing rates of the cercal interneurons we have been discussing are proportional to the Cartesian components of the wind direction vector, a similar sum should allow us to reconstruct the wind direction from a knowledge of the interneuron firing rates, except that four, not two, terms must be included. If r_a is the firing rate of neuron a, an estimate of the wind direction on any given trial can be obtained from the direction of the vector

$$\vec{v}_{\text{pop}} = \sum_{a=1}^{4} \left(\frac{r}{r_{\max}} \right)_a \vec{c}_a \tag{3.23}$$

This decoding scheme works quite well. Figure 3.5B shows the root-mean-square difference between the direction determined by equation 3.23 and the actual wind direction that evoked the firing rates. The typical difference between the decoded and actual wind directions is 6° except for dips at the angles corresponding to the preferred directions of the neurons. The average error is close to the theoretical limit imposed by the variability of the neuronal responses, as computed by Theunissen and Miller (1991). The decoding method used in figure 3.5 is called the vector method, and the vector computed from Equation 3.23 is known as the population vector. The vector method is a particularly simple approach to decoding, and it can produce accurate results if the neuronal tuning curves are cosine functions and the preferred directions of the neurons are appropriately arranged.

As discussed in chapter 1, tuning curves of certain neurons in the primary motor cortex (M1) of the monkey can be described by cosine functions of arm movement direction. Thus, a vector decomposition similar to that of the cercal system appears to take place in M1. Many M1 neurons have nonzero offset rates so they can represent the cosine function over most or all of its range. When an arm movement is made in the direction represented by a vector of unit length, $\vec{v}$, the average firing rates for such an M1 neuron labeled by an index a (assuming that it fires over the entire range of angles) can be written as

$$\left\langle \left(\frac{r - r_0}{r_{\max}} \right)_a \right\rangle = \left(\frac{f(s) - r_0}{r_{\max}} \right)_a = \vec{v} \cdot \vec{c}_a \tag{3.24}$$

for a set of preferred direction vectors $\vec{c}_a$. Since these firing rates represent the full cosine function, it would, in principle, be possible to encode all movement directions in three dimensions using just three neurons. Instead, many thousands of M1 neurons have arm-movement related tuning curves resulting in a highly redundant representation. These neurons encode additional movement-related quantities, for example, their firing rates depend on the initial position of the arm relative to the body as well as on movement direction. This complicates the interpretation of their activity as reporting movement direction in a particular coordinate system.

Unlike the cercal interneuron, M1 neurons do not have preferred directions that form a Cartesian coordinate system so the Cartesian sum, equation 3.23, is not the correct way to reconstruct $\vec{v}$ from its projections. Instead, the preferred directions of the neurons appear to point in all directions with roughly equal probability. As we see later in this chapter, corrections should be made for the nonorthogonal directions along which the projections are being made. However, if the preferred directions point uniformly in all directions and the number of neurons N is sufficiently large, the population vector

$$\vec{v}_{\text{pop}} = \sum_{a=1}^{N} \left(\frac{r - r_0}{r_{\text{max}}} \right)_a \vec{c}_a \tag{3.25}$$

will, on average, point in a direction parallel to the arm movement direction vector $\vec{v}$. If we substitute equation 3.24 into 3.25 and average, we find

$$\langle \vec{v}_{\text{pop}} \rangle = \sum_{a=1}^{N} (\vec{v} \cdot \vec{c}_a) \vec{c}_a \,. \tag{3.26}$$

We leave as an exercise (problem 3) the proof that $\langle \vec{v}_{\text{pop}} \rangle$ is approximately parallel to $\vec{v}$ if a large enough number of neurons is included in the sum, and if their preferred direction vectors point randomly in all directions with equal probability.

Georgopoulos and collaborators (1986 & 1988) have used the population vector method to reconstruct arm movement directions from the activity of neurons in M1 and other areas of the brain recorded while monkeys performed reaching tasks. Figure 3.6 compares the population vector constructed from equation 3.25 to the actual directions of arm movements.

Optimal Decoding Methods

The vector method is useful in certain applications but it is neither a general nor an optimal way to reconstruct a stimulus from the firing rates of a population of neurons. Optimal decoding methods make use of the conditional probability that a stimulus parameter takes a value between s and $s+\Delta s$ given that the set of N encoding neurons fired at given rates $\mathbf{r}$. The probability density needed for a continuous stimulus parameter, $p[s|\mathbf{r}]$, can be obtained from the encoding probability density $p[\mathbf{r}|s]$ by the continuous version of Bayes theorem (equation 3.2),

$$p[s|\mathbf{r}] = \frac{p[\mathbf{r}|s]p[s]}{p[\mathbf{r}]} \,. \tag{3.27}$$

We consider two approaches that extract an estimate of the stimulus from $p[s|\mathbf{r}]$ in different ways. These are called Bayesian and maximum *a posteriori*, or MAP, inference. We also discuss a special case of MAP called maximum likelihood, or ML, inference.

Bayesian inference uses the full probability density $p[s|\mathbf{r}]$ to determine an estimate of the stimulus that minimizes a particular loss function $L(s, s_{\text{bayes}})$ that quantifies the 'cost' of reporting the estimate s_{bayes} when the true answer is s. The value of s_{bayes} is chosen to minimize the expected loss averaged over all stimuli for a given set of rates

$$\int ds\, L(s, s_{\text{bayes}}) p[s|\mathbf{r}] = \text{minimum} \,. \tag{3.28}$$

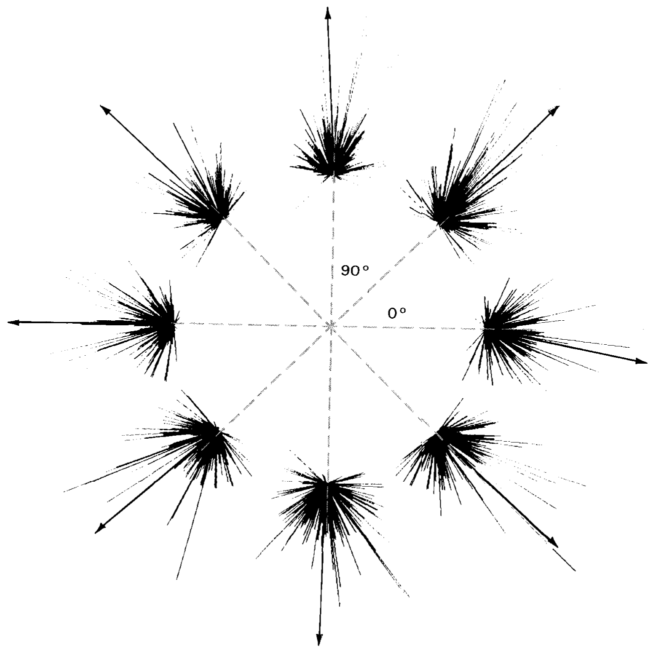

Figure 3.6: Comparison of population vectors with actual arm movement directions. Results are shown for eight different movement directions. Actual arm movement directions are radially outward at angles that are multiples of 45°. The groups of lines without arrows show the preferred direction vectors of the recorded neurons multiplied by their firing rates. Vector sums of these individual terms for each movement direction are indicated by the arrows. The fact that the arrows point approximately radially outward shows that the population vector reconstructs the actual movement direction fairly accurately. (Figure from Kandel *et al*, 1991)

If the loss function is the squared difference between the estimate and the true value, $L(s, s_{\text{bayes}}) = (s - s_{\text{bayes}})^2$, the estimate that minimizes the expected loss is the mean

$$s_{\text{bayes}} = \int ds\, p[s|\mathbf{r}]s \, . \tag{3.29}$$

If the loss function is the absolute value of the difference, $L(s, s_{\text{bayes}}) = |s - s_{\text{bayes}}|$, s_{bayes} is the median rather than the mean of the distribution $p[s|\mathbf{r}]$. The loss function provides a way of defining the optimality criterion for decoding analogous to the loss computation discussed previously for optimal discrimination.

Maximum *a posteriori* (MAP) inference ignores the complexities of the loss function and simply chooses the stimulus value that maximizes the conditional probability density of the stimulus,

$$p[s_{\text{MAP}}|\mathbf{r}] = \text{maximum} \, . \tag{3.30}$$

The MAP approach is thus to choose as the estimate $s_{\rm MAP}$ the most likely stimulus value for a given set of rates. If the prior density of the stimulus $p[s]$ is flat (independent of s), $p[s|\mathbf{r}]$ and $p[\mathbf{r}|s]$ have the same dependence on s, because the additional factor $p[\mathbf{r}]$ in equation 3.27 is independent of s. In this case, the MAP algorithm is equivalent to maximizing the likelihood function,

$$p[\mathbf{r}|s_{\rm ML}] = \text{maximum} \tag{3.31}$$

where $s_{\rm ML}$ is called the maximum likelihood estimate. Later, we will define the sense in which ML and MAP estimation are optimal decoding strategies.

Previously we applied the vector decoding method to the cercal system of the cricket. Figure 3.7 shows the root-mean-square difference between the true and estimated wind directions for the cercal system using ML and Bayesian methods. For the cercal interneurons, the response probability density $p[\mathbf{r}|s]$ is a product of four Gaussians with means and variances given by the data points and error bars in figure 3.4. The Bayesian estimate in figure 3.7 is based on the squared-difference loss function. Comparing these results with figure 3.5B shows the improved performance of these methods relative to the vector method. The vector method is ideally suited for this system so the degree of improvement is not large. Much more dramatic differences occur in systems less suited to vector decoding (such as that shown in figure 3.13).

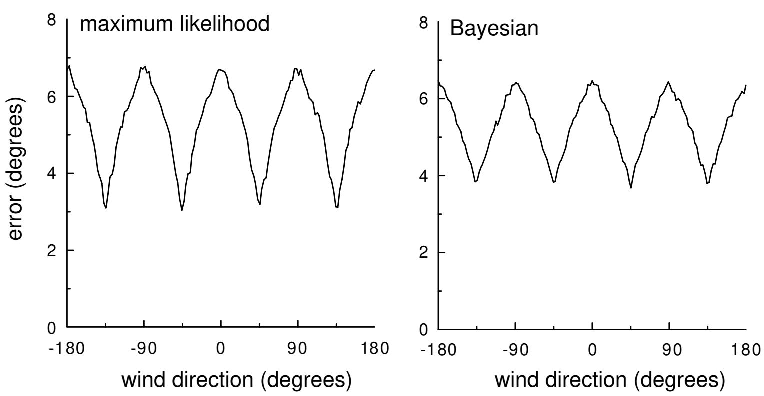

Figure 3.7: Maximum likelihood and Bayesian estimation errors for the cricket cercal system. ML and Bayesian estimates of the wind direction were compared with the actual stimulus value for a large number of simulated firing rates. Firing rates were generated and estimation was made using Gaussian firing rate probability densities with means and standard deviations extracted from the data. The error shown is the root-mean-squared difference between the estimated and actual stimulus angles. (Adapted from Salinas and Abbott, 1994)

An instructive example of population coding and decoding is provided by an array of N neurons with preferred values distributed uniformly across the full range of possible stimuli. An example of such an array for Gaussian tuning curves,

$$f_a(s) = r_{\max} \exp\left(-\frac{1}{2}\left(\frac{s - s_a}{\sigma_a}\right)^2\right) \tag{3.32}$$

is shown in figure 3.8. In this example, each neuron has a tuning curve with a different preferred value s_a and potentially a different width σ_a (although all the curves in figure 3.8 have the same width). If the tuning curves are evenly and densely distributed across the range of s values, the sum of all tuning curves $\sum f_a(s)$ will be approximately independent of s. The roughly flat line in figure 3.8 is proportional to this sum. The constancy of the sum over tuning curves will be useful in the following analysis.

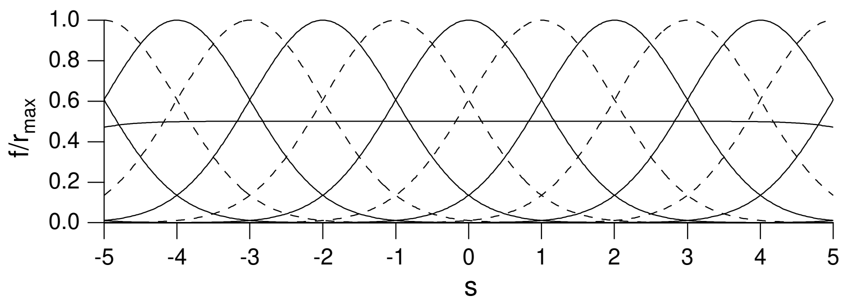

Figure 3.8: An array of Gaussian tuning curves spanning stimulus values from -5 to 5. The peak values of the tuning curves fall on the integer values of s and the tuning curves all have $\sigma_a = 1$. For clarity, the curves are drawn alternately with dashed and solid lines. The approximately flat curve at the value 0.5 is $1/5$ the sum of the tuning curves shown, indicating that this sum is roughly independent of s.

Tuning curves give the mean firing rates of the neurons across multiple trials. In any single trial, measured firing rates will vary from their mean values. To implement the Bayesian, MAP, or ML approaches, we need to know the conditional firing-rate probability density $p[\mathbf{r}|s]$ that describes this variability. If the firing rates are determined by counting spikes and the variability can be described by the Poisson model discussed in chapter 1, the expected number of spikes for neuron a with response tuning curve $f_a(s)$ over a trial of duration T is $f_a(s)T$. For the homogeneous Poisson model with this mean rate, the probability of neuron a firing n_a spikes is

$$P[n_a|s] = \frac{(f_a(s)T)^{n_a}}{n_a!} \exp(-f_a(s)T) \,. \tag{3.33}$$

We can express the number of spikes in terms of a firing rate r_a by writing $n_a = r_a T$. Since this rate takes discrete values, we can use the conditional probability $P[\mathbf{r}|s]$ rather than the density $p[\mathbf{r}|s]$ to describe the response variability of the population in this case. If we assume that each neuron fires independently, the firing-rate probability for the population is the product of the individual probabilities

$$P[\mathbf{r}|s] = \prod_{a=1}^{N} \frac{(f_a(s)T)^{r_a T}}{(r_a T)!} \exp(-f_a(s)T) \, . \tag{3.34}$$

The assumption of independence simplifies the calculations considerably.

The filled circles in figure 3.9 show a set of randomly generated firing rates for the array of Gaussian tuning curves in figure 3.8 for $s=0$. This figure also illustrates a useful way of visualizing population responses; plotting the response of each cell as a function of its preferred stimulus value. The dashed curve in figure 3.9 is the tuning curve for the neuron with $s_a = 0$. Since the tuning curves are translation invariant, the values of the dashed curve at $s = -5, -4, \ldots, 5$ are the mean activities of the cells with preferred values at those locations for a stimulus at $s=0$.

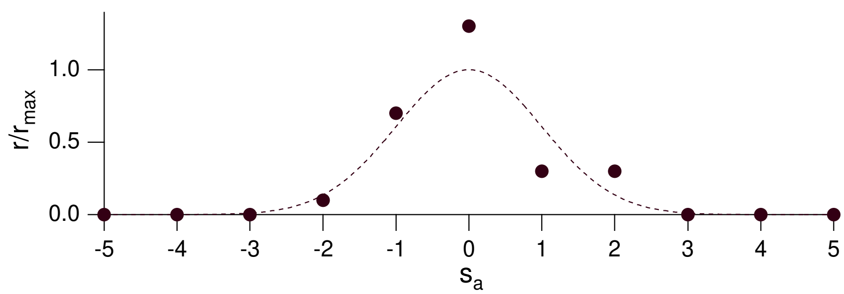

Figure 3.9: Simulated responses of 11 neurons with the Gaussian tuning curves shown in figure 3.8 to a stimulus value of zero. Firing rates for a single trial, generated using the Poisson model, are plotted as a function of the preferred stimulus values of the different neurons in the population (filled circles). The dashed curve shows the tuning curve for the neuron with $s_a = 0$. Its heights at integer values of s_a are the average responses of the associated cells.

To apply the MAP, ML and Bayesian estimation algorithms, we only need to consider the terms in $P[\mathbf{r}|s]$ that depend on s. Because equation 3.34 involves a product, it is convenient to take logarithms and write

$$\ln P[\mathbf{r}|s] = T \sum_{a=1}^{N} r_a \ln\big(f_a(s)T\big) + \text{const} \tag{3.35}$$

where const represents all the terms independent of s, including (as we discussed above) $\sum f_a(s)$. Since maximizing a function and maximizing its logarithm are equivalent, we can use the logarithm of the conditional probability in place of the actual probability in ML decoding.

The ML estimated stimulus, s_{ML}, is the stimulus that maximizes equation 3.35. Setting the derivative to zero, we find that s_{ML} is determined by

$$\sum_{a=1}^{N} \frac{r_a f_a'(s_{\mathrm{ML}})}{f_a(s_{\mathrm{ML}})} = 0 \tag{3.36}$$

where the prime denotes a derivative. If the tuning curves are the Gaussians of equation 3.32, this equation can be solved explicitly using the result $f_a'(s)/f_a(s) = (s_a - s)/\sigma_a^2$,

$$s_{\mathrm{ML}} = \frac{\sum r_a s_a/\sigma_a^2}{\sum r_a/\sigma_a^2} \,. \tag{3.37}$$

If all the tuning curves have the same width this reduces to

$$s_{\mathrm{ML}} = \frac{\sum r_a s_a}{\sum r_a} \,, \tag{3.38}$$

which is a simple estimation formula with an intuitive interpretation as the firing-rate weighted average of the preferred values of the encoding neurons. The numerator of this expression is reminiscent of the population vector. Even though it gives the ML estimate for a population of neurons with Poisson variability, equation 3.38 has some undesirable properties that can cause problems if the Poisson model is not exactly true. Consider a neuron with a preferred stimulus value s_a that is much greater than the actual stimulus value s. Since $s_a \gg s$, the average firing rate of this neuron is essentially zero. For a Poisson distribution, zero rate implies zero variability. If, however, this neuron fires one or more spikes on a trial due to a non-Poisson source of variability, this will cause a large error in the estimate because of the large weighting factor s_a.

If the prior distribution over s is flat, the MAP and ML algorithms are the same, but the MAP estimate s_{MAP} may differ from s_{ML} if $p[s]$ is not constant. The MAP estimate is computed from the distribution $p[s|\mathbf{r}]$ determined by Bayes theorem. In terms of the logarithms of the probabilities, $\ln p[s|\mathbf{r}] = \ln P[\mathbf{r}|s] + \ln p[s] - \ln P[\mathbf{r}]$. This last term is independent of s and can be absorbed into a new constant (const), so we can write

$$\ln p[s|\mathbf{r}] = T \sum_{a=1}^{N} r_a \ln\big(f_a(s)\big) + \ln p[s] + \text{const} \,. \tag{3.39}$$

Minimizing this determines the MAP estimate,

$$T \sum_{a=1}^{N} \frac{r_a f_a'(s_{\mathrm{MAP}})}{f_a(s_{\mathrm{MAP}})} + \frac{p'[s_{\mathrm{MAP}}]}{p[s_{\mathrm{MAP}}]} = 0 \,. \tag{3.40}$$

If the prior distribution over the stimulus is itself Gaussian with mean s_{prior} and variance σ_{prior}, and we use the Gaussian array of tuning curves,

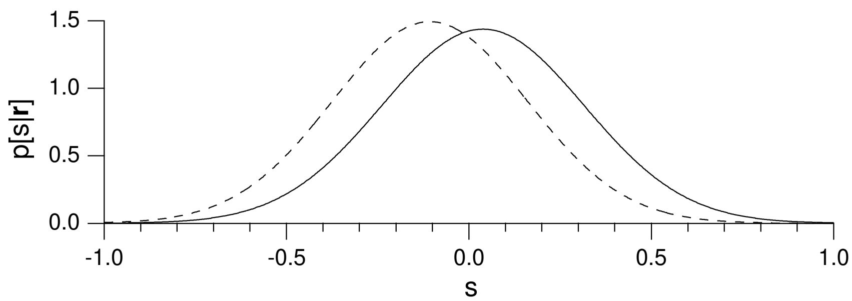

Figure 3.10: Probability densities for the stimulus $s=0$ given the tuning curves of figure 3.8 and the firing rates shown in figure 3.9. The solid curve is $p[s|\mathbf{r}]$ when the prior distribution of stimulus values is flat and the dashed curve is for a Gaussian prior distribution with a mean of -2 and variance of 1. The peak of the solid and dashed curves are at 0.0385 and -0.107 respectively.

equation 3.40 yields

$$s_{\text{MAP}} = \frac{T\sum r_a s_a/\sigma_a^2 + s_{\text{prior}}/\sigma_{\text{prior}}^2}{T\sum r_a/\sigma_a^2 + 1/\sigma_{\text{prior}}^2} . \tag{3.41}$$

Figure 3.10 compares the conditional stimulus probability densities $p[s|\mathbf{r}]$ for a flat prior distribution (solid curve) and for a Gaussian prior distribution with $s_{\text{prior}} = -2$ and $\sigma_{\text{prior}} = 1$, using the firing rates given by the filled circles in figure 3.9. When the prior distribution is flat, $p[s|\mathbf{r}]$ is peaked near the true stimulus value of zero. The effect of the prior distribution is to shift the curve toward the value -2 where the density of the prior has its maximum, and to decrease its width by a small amount (this can be seen by noting that the peak of the dashed curve is slightly higher and both curves have unit area). The decreased width is due to the added information that the prior distribution provides. The functional form of the curves in figure 3.10 can be computed from equations 3.32 and 3.39 as Gaussians with variances $1/(T\sum r_a/\sigma_a^2)$ (flat prior) and $1/(T\sum r_a/\sigma_a^2 + 1/\sigma_{\text{prior}}^2)$ (Gaussian prior).

Fisher Information

The maximum likelihood estimate has important properties that make it a standard of comparison for other decoding methods. These properties are exact only in the limit when the number of neurons goes to infinity (assuming an appropriate distribution of tuning curves over the range of encoded stimulus values), but they often apply to a good approximation even for reasonably small numbers of neurons. In the large N limit, the prior information $p[s]$ often has less impact than the likelihood, so the MAP and ML methods typically coincide.

The average over trials of any estimate s_{est}, when the true stimulus value is s, can be written as $\langle s_{\text{est}} \rangle = s + b(s)$ where $b(s)$ is called the bias. The average squared estimation error of such an estimator for this stimulus value is

$$\langle (s_{\text{est}} - s)^2 \rangle = \langle (s_{\text{est}} - \langle s_{\text{est}} \rangle + b(s))^2 \rangle = \sigma_{\text{est}}^2(s) + b^2(s) \qquad (3.42)$$

where $\sigma_{\text{est}}^2(s) = \langle (s_{\text{est}} - \langle s_{\text{est}} \rangle)^2 \rangle$ is the trial-to-trial variance of the estimator for stimulus value s. In general minimizing this error involves a trade-off between minimizing the bias and the variance of the estimator. If, however, the estimator is unbiased so that $b(s) = 0$, the estimation error is minimized simply by minimizing the variance $\sigma_{\text{est}}^2(s)$.

The variance of any estimator with bias $b(s)$ satisfies an inequality known as the Cramér-Rao bound (a proof is given in Appendix B),

$$\sigma_{\text{est}}^2(s) \geq \frac{(1 + b'(s))^2}{I_{\text{F}}(s)} \qquad (3.43)$$

where $b'(s)$ is the derivative of $b(s)$. The quantity $I_{\text{F}}(s)$ is the Fisher information, which is related to $p[\mathbf{r}|s]$ by

$$I_{\text{F}}(s) = \left\langle -\frac{\partial^2 \ln p[\mathbf{r}|s]}{\partial s^2} \right\rangle = \int d\mathbf{r}\, p[\mathbf{r}|s] \left(-\frac{\partial^2 \ln p[\mathbf{r}|s]}{\partial s^2} \right) . \qquad (3.44)$$

The Fisher information can be written in other equivalent ways:

$$I_{\text{F}}(s) = \left\langle \left(\frac{\partial \ln p[\mathbf{r}|s]}{\partial s} \right)^2 \right\rangle = \int d\mathbf{r}\, p[\mathbf{r}|s] \left(\frac{\partial \ln p[\mathbf{r}|s]}{\partial s} \right)^2 \qquad (3.45)$$

or

$$I_{\text{F}}(s) = \int d\mathbf{r}\, \frac{1}{p[\mathbf{r}|s]} \left(\frac{\partial p[\mathbf{r}|s]}{\partial s} \right)^2 . \qquad (3.46)$$

We leave it as an exercise (problem 7) to prove the equivalence of these different expressions. For an unbiased estimator with $b(s) = 0$, the Cramér-Rao bound sets a limit on estimation accuracy, since, in this case, the average squared estimation error is $\sigma_{\text{est}}^2 \geq 1/I_{\text{F}}(s)$ by equation 3.43. This bound makes the Fisher information a useful measure of coding accuracy.

The significance of the ML estimate arises from the fact that, in the limit of large N, it is unbiased and saturates the Cramér-Rao bound. In other words, the variance of the ML estimate is given asymptotically (for large N) by $\sigma_{\text{ML}}^2(s) = 1/I_{\text{F}}(s)$ (for a proof see Cox and Hinkley, 1974). In addition, the ML estimate obeys a central limit theorem, so that s_{ML} is Gaussian distributed in the appropriate limit. Note that the Cramér-Rao lower bound does not prevent a biased estimator from having a lower variance than the value $1/I_{\text{F}}(s)$ of the ML estimate (for example, an estimate that never changes has zero variance). For a biased estimator, the average

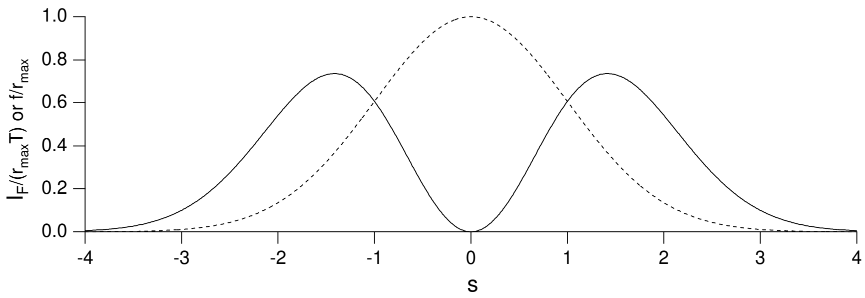

Figure 3.11: The Fisher information for a single neuron with a Gaussian tuning curve centered at $s=0$ with $\sigma_a=1$ and Poisson variability. The Fisher information (solid curve) has been divided by $r_{\max}T$, the peak firing rate of the tuning curve times the duration of the trial. The dashed curve shows the tuning curve scaled by $r_{\max}$. Note that the Fisher information is greatest where the slope of the Gaussian is high, and vanishes at $s=0$ where the tuning curve peaks.

squared estimation error and the variance are not equal, and the estimation error can be either larger or smaller than the ML variance.

The Fisher information for a population of neurons with uniformly arrayed tuning curves (the Gaussian array in figure 3.8, for example) can be computed from equation 3.35. Recall that $\sum f_a(s)$ is independent of s for such an array, so any derivatives of this sum, such as $\sum f_a'(s)$, can be set to zero. Using this, we find (Paradiso, 1988; Seung and Sompolinsky, 1993)

$$I_{\rm F}(s) = T\sum_{a=1}^{N}\langle r_a\rangle\left(\frac{f_a'(s)}{f_a(s)}\right)^2 = T\sum_{a=1}^{N}\frac{\left(f_a'(s)\right)^2}{f_a(s)} \tag{3.47}$$

since $\langle r_a\rangle = f_a(s)$. Therefore, under the assumption of independent Poisson variability in the measured firing rates, each neuron contributes an amount to the Fisher information proportional to the square of its tuning curve slope (the signal) and inversely proportional to the average firing rate (the noise) for the particular stimulus value being estimated. Highly sloped tuning curves give firing rates that are sensitive to the precise value of the stimulus. Figure 3.11 shows the contribution to this sum from a single neuron with a Gaussian tuning curve, the neuron with $s_a=0$ in figure 3.8. For comparison purposes, a dashed curve proportional to the tuning curve is also plotted. Note that the Fisher information vanishes for the stimulus value that produces the maximum average firing rate because $f_a'(s)=0$ at this point. A neuron at the peak of its tuning curve is relatively unaffected by small changes in the stimulus. Neurons carry the most Fisher information in regions of their tuning curves where average firing rates are rapidly varying functions of the stimulus value.

If we substitute Gaussian tuning curves into equation 3.47 and use the fact that, for a dense, uniform distribution of tuning curves all with $\sigma_a = \sigma$,

$$\sum_{a=1}^{N} \frac{(f_a'(s))^2}{f_a(s)} = \sum_{a=1}^{N} \frac{r_{\max}(s-s_a)^2}{\sigma^4} \exp\left(-\frac{1}{2}\left(\frac{s-s_a}{\sigma}\right)^2\right) \approx \frac{\sqrt{2\pi}\rho\sigma r_{\max}}{\sigma^2}, \tag{3.48}$$

we find

$$I_{\mathrm{F}}(s) = \frac{\sqrt{2\pi}\rho\sigma r_{\max} T}{\sigma^2}. \tag{3.49}$$

Here ρ is the number of neurons with preferred stimulus values s_a lying within a unit range of s values. This determines the density with which the neurons cover the range of stimulus values. The number of neurons that respond to a given stimulus value is roughly $\rho\sigma$, and the Fisher information is proportional to this number divided by the square of the tuning curve width. The corresponding variance $1/I_{\mathrm{F}}$ is inversely proportional to the number of responding neurons.

Instead of a stimulus that is parameterized by a single variable s, consider a stimulus described by D parameters and suppose that the response tuning curves are products of identical Gaussians for all of these parameters. If the tuning curves cover the D-dimensional space of stimulus values with a uniform density ρ, the number of responding neurons for any stimulus value is proportional to $\rho\sigma^D$ and the Fisher information is

$$I_{\mathrm{F}} = \frac{(2\pi)^{D/2}\rho\sigma^D r_{\max} T}{\sigma^2} = (2\pi)^{D/2}\rho\sigma^{D-2} r_{\max} T. \tag{3.50}$$

This equation, which reduces to the result given above if $D = 1$, allows us to examine the effect of changing the tuning curve width on encoding accuracy. As noted by Zhang and Sejnowski (1998), the effect depends on the dimensionality of the encoded stimulus. When tuning curve width is changed, the tradeoff between the encoding accuracy of individual neurons and the number of responding neurons depends on the dimension of the stimulus space. Narrowing the tuning curves (making σ smaller) increases the Fisher information for $D = 1$, decreases it for $D > 2$, and has no impact if $D = 2$.

Optimal Discrimination

In the first part of this chapter, we considered discrimination between two values of a stimulus. An alternative to the procedures discussed there is simply to decode the responses and discriminate on the basis of the estimated stimulus values. Consider the case of discriminating between s and $s+\Delta s$ for small Δs. For large N, the average value of the difference between the ML estimates for the two stimulus values is equal to Δs (because the estimate is unbiased) and the variance of each estimate (for small

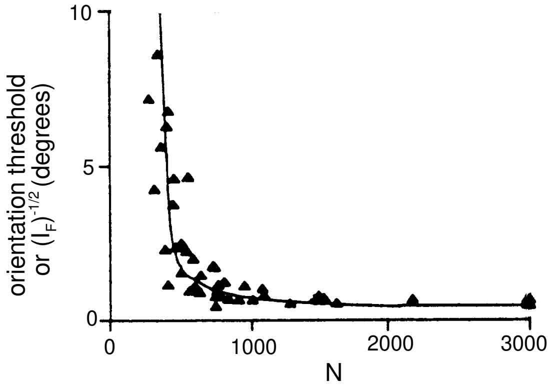

Figure 3.12: Comparison of Fisher information and discrimination thresholds for orientation tuning. The solid curve is the minimum standard deviation of an estimate of orientation angle from the Cramér-Rao bound, plotted as a function of the number of neurons (N) involved in the estimation. The triangles are data points from an experiment that determined the threshold for discrimination of the orientation of line images by human subjects as a function of line length and eccentricity. An effect number of neurons was computed for the different line lengths and eccentricities using the cortical magnification factor (see chapter 2). (Adapted from Paradiso, 1988)

Δs) is $1/I_F(s)$. Thus, the discriminability of the ML based test is

$$d' = \Delta s \sqrt{I_F(s)} \,. \tag{3.51}$$

The larger the Fisher information, the higher the discriminability. We leave to exercise 9 the proof that for small Δs, this discriminability is the same as that of the likelihood ratio test $Z(\mathbf{r})$ defined in equation 3.20.

Discrimination by ML estimation requires maximizing the likelihood, which may be computationally challenging. The likelihood ratio test may be simpler, especially for Poisson variability, since, for small Δs, the likelihood ratio test Z defined in equation 3.20 is a linear function of the firing rates in this case,

$$Z = T \sum_{a=1}^{N} \frac{r_a f_a'(s)}{f_a(s)} \,. \tag{3.52}$$

Figure 3.12 shows an interesting comparison of the Fisher information for orientation tuning in the primary visual cortex and human orientation discrimination thresholds. Agreement like this can occur for difficult tasks, like discrimination at threshold, where the performance of a subject may be limited by basic constraints on neuronal encoding accuracy.

Optimal Linear Decoding

Optimal decoding using the Bayesian, MAP, or ML algorithms requires knowledge of the full conditional probability distribution for the stimulus. If this is not available, simpler decoding methods that require less information must be used. Among the simplest schemes is an estimate which depends only linearly on the firing rates:

$$s_{\mathrm{OLE}} = \sum_{a=1}^{N} r_a w_a \tag{3.53}$$

where the weights w_a are chosen to optimize the accuracy of the estimate. To find the optimal linear estimator (OLE), we use a Bayesian approach and minimize the average loss using the squared difference between the estimated and true stimulus values averaged over trials as a measure of accuracy. The quantity to be to be minimized is

$$E = \int ds\, p[s] \left\langle \left(s - \sum_{a=1}^{n} r_a w_a \right)^2 \right\rangle . \tag{3.54}$$

Setting the derivative of this error with respect to w_a to zero gives the condition (this computation is the matrix analog of the derivation of the optimal linear kernel for firing-rate estimation in chapter 1)

$$\int ds\, p[s] s \langle r_a \rangle = \int ds\, p[s] \sum_{b=1}^{N} \langle r_a r_b \rangle\, w_b . \tag{3.55}$$

To solve for the weights we must invert the stimulus-averaged correlation matrix of the firing rates

$$M_{ab} = \int ds\, p[s]\, \langle r_a r_b \rangle \tag{3.56}$$

to obtain

$$w_a = \sum_{b=1}^{N} [M^{-1}]_{ab} \int ds\, p[s] s \langle r_b \rangle . \tag{3.57}$$

Considering its simplicity, the optimal linear estimator (OLE) can perform remarkably well. Figure 3.13 compares an OLE decoding of arm movement direction from the activity of neurons recorded in the primary motor cortex of a monkey performing reaching movements in various directions with the results of the vector decoding algorithm. The root-mean-square decoding error is plotted as a function of the number of neurons used in the analysis. The OLE performs considerably better than the vector method, especially when a relatively small number of neurons is used.

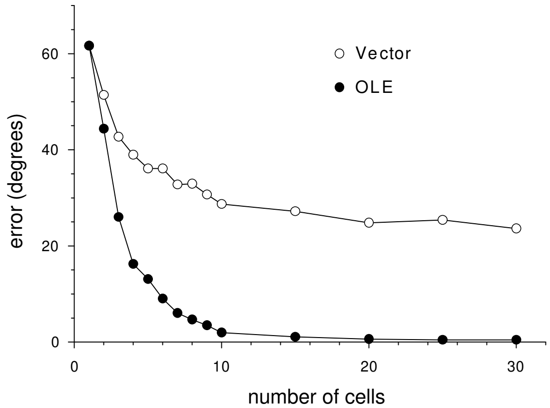

Figure 3.13: Comparison of the OLE and vector methods for reconstruction of arm movement directions from recordings of neurons in the monkey motor cortex. The plot is from an analysis by Pellizzer and Georgopoulos. Each plotted point represents the root-mean-square average angular decoding error as a function of the number of cells used in the analysis. (Adapted from Salinas and Abbott, 1994)

3.4 Spike Train Decoding

The decoding methods we have considered thus far estimate or discriminate static stimulus values on the basis of time-averaged firing rates of one or more neurons. Time-averaged firing rates do not provide sufficient information for reconstructing a stimulus that varies during the course of a trial. Instead, we can estimate such a stimulus directly from the sequence of firing times t_i for $i = 1, 2, \ldots, n$ of the spikes that it evokes. One method for doing this is similar to the Wiener kernel approach used to estimate the firing rate from the stimulus in chapters 1 and 2. It is also related to the definition of a firing rate using a sliding window function introduced in chapter 1. For simplicity, we restrict our discussion to the decoding of a single neuron.

Spike train decoding attempts to construct an estimate of the stimulus at time t from the sequence of spikes evoked up to that time. There are paradoxical aspects of this procedure. The firing of an action potential at time t_i is only affected by the behavior of the stimulus $s(t)$ prior to that time, $t < t_i$, and yet, in spike decoding, we attempt to extract information from this action potential about the value of the stimulus at a later time $t > t_i$. The evoked spikes tell us about the past behavior of the stimulus and, in spike decoding, we attempt to use this information to predict the current stimulus value. Clearly, this requires that the stimulus have some form of temporal correlation so that past behavior provides information about the current stimulus value. To make the decoding task easier, we can introduce a prediction delay, τ_0, and attempt to construct from spikes occurring prior to time t, an estimate of the stimulus at time $t - \tau_0$. Such a delayed estimate uses a combination of spikes that were fired in response to the stimulus being estimated (those with $t - \tau_0 < t_i \leq t$), and spikes that have not yet reacted to the stimulus at the time it is being estimated, but that can be used to predict it on the basis of stimulus correlations (those with $t_i \leq t - \tau_0$). The estimation task gets easier as τ_0 is increased, but this delays the decoding and makes the result less behaviorally relevant. We will consider decoding with an arbitrary delay and later discuss how to set a specific value for τ_0.

The stimulus estimate is constructed as a linear sum over all spikes. A spike occurring at time t_i contributes a kernel $K(t - t_i)$, and the total estimate is obtained by summing over all spikes,

$$s_{\text{est}}(t - \tau_0) = \sum_{i=1}^{n} K(t - t_i) - \bar{r} \int d\tau\, K(\tau)\,. \tag{3.58}$$

The last term, with $\bar{r} = \langle n \rangle / T$ the average firing rate over the trial, is included to impose the condition that the time average of s_{est} is zero (recall that we imposed this condition on s in chapter 1). The sum in equation 3.58 includes all spikes so the constraint that only those spikes occurring prior to the time t should be included must be imposed by requiring $K(t - t_i) = 0$ for $t - t_i \leq 0$. We will ignore this 'causality' constraint for now, but return to it later in the discussion.

Equation 3.58 can be written in a compact way by using the neural response function $\rho(t) = \sum \delta(t - t_i)$ introduced in chapter 1,

$$s_{\text{est}}(t - \tau_0) = \int d\tau\; (\rho(t - \tau) - \bar{r})\, K(\tau)\,. \tag{3.59}$$

Using this form of the estimate, the construction of the optimal kernel K proceeds very much like the construction of the optimal kernel for predicting firing rates in chapter 1. We choose K so that the squared difference between the estimated stimulus and the actual stimulus, averaged over both time and trials,

$$E = \frac{1}{T} \int_0^T dt \left\langle \left(\int d\tau\; (\rho(t - \tau) - \bar{r})\, K(\tau) - s(t - \tau_0) \right)^2 \right\rangle \tag{3.60}$$

is minimized. The calculation proceeds as in Appendix B of chapter 1 (see exercise 10), and the result is that K obeys the equation

$$\int d\tau'\, Q_{\rho\rho}(\tau - \tau') K(\tau') = Q_{\rho s}(\tau - \tau_0)\,. \tag{3.61}$$

where $Q_{\rho\rho}$ is the spike-train autocorrelation function,

$$Q_{\rho\rho}(\tau' - \tau) = \frac{1}{T} \int_0^T dt \left\langle (\rho(t) - \bar{r})(\rho(t + \tau' - \tau) - \bar{r}) \right\rangle, \tag{3.62}$$

as defined in chapter 1, and $Q_{\rho s}$ is the correlation of the average neural response function and the stimulus, which is related to the spike-triggered average C, both introduced in chapter 1,

$$\frac{1}{T}\left\langle \sum_{i=1}^{n} s(t_i + \tau - \tau_0) \right\rangle = Q_{\rho s}(\tau - \tau_0) = \bar{r}C(\tau_0 - \tau)\,. \tag{3.63}$$

At this point in the derivation of the optimal linear kernel for firing-rate prediction in chapter 1, we chose the stimulus to be uncorrelated so that an integral equation similar to 3.61 simplified. This could always be done because we have complete control over the stimulus in this type of experiment. However, we do not have similar control of the neuron, and must deal with whatever spike train autocorrelation function it gives us. If the spike train happens to be uncorrelated like a Poisson spike train, so that

$$Q_{\rho\rho}(\tau) = \bar{r}\delta(\tau) \tag{3.64}$$

we find from equation 3.61 that

$$K(\tau) = \frac{1}{\bar{r}} Q_{\rho s}(\tau - \tau_0) = C(\tau_0 - \tau) = \frac{1}{\langle n \rangle}\left\langle \sum_{i=1}^{n} s(t_i + \tau - \tau_0) \right\rangle . \tag{3.65}$$

This is the average value of the stimulus at time $\tau - \tau_0$ relative to the appearance of a spike. Since $\tau - \tau_0$ can be either positive or negative, stimulus estimation, unlike firing rate estimation, involves both forward and backward correlation and the average values of the stimulus both before and after a spike.

When the spike train autocorrelation function is not a δ function, the solution for K can be expressed as an inverse Fourier transform,

$$K(\tau) = \frac{1}{2\pi}\int d\omega\, \tilde{K}(\omega) \exp(-i\omega\tau) \tag{3.66}$$

where, as shown in Appendix C,

$$\tilde{K}(\omega) = \frac{\tilde{Q}_{\rho s}(\omega)\exp(i\omega\tau_0)}{\tilde{Q}_{\rho\rho}(\omega)}\,. \tag{3.67}$$

By itself, the numerator in this expression reproduces the expression $Q_{\rho s}(\tau - \tau_0)$ in equation 3.65. The role of the denominator is to correct for any correlations in the response spike train.

Because we have not explicitly imposed the causality constraint, the kernel computed from equations 3.65 or from 3.66 and 3.67 will not, in general, satisfy $K(\tau) = 0$ for $\tau < 0$. This constraint can be imposed by choosing the time-delay factor τ_0 appropriately and by restricting K. The discussion is simpler if we consider uncorrelated spike trains so that the kernel is given by equation 3.65, but the analysis applies to the general case as well. The spike triggered average $C(\tau)$ vanishes for sufficiently large values of τ because the spike occurrence does not depend on the behavior of

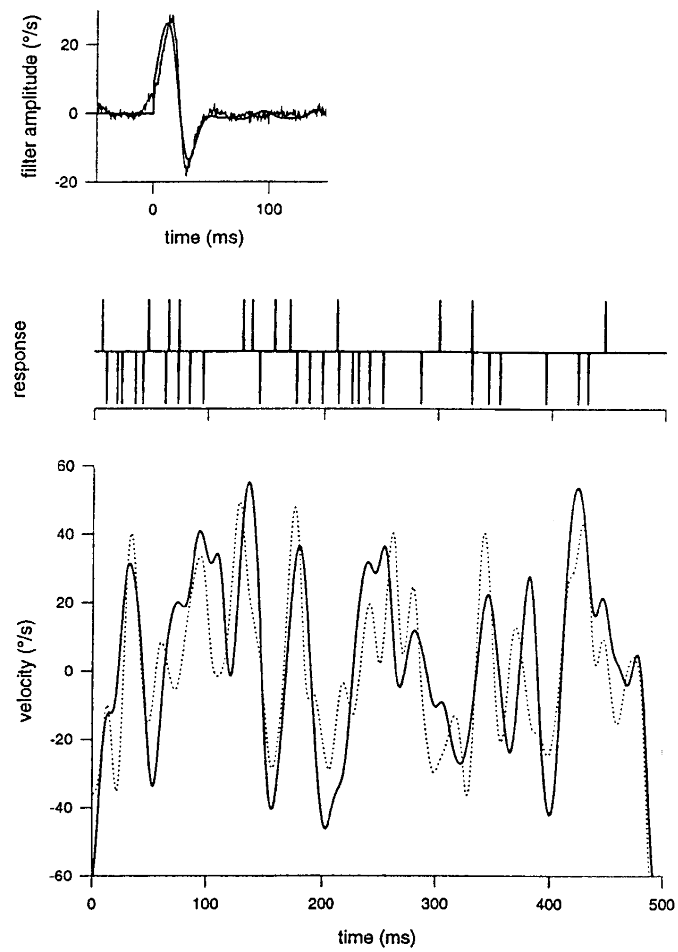

Figure 3.14: Decoding the stimulus from an H1 neuron of the fly. The upper panel is the decoding kernel. The jagged curve is the optimal acausal filter and the smooth curve is a kernel obtained by expanding in a causal set of basis functions. In both cases, the kernels are shifted by $\tau_0 = 40$ ms. The middle panel shows typical responses of the H1 neuron to the stimulus $s(t)$ (upper trace) and $-s(t)$ (bottom trace). The dashed line in the lower panel shows the actual stimulus and the solid line is the estimated stimulus from the optimal linear reconstruction. (Adapted from Rieke *et al*, 1997)

a stimulus in the distant past. For most neurons, $C(\tau) \approx 0$ for τ greater than a few hundred ms. The decoding kernel given by equation 3.65 is $K(\tau) = C(\tau_0 - \tau)$ and we can assure that this is small for negative values of τ by choosing τ_0 large enough. Unfortunately, this requires a fairly large prediction delay. Correlations in the spike train, which require equations 3.66 and 3.67 to be used to determine the decoding kernel, may shorten the required delay. In either case, once the value of τ_0 has been determined, any residual nonzero value of the kernel can be eliminated by replacing $K(\tau)$ by $\Theta(\tau)K(\tau)$ where $\Theta(\tau)$ is the Heaviside function which is one for positive values and zero for negative values of its argument.

Another way of imposing causality on the decoding kernel is to expand $K(\tau)$ as a weighted sum of causal basis functions (functions that vanish for negative arguments and span the space of functions satisfying the causal constraint). The optimal weights are determined by minimizing the estimation error. This approach has the advantage of producing a truly optimal kernel for any desired value of τ_0. A simpler but non-optimal approach is to consider a fixed functional form for $K(\tau)$ that vanishes for $\tau \leq 0$ and is characterized by a number of free parameters that can be determined by minimizing the decoding error.

Figure 3.14 shows an example of spike train decoding for the H1 neuron of the fly discussed in chapter 1. The top panel shows two versions of the reconstruction kernel used in the decoding, and the bottom panel compares the reconstructed stimulus velocity with the actual stimulus velocity. The smoother reconstruction kernel is computed by expanding $K(\tau)$ in a causal basis with $\tau_0 = 40$ ms. The noisier kernel is the optimal acausal filter shifted by $\tau_0 = 40$ ms. The reconstruction in the bottom panel uses the acausal filter.

The middle panel in figure 3.14 points out one further wrinkle in the procedure. Flies have two H1 neurons, one on each side of the body, that respond maximally to motion in opposite directions. As is often the case, rectification prevents a single neuron from encoding both directions of motion simultaneously. In the experiment described in the figure, rather than recording from both H1 neurons, Bialek *et al* (1991) recorded from a single H1 neuron, but presented both the stimulus $s(t)$ and its negative, $-s(t)$. The two rows of spikes in the middle panel show sample traces for each of these presentation. This procedure provides a reasonable approximation of recording both H1 neurons provided that there are no correlations between their activities. Instead of equation 3.58, reconstruction is based on a double sum (representing, as for discrimination in area MT, one neuron and one antineuron),

$$s_{\mathrm{est}}(t) = \sum_{i=1}^{n} K(t - t_i) - \sum_{i=1}^{n'} K(t - t'_i) - (\bar{r} - \bar{r}') \int d\tau\, K(\tau)\,. \tag{3.68}$$

The spike times, spike number, and rate marked with primes represent the response to the inverted stimulus $-s(t)$.

Stimulus reconstruction using this technique is quite accurate. Whereas the monkey has a very large population of MT neurons reporting on motion directions, the fly only has two H1 neurons, so it may be less surprising that they support excellent inference all by themselves. On the other hand, the decoding from H1 reconstructs a time-dependent stimulus which is more challenging than a binary discrimination. Finally, it is worth remembering that both the fly and monkey experiments involve stimuli that are extremely simplified compared to the naturally occurring visual images that must be interpreted during normal behavior.

3.5 Appendices

A) The Neymann-Pearson Lemma

Consider the difference $\Delta\beta$ in the power of two tests that have identical sizes α. One uses the likelihood ratio $l(r)$, and the other uses a different test function $f(r)$. For a test $f(r)$,

$$\alpha(z) = \int_{f(r)\geq z} dr\, p[r|-] \quad \text{and} \quad \beta(z) = \int_{f(r)\geq z} dr\, p[r|+] \,. \tag{3.69}$$

Similar equations hold for the test $l(r)$. Comparing the two β's, we find

$$\Delta\beta = \int_{l(r)\geq z} dr\, p[r|+] \quad - \quad \int_{f(r)\geq z} dr\, p[r|+] \,. \tag{3.70}$$

Eliminating terms that cancel between the two integrals, we can write

$$\Delta\beta = \int_{\substack{l(r)\,\geq\, z\\ f(r)\,<\, z}} dr\, p[r|+] \quad - \quad \int_{\substack{l(r)\,<\, z\\ f(r)\,\geq\, z}} dr\, p[r|+] \,. \tag{3.71}$$

Using the definition $l(r) = p[r|+]/p[r|-]$ this can be re-expressed as

$$\Delta\beta = \left[\int_{\substack{l(r)\,\geq\, z\\ f(r)\,<\, z}} dr\, l(r) p[r|-] \quad - \quad \int_{\substack{l(r)\,<\, z\\ f(r)\,\geq\, z}} dr\, l(r) p[r|-]\right] \tag{3.72}$$

This leads to the inequality

$$\Delta\beta \geq z\left[\int_{\substack{l(r)\,\geq\, z\\ f(r)\,<\, z}} dr\, p[r|-] \quad - \quad \int_{\substack{l(r)\,<\, z\\ f(r)\,\geq\, z}} dr\, p[r|-]\right] \tag{3.73}$$

due to the conditions of $l(r)$ within the integrals. Putting back the terms that we canceled at the first step of this derivation, we find

$$\Delta\beta \geq z\left[\int_{l(r)\geq z} dr\, p[r|-] \quad - \quad \int_{f(r)\geq z} dr\, p[r|-]\right] . \tag{3.74}$$

By definition, these integrals are the sizes of the two tests, which are equal. Thus $\Delta\beta \geq 0$, showing that no test can be better than the likelihood ratio $l(r)$, at least in the sense of maximizing the power for a given size.

B) The Cramér-Rao Bound

The Cramér-Rao lower bound for an estimator s_{est} comes from differentiating the equation for its bias. By definition,

$$s + b(s) = \langle s_{\mathrm{est}} \rangle = \int d\mathbf{r}\, p[\mathbf{r}|s] s_{\mathrm{est}}\,. \tag{3.75}$$

Differentiating with respect to s gives

$$\begin{aligned} 1 + b'(s) &= \int d\mathbf{r}\, \frac{\partial p[r|s]}{\partial s} s_{\mathrm{est}} \\ &= \int d\mathbf{r}\, p[\mathbf{r}|s] \frac{\partial \ln p[\mathbf{r}|s]}{\partial s} s_{\mathrm{est}} \\ &= \int d\mathbf{r}\, p[\mathbf{r}|s] \frac{\partial \ln p[\mathbf{r}|s]}{\partial s} (s_{\mathrm{est}} - \langle s_{\mathrm{est}} \rangle)\,. \end{aligned}$$

The last equality follows from the identity

$$\int d\mathbf{r}\, p[\mathbf{r}|s] \frac{\partial \ln p[\mathbf{r}|s]}{\partial s} \langle s_{\mathrm{est}} \rangle = \langle s_{\mathrm{est}} \rangle \int d\mathbf{r}\, \frac{\partial p[\mathbf{r}|s]}{\partial s} = 0 \tag{3.76}$$

because $\int d\mathbf{r}\, p[\mathbf{r}|s] = 1$. Defining

$$\sigma^2_{\mathrm{est}}(s) = \int d\mathbf{r}\, p[\mathbf{r}|s] (s_{\mathrm{est}} - \langle s_{\mathrm{est}} \rangle)^2 = \left\langle (s_{\mathrm{est}} - \langle s_{\mathrm{est}} \rangle)^2 \right\rangle \tag{3.77}$$

and using the definition 3.45 we have, from the Cauchy-Schwarz inequality,

$$\sigma^2_{\mathrm{est}}(s) I_{\mathrm{F}} \geq \left(\int d\mathbf{r}\, p[\mathbf{r}|s] \frac{\partial \ln p[\mathbf{r}|s]}{\partial s} (s_{\mathrm{est}} - \langle s_{\mathrm{est}} \rangle) \right)^2 = (1 + b'(s))^2 \tag{3.78}$$

which, when rearranged, is the Cramér-Rao bound of equation 3.43. The Cauchy-Schwarz inequality states that for any two quantities A and B

$$\langle A^2 \rangle \langle B^2 \rangle \geq \langle AB \rangle^2\,. \tag{3.79}$$

To prove this inequality, note that

$$\left\langle \left(\langle B^2 \rangle A - \langle AB \rangle B \right)^2 \right\rangle \geq 0 \tag{3.80}$$

because it is the average value of a square. Computing the square gives

$$\langle B^2 \rangle^2 \langle A^2 \rangle - \langle AB \rangle^2 \langle B^2 \rangle \geq 0 \tag{3.81}$$

from which the inequality follows directly.

C) The Optimal Spike-Decoding Filter

The optimal linear kernel for spike train decoding is determined by solving equation 3.61. This is done by taking the Fourier transform of both sides of the equation, that is, multiplying both sides by $\exp(i\omega\tau)$ and integrating over τ,

$$\int d\tau \, \exp(i\omega\tau) \int d\tau' \, Q_{\rho\rho}(\tau - \tau') K(\tau') = \int d\tau \, \exp(i\omega\tau) Q_{\rho s}(\tau - \tau_0) \, . \quad (3.82)$$

By making the replacement of integration variable $\tau \to \tau + \tau_0$, we find that the right side of this equation is

$$\exp(i\omega\tau_0) \int d\tau \, \exp(i\omega\tau) Q_{\rho s}(\tau) = \exp(i\omega\tau_0) \tilde{Q}_{\rho s}(\omega) \quad (3.83)$$

where $\tilde{Q}_{\rho s}(\omega)$ is the Fourier transform of $Q_{\rho s}(\tau)$. The integral of the product of two functions that appears on the left side of equations 3.61 and 3.82 is called a convolution. To evaluate the Fourier transform on the left side of equation 3.82, we make use of an important theorem stating that the Fourier transform of a convolution is the product of the Fourier transforms of the two functions involved (see chapter 12). Thus, according to this theorem

$$\int d\tau \, \exp(i\omega\tau) \int d\tau' \, Q_{\rho\rho}(\tau - \tau') K(\tau') = \tilde{Q}_{\rho\rho}(\omega) \tilde{K}(\omega) \quad (3.84)$$

where $\tilde{Q}_{\rho\rho}(\omega)$ and $\tilde{K}(\omega)$ are the Fourier transforms of $Q_{\rho\rho}(\tau)$ and $K(\tau)$ respectively. In other words,

$$\tilde{Q}_{\rho\rho}(\omega) = \int d\tau \, \exp(i\omega\tau) Q_{\rho\rho}(\tau) \text{ and } \tilde{K}(\omega) = \int d\tau \, \exp(i\omega\tau) K(\tau) \, . \quad (3.85)$$

Putting the left and right sides of equation 3.82 together as we have evaluated them, we find that

$$\tilde{Q}_{\rho\rho}(\omega) \tilde{K}(\omega) = \exp(i\omega\tau_0) \tilde{Q}_{\rho s}(\omega) \, . \quad (3.86)$$

Equation 3.67 follows directly from this result, and equation 3.66 then determines $K(\tau)$ as the inverse Fourier transform of $\tilde{K}(\omega)$.

3.6 Further Reading and References

Further Reading

Cox, DR & Hinckley, DV (1974) *Theoretical Statistics.* London:Chapman & Hall.

Gabbiani F & Koch C (1998) Principles of spike train analysis. In C Koch and I Segev (editors) *Methods in Neuronal Modeling: From Ions to Networks* Cambridge MA:MIT Press.

Green, DM & Swets, JA (1966) *Signal Detection Theory and Psychophysics*. Los Altos, CA:Peninsula Publishing.

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

References

Bialek W, Rieke F, de Ruyter van Steveninck RR & Warland D (1991) Reading a neural code. *Science* **252**:1854-1857.

Britten, KH, Shadlen, MN, Newsome, WT and Movshon, JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience* **12**:4745-4765.

Georgopoulos, AP, Kettner, RE & Schwartz, AB (1988) Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. *Neuroscience* **8**:2928-2937.

Georgopoulos, AP, Schwartz, AB & Kettner, RE (1986) Neuronal population coding of movement direction. *Science* **243**:1416-1419.

Humphrey DR, Schmidt, EM & Thompson, WD (1970) Predicting measures of motor performance from multiple cortical spike trains. *Science* **170**:758-761.

Kandel, ER, Schwartz, JH & Jessel, TM (1991) *Principles of Neural Science*. New York:Elsevier.

Lee, C, Rohrer, WH & Sparks, DL (1988) Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature* **332**:357-360.

Lee, D, Port, NL, Kruse, W & Georgopoulos, AP (1998) Variability and correlated noise in the discharge of neurons in motor and parietal areas of the primate cortex. *Journal of Neuroscience* **18**:1161-1170.

Lewis JE & Kristan WB (1998) A neuronal network for computing population vectors in the leech. *Nature* **391**:76-9.

Newsome, WT, Britten, KH & Movshon, JA (1989) Neural correlates of a perceptual decision. *Nature* **341**:52-54.

Paradiso, MA (1988) A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biological Cybernetics* **58**:35-49.

Salinas, E & Abbott, LF (1994) Vector reconstruction from firing rates. *Journal of Computational Neuroscience* **1**:89-107.

Sanger, TD (1994) Theoretical considerations for the analysis of population coding in motor cortex. *Neural Computation* **6**:29-37.

Seung, HS & Sompolinsky, H (1993) Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences of the United States of America* **90**:10749-10753.

Theunissen, FE & Miller, JP (1991) Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons. *Journal of Neurophysiology* **66**:1690-1703.

van Gisbergen, JAM, Van Opstal, AJ & Tax, AMM (1987) Collicular ensemble coding of saccades based on vector summation. *Neuroscience* **21**:541-555.

Zhang K & Sejnowski T (1998) Neural tuning: to sharpen or broaden? *Neural Computation* (in press).

Zohary E, Shadlen MN, & Newsome, WT (1994) Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* **370**:140-143.

3.7 Exercises

1. Suppose that the probability that a neuron responds with a firing rate between r and $r + \Delta r$ to a plus or a minus stimulus is $p[r|\pm]\Delta r$ where

$$p[r|\pm] = \frac{1}{\sqrt{2\pi}\sigma_r} \exp\left(-\frac{1}{2}\left(\frac{r - \bar{r}_\pm}{\sigma_r}\right)^2\right) .$$

   Choose the two mean rate parameters $\bar{r}_+$ and $\bar{r}_-$ and the single variance σ_r^2 so that these distributions rarely produce negative rates. Plot ROC curves for the discrimination of the plus and minus stimuli on the basis of this neuron for different values of the discriminability

$$d' = \frac{\bar{r}_+ - \bar{r}_-}{\sigma_r} .$$

   By simulation, determine the fraction of correct discriminations that can be made in a two-alternative forced choice task involving discriminating between plus-then-minus and minus-then-plus presentations of two stimuli. Show that the fractions of correct answer for different values of d' are equal to the areas under the corresponding ROC curves.

2. Model the responses of the cercal system of the cricket by using cosine tuning curves to determine mean response rates and a Poisson model to produce variability. Generate a large number of responses for a variety of wind directions randomly, use the vector method to decode them, and compare the decoded direction with the actual direction used to generated the responses to determine the decoding accuracy. Plot the root-mean-square decoding error as a function of wind direction.

3. Show that if an infinite number of unit vectors $\vec{c}_a$ are chosen from a probability distribution that is independent of direction, $\sum(\vec{v}\cdot\vec{c}_a)\vec{c}_a \propto \vec{v}$ for any vector $\vec{v}$. How does the sum approach this limit for a finite number of terms?

4. Repeat exercise 2 for the case of M1 neurons and arm movement direction (restricted to arm movements in a two-dimensional plane). Study how the accuracy of the vector decoding method depends on the number of neurons used. Compare these results with those obtained using ML and Bayesian decoding methods. For the Bayesian method, use the estimate given by equation 3.29.

5. Repeat exercise 4 including a Gaussian prior distribution for arm direction probabilities. Compare the results for both the predicted directions and the accuracy with those of exercise 4.

6. Show that the formulas for the Fisher information in equations 3.45 and 3.46 are equivalent to 3.44. Use the fact that $\int d\mathbf{r}\, P[\mathbf{r}|s] = 1$.

7. Show that the Bayesian estimator that minimizes the expected average value of the the loss function $L(s, s_{\text{bayes}}) = (s - s_{\text{bayes}})^2$ is the mean given by equation 3.29 and that the median corresponds to minimizing the the expected loss for $L(s, s_{\text{bayes}}) = |s - s_{\text{bayes}}|$.

8. The discriminability for the variable Z defined in equation 3.20 is the difference between the average Z values for the two stimuli $s + \Delta s$ and s divided by the standard deviation of Z. The average of the difference in Z values is

$$\langle \Delta Z \rangle = \int d\mathbf{r}\, \frac{\partial \ln P[\mathbf{r}|s]}{\partial s} \left(P[\mathbf{r}|s + \Delta s] - P[\mathbf{r}|s] \right) .$$

Show that for small Δs, $\langle \Delta Z \rangle = I_{\text{F}}(s)\Delta s$. Also prove that the average value of Z,

$$\langle Z \rangle = \int d\mathbf{r}\, P[\mathbf{r}|s] \frac{\partial \ln P[\mathbf{r}|s]}{\partial s}$$

is zero and that the variance of Z is $I_{\text{F}}(s)$. Computing the ratio, we find from these results that $d' = \Delta s \sqrt{I_{\text{F}}(s)}$ which gives the discriminability 3.51 of the ML estimator.

9. Use the electric fish model from chapter 1 exercise 9 to generate a spike train response to a stimulus $s(t)$ of your choosing. Decode the spike train and reconstruct the stimulus using an optimal linear filter. Compare the optimal decoding filter with the optimal kernel for rate prediction, $D(\tau)$. Determine the average squared error of your reconstruction of the stimulus. Examine the effect that various static nonlinearities in the model for the firing rate that generates the spikes have on the accuracy of the decoding.

10. Using the methods of Appendix B of chapter 1, minimize 3.60 to obtain 3.61. Then, solve equation 3.61 to derive the result in equations 3.66 and 3.67 using the method of Fourier transforms as in excercise 7 of chapter 1.